Automated Road Environment Perception for Safety Assessment with Tuning-Free Vision Foundation Models

Shuyi Yin

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Yinhai Wang, Chair

Xuegang Ban

Edward McCormack

Qing Shen

Program Authorized to Offer Degree:

Civil and Environmental Engineering

University of Washington

Abstract

Automated Road Environment Perception for Safety Assessment with Tuning-Free Vision Foundation Models

Shuyi Yin

Chair of the Supervisory Committee:
Yinhai Wang
Department of Civil and Environmental Engineering

Persistent roadway fatalities in the United States expose the limitations of reactive safety practice: engineering countermeasures are usually deployed after crashes accumulate and patterns emerge. The Safe System philosophy seeks to break this cycle by insisting that road environment be designed to proactively anticipate human error, and the International Road Assessment Programme (iRAP) Star Ratings materializes this philosophy by scoring every 100-meter segment on its ability to prevent human errors from happening and to tolerate these mistakes if they occur. Yet producing these ratings still relies on labor-intensive image annotation that covers only a few dozen kilometers per annotator per day. Meanwhile, vast archives of street-level imagery already blanket road networks at scale, and recent vision foundation models can segment objects they are not explicitly trained to recognize. Closing the gap between this technological capacity and today's

slow-moving manual workflows—so that Star Ratings can be generated quickly, consistently, and at scale—forms the core mission of this dissertation.

To do so, this research pursues three tightly linked objectives. First, it diagnoses why existing computer vision pipelines fall short of iRAP needs, formalizes road environment perception as a domain task for safety assessment, and identifies the Segment Anything Model (SAM) as a promising backbone for exploration. Secondly, it designs pipelines that recast infrastructure recognition as a semantic segmentation task, enriching SAM with either domain-aware priors that guide mask labeling or tailored prompts that map infrastructure class descriptions, all while preserving the model's zero-shot versatility. Thirdly, it integrates these methodological proposals into real-world workflows in elements of Safe System framework, demonstrating end-to-end scalability in agency contexts.

These goals materialize in three core contributions. Firstly, this dissertation demonstrates that emerging transportation big data sources, such as street-level images and connected vehicle data, streamline and automate key steps in road safety assessment, reducing manual effort while preserving analytical rigor for road infrastructure classes. Secondly, this dissertation introduces two vision pipelines that leverage tuning-free foundation models, augmented with domain priors and engineered class prompts, to steer semantic segmentation for infrastructure element recognition without task-specific fine-tuning. Grid-based sampling strategy first derives SAM's class-agnostic masks, then injects domain knowledge through uncertainty-weighted votes from specialized domain models, and finally employs an iterative sampler to close coverage gaps for road infrastructure object classes, boosting their performances on the research dataset. Target sampling accelerates segmentation of high-priority infrastructure elements by pairing SAM with a multi-modal foundation model that converts engineered class descriptions into informative box

and point prompts, eliminating post-hoc voting and achieving zero-shot transferability. Thirdly, this dissertation integrates the proposed vision pipelines into a unified probabilistic framework that quantifies recognition uncertainty of speed limit signs in street-level images, fuses these observations with connected vehicle data, models sampling- and penetration-rate variability, and estimates link-level posterior distributions of speed compliance risk at network scale.

In conclusion, this research underscores the imperative for automated safety assessment that places Safe Roads at the center of the Safe System paradigm, recasts that need as a road environment perception challenge, and formalizes the challenge as a tractable semantic segmentation problem. By combining data-driven priors, domain-tuned prompts, and uncertainty-aware fusion, the proposed techniques harness vision foundation models to convert ubiquitous street-level imagery into a scalable source for iRAP-aligned safety assessments. The resulting pipelines accelerate network-wide safety assessments, have the potential to lower survey costs, and open new transportation-domain research frontiers on prompt engineering, foundation model customization, and multi-modal risk analysis.

Contents

Listing of Figures

List of Tables

Acknowledgments

My Ph.D. study is a "journey to the west", full of uncertainties, challenges, discoveries, and rewards. By working through it, I learned so much, not only about knowledge and skills, but also about vision and courage. I am so proud to grow into a better version of myself: a lot of self-reflections would not have been triggered had I not embarked on this journey. Pursuing this doctorate pushed me to generate original ideas, scrutinize long-held assumptions, and uncover intellectual and personal strengths I did not I possessed. My most sincere appreciation goes to everyone who has helped and accompanied me on this odyssey of growth and discovery.

First and foremost, I want to express my gratitude to my advisor and chair of committee Dr. Yinhai Wang for his invaluable guidance and consistent support during my Ph.D. study. Prof. Wang not only offered me tremendous freedom to conduct research on my own pace and to explore a wide range of topics, but also taught me to identify and adopt forward-looking scholarly vision. I would also like to thank Dr. Jeff Ban, Dr. Edward McCormack and Dr. Qing Shen for serving on my committee and for offering me valuable guidance during my

studies and enhancing this dissertation.

My colleagues at the STAR Lab offered me friendship and collaboration, especially Dr. Zhiyong Cui, Dr. John Ash, Dr. Ruimin Ke, Dr. Ziyuan Pu, Dr. Yifan Zhuang, Dr. Chenxi Liu, Dr. Meixin Zhu, Dr. Wei Sun, Dr. Dennis Meng-Ju Tsai, Dr. Sam Ricord, Luyang Gong, Yifan Ling, Ollie Wiesner, Nutvara Jantarathaneewat. I learned so much from you in our lab meetings, research discussions, proposal writings, course lectures, teaching sessions, conference presentations, among all on-campus and off-campus activities we participated in. I will always cherish our memories together. In particular, I want to thank Dr. Yifan Zhuang for his generous mentorship—from helping me select the right GPU components when building my research desktop computer, to recommending essential deep learning papers when I felt overwhelmed by the rapidly expanding literature, to translating abstract research concepts in my mind into actionable steps, and finally to sharing invaluable insights about industry work life that helped shape my career decisions. I wish to express my equally profound gratitude to Dr. Zhiyong Cui, with whom I spent countless days cooking together while chatting about wild research ideas, with whom I wrote endless NSF/NCHRP/DoT/-DoE proposals in endless nights while scratching our heads, with whom I taught a class at our department, and from whom I learned what it truly meant to be a researcher. I wish we could publish a serious breakthrough paper together to honor this friendship.

I would also like to extend my heartfelt thanks to friends who patiently listened to my complaints even when they made little sense. Although we have not seen each other for the last six years, I always feel like you are next door for a chat, a dinner, a walk. Talking to Dr. Yuhan Bao has always been calming: thank you for all the text chats during the pandemic, for practicing Feng Shui divination for my apartment, and for introducing mindfulness practices that

promote calm and self-reflection. Talking to Dr. Jianpeng Cao has always been encouraging: thank you for demonstrating remarkable resilience on your PhD journey and sharing those stories that illuminated my own path during moments of doubt. Your experiences became both compass and comfort as I navigated similar challenges.

At last, I want to thank my family for supporting me throughout this journey. Words cannot fully express how much you mean to me. Thank you, Mr. Yin and Ms. Zhang, and the entire family—it has been a privilege and a blessing to grow up surrounded by such love. Your unwavering support and unconditional love have shaped not only my academic path but the person I have become. Finally, thank you Shiyu Li for believing in me and sharing this splendid piece of life with me. Your presence has been my constant source of strength and joy. This accomplishment belongs to both of us.

1

Introduction

1.1 Motivation

A persistent road-safety crisis continues to imperil communities across the United States, demanding scalable and innovative solutions. Despite advancements in vehicle technology and intelligent transportation systems (ITS), fatality rates have stagnated over the past decade, with a troubling reversal of previously declining trends [130]. This situation stands in stark con-

trast to peer nations that have achieved substantial reductions in traffic fatalities by adopting comprehensive safety frameworks like the Safe System Approach[193]. The disparity suggests a critical need to reevaluate how road safety is assessed and addressed in the country.

At the heart of proactive safety interventions as proposed in Safe System Approach and similar protocols lies the ability to correctly identify road environment and infrastructure and assess their contributions to crash risk[52]. These physical features of roadway environments, from pedestrian facilities and roadside objects to traffic control devices and intersection configurations, define the context that road users perceive and where their behaviors unfold. These elements can either mitigate or exacerbate the consequences of inevitable human errors, ultimately determining the outcome of a mistake.

Safety assessments start by locating roadway features listed in protocols—such as the International Road Assessment Programme (iRAP)[79]—that materialize broader conceptual frameworks into concrete checklists. Traditional approaches rely on trained personnel to review collected imagery that combine visual judgment with manual measurement tools. Therefore, such efforts are constrained by their resource-intensive nature, limited geographic coverage, and inherent subjectivity. These approaches, despite being thorough, cannot scale to comprehensively cover the vast road network across diverse geographies or to keep pace with evolving roadway infrastructure. The resulting patchwork of safety knowledge creates critical gaps in the understanding of infrastructure safety risk, ultimately hindering targeted interventions that could save lives.

The proliferation of street-level images (SLIs) presents an unprecedented opportunity to transform infrastructure safety assessment[208]. These widely available visual data sources, which are collected by mapping services, agency vehicles, and even crowd-sourced platforms,

capture the contextual richness of roadway environments at scales previously unimaginable. The visual documentation of roadway features across entire networks creates a digital record that could theoretically support comprehensive safety evaluations without the constraints of traditional resource-intensive procedures.

However, the promise of this emerging transportation big data source remains largely unfulfilled. Converting unstructured visual data into structured safety metrics requires overcoming significant technical challenges. Infrastructure elements appear in SLIs with considerable variation in perspective, lighting, occlusion, and geographic context. These elements often blend into complex backgrounds, exhibit fragmented visibility, or appear as part of intricate scenes that defy simple common approaches [141]. Conventional computer vision models, while effective and specialized in specific use cases, struggle with their poor capacity to generalize beyond their training datasets, not to mention the nuanced requirements of infrastructure safety assessment protocols.

Recent breakthroughs in computer vision, particularly the emergence of vision foundation models with remarkable generalization capabilities, offer new possibilities for automated infrastructure assessment. These models are trained on visual and multi-modal data of unprecedented scales and diversity [101,116] and are found to demonstrate zero-shot recognition abilities that could potentially overcome barriers that have limited previous approaches. When properly adapted to road safety assessment contexts, these models could enable consistent, objective, and comprehensive evaluation of safety features across diverse roadway environments. The automated perception of roadway features from SLIs can greatly enhance existing safety databases by populating previously unavailable features without intensive manual collection. This enables a more complete picture of infrastructure performance, allowing

practitioners to model safety outcomes with precision and target interventions where they will have maximum impact.

1.2 Background

1.2.1 Road safety assessment and iRAP Star Ratings

The road safety situation in the United States is alarmingly serious. Over the past decade, the country has experienced a tragic loss of lives due to transportation incidents[130]. The majority of these deaths have occurred on streets, roads, and highways, reflecting a widespread issue in roadway traffic. The year 2021 was particularly alarming, recording around 43 thousand fatalities in motor vehicle crashes, which included a considerable number of pedestrian deaths. This uptick in roadway fatalities signifies a distressing shift from a previously consistent 30-year trend of decreasing fatality rates, with an especially sharp increase observed in 2020 and 2021. What is worse, more recent data from National Highway Traffic Safety Administration (NHTSA) show that even when the pandemic was ending, this number of roadway fatalities remained high in 2022[129]. Another concerning aspect of the situation is the stagnation in the improvement of the fatality rate per hundred million vehicle miles traveled, which has shown little improvement over the last decade. These statistics underscore the urgent need for effective and comprehensive strategies to address and mitigate this road safety crisis. The troubling trend in road fatalities and the disparities in the impact of these incidents highlight the complexity of the issue and the necessity for multifaceted solutions[130]. Improving road safety is imperative to prevent further loss of life and to ensure safer travel for all road users.

Safe System Approach is such a framework that aims to eliminate fatal and serious injuries for all road users[52]. This conceptual framework has been adopted and proved effective in

countries such as Norway, France, Sweden, Netherlands, and Australia by achieving 30-70% reductions in traffic fatalities over the last 20 years[193]. While these countries now have the lowest rates of roadway traffic fatalities per population worldwide, fatalities in the US were only reduced by 5.6% during the same period. A core idea of Safe System Approach is to refocus the design and operation of transportation systems to proactively account for human errors and reduce impact forces to tolerable levels when crashes do occur[52]. Its vision includes five elements: safe speeds, safe roads, safe road users, safe vehicles, and post-crash care. In particular, *safe roads* seeks to implement features appropriate for the intended and actual road use as well as the speed environment to reduce the likelihood of error and the consequences of error; *safe speeds* is based on the fact that lower vehicle speed means reduced impact forces, provides additional time for drivers to react and stop; and *safe road users* considers all road users equally regardless of their travel modes and urges all users to comply with rules and thus act within the limits of the road system's design. Among them, *safe roads* is the foundation of road safety as they provide the physical environment that constrains and guides driving behavior, setting the boundaries within which all road users must operate. Unlike other elements, road infrastructure represents a relatively fixed constraint that directly shapes risk exposure and influences both crash likelihood and severity outcomes. The design qualities of the roadway—including alignment, cross-section, intersection configuration, and roadside objects—establish the baseline safety conditions that other interventions must work within. This element is also where transportation authorities can act most proactively upon through planning, design, and implementation, while other elements all involve user behavior that is inherently more variable and challenging to consistently influence through policy interventions alone.

In order to achieve Safe System and modal priority decision-making as well as performance tracking, it is critical to collect low-cost data and set up safety performance metrics. The International Road Assessment Programme (iRAP) Star Rating provides such a protocol that facilitates this objective through a structured assessment of roadway infrastructure. It is widely acknowledged and applied internationally, e.g., two of the Voluntary Global Road Safety Performance Targets agreed by United Nations (UN) Member States are to ensure all new roads are built to a 3-star or better standard for all road users (Target 3), and that 75% of all travel is conducted on the equivalent of 3-star or better roads for all road users by 2030 (Target 4)[176]. UN estimates that 450,000 lives will be saved every year if these targets are achieved in practice. The Star Ratings protocol quantitatively evaluates road safety risks through systematic coding of approximately 50 distinct road attributes at 100-meter intervals. This evidence-based approach generates safety ratings from 1 to 5 stars for four road user categories: vehicle occupants, motorcyclists, pedestrians, and bicyclists, with 5 stars representing optimal safety conditions. iRAP variables can be grouped into six categories: roadside objects (hazard proximity, rumble strips, shoulders); midblock design (carriageway configuration, lane configuration, width, curvature, median type, skid resistance); intersection configurations (geometry, control mechanisms, turning provisions); vulnerable road user facilities and land use (crossing infrastructure, sidewalks, motorcycle and bicycle paths); speed management (speed limits, differentials, management measures); and flow environment (annual average daily traffic, motorcycle/truck percentage, bike flow). These coded attributes are processed through algorithms based on crash modification factors to calculate risk scores that reflect both crash likelihood and potential severity. The derived Star Ratings provide quantifiable metrics for identifying high-risk road segments, prioritizing safety investments,

and tracking infrastructure safety performance over time. Despite its robustness, traditional iRAP coding practices rely on manual annotation. Human annotators are only able to process only 15-25km per day[79] and require extensive quality management protocols to ensure consistency. The process demands specialized training, technical infrastructure for viewing geo-referenced imagery, and careful management of coder fatigue, collectively creating substantial resource demands and scalability constraints. These limitations are particularly problematic in resource-constrained environments where road safety assessment is often most urgently needed. The considerable resources required for human annotation—including training costs, quality assurance processes, and specialized equipment—have stimulated interest in developing automated approaches that maintain analytical rigor while reducing resource requirements and human variability in the coding process, thereby enhancing the accessibility and scalability of road safety assessment.

1.2.2 Emerging transportation big data

Emerging transportation big data offer unique potentials and reliable inputs for safety assessments by their variety, volume, value, velocity, and veracity. Firstly, the variety of emerging big data sources enables a vast pool of features, where each source offers unique angles to describe road infrastructure and user activities. Satellite and aerial images provide comprehensive overhead views of the Earth's surface. This capability allows for easy measurement of distances and is particularly useful for capturing geometric information and land use patterns at scale. These images are instrumental in modeling roadway design[191], evaluating roadway markings[23], and collecting pedestrian infrastructure data[76]. On the other hand, street-level images (SLIs), captured by mobile mapping systems, offer a detailed, localized view from the

ground level. They provide rich contextual information essential for understanding and assessing accessibility issues, signage visibility, and interactions among road users[144,121]. Thus, while satellite and aerial images capture extensive geometric and land use data in a single snapshot, SLIs deliver granular, localized traffic and contextual information. Another example of how emerging transportation data sources complement each other is seen between SLIs and mobile LiDAR. SLIs are accessible and reliable for scene understanding, yet they often lack the attributes necessary to derive accurate spatial information. For instance, calculating size of objects and distances between an object and the camera in images generally requires camera parameters. In large crowd-sourced datasets, however, access to such data is not guaranteed. Sophisticated methods are thus needed, e.g., (1) leveraging known reference objects like standard traffic signs and vehicles whose dimensions are known[166], (2) analyzing vanishing points where parallel lines like roads converge[103], and (3) performing feature matching across multiple photos taken from different angles[50]. Yet, these methods are based on assumptions that may not always hold in all scenarios. On the contrary, mobile LiDAR collects three-dimensional point clouds by measuring distances with laser light. This three-dimensional data allows for measurements of distances, elevations, and other critical dimensions[22], and help differentiate between objects that might appear similar in two-dimensional images, such as poles and tree trunks, by providing depth information[170]. Moreover, LiDAR systems do not rely on ambient light and can operate in low-light conditions or at night, unlike cameras which require sufficient lighting to capture useful images[26]. This makes LiDAR a valuable tool for continuous data collection across different times of day and in varying weather conditions. Due to these unique capabilities, LiDAR has been applied to studying pavement condition[47,19], assessing sight distance[92,98], extracting road geometry features[161,113], and iden-

tifying infrastructure objects[77,114], among others.

Secondly, the coverage of emerging big data sources enables a full understanding of roadway dynamics over various regions and time periods, while also capturing a wider range of population activities. Traffic sensors, with improved robustness and reduced sizes, are mounted on mobile sources, enabling data collection anywhere these vehicles can travel[25,164]. This mobility overcomes the traditional limitation of sensors being fixed to installed locations. An example is the Location-based Service (LBS) data collection, which is facilitated by GPS sensors built into mobile devices. Such datasets, collected on scales from cities to entire countries, surpass the spatial limitations of traditional traffic datasets, which are typically confined to highways and major arterials[88,137]. For instance, despite extensive studies and adoption, loop detectors are restricted to instrumented locations on the highway network due to installation and maintenance costs[43,99]. The expanded spatial coverage from newer technologies offers a comprehensive view of traffic patterns, human mobility, and transportation trends across entire regions, not just at spot locations on major roads. Additionally, advancements in technology provide extended temporal coverage through devices like cameras that continuously stream and record traffic data[57,13], eliminating the need for fixed analysis windows typically used to define peak hours[173]. This allows for analysis of varying patterns at different time of day and day of the week, which may arise from distinct causal factors and mechanisms[74,205]. Moreover, the variety of user activities recorded by emerging data sources provides a contrast to active data collection, which targets specific behaviors. Sources like LBS and video cameras passively gather data with little bias, allowing for effective extraction of targeted events post-collection. This approach reduces the marginal cost of capturing rare events, such as near-misses used in traffic safety analysis. These near-miss events, which involve sudden evasive

maneuvers to avoid conflicts and can escalate into collisions, are not included in traditional data sources like crash reports but are increasingly identified in new datasets and found valuable in safety analysis. In LBS data, near-misses are detected by analyzing speeds between consecutive trajectory points against pre-defined thresholds[80]. In video data, they are identified through metrics like time-to-collision (TTC), typically compared with thresholds, typically between 1.5 seconds to 5 seconds[96].

Thirdly, fine spatiotemporal granularity of emerging big data sources reveals real-time factors critical for safety analysis. These dynamic factors are highly sought after by statistical models but always missing in traditional roadway inventory datasets, which typically only have Average Annual Daily Traffic (AADT) surveyed once several years[206]. Telemetics and LBS data are widely used to characterize traffic patterns and profiles[160,221], recording vehicle status variables and GPS pings at sub-minute frequencies via smartphones[62,160] or even better by vehicular onboard sensors[48]. Associated spatial error radius can also be controlled to several meters in most situations. This brings three key benefits for creating traffic dynamic feature creation: (1) enhanced spatial resolution, which reduces errors in matching trajectory points to roadway networks and improves the quality of features aggregated by segment or intersection; (2) high-resolution status data, allowing for flexible aggregation with window size ranging from minutes to hours; (3) the ability to generate new indicator variables such as acceleration, which is approximated linearly by dividing the difference in speed over time interval, enabling the identification of near-miss events such as hard braking[80]; and yaw rate, which measures the lateral acceleration of the vehicle in degrees per second and is typically collected by specialized sensors[18], can be approximated as the linear rate of change in heading and applied to a few types of vehicular conflict modeling[81]. Thus, this granularity opens up

a pool of safety variables and they are of high quality. Compared to traditional traffic measures from loop detectors, AADT, and features recorded in crash reports[206,33], these newly available safety variables offer deeper insights into traffic conditions[154], closely aligned with real-time events. Similarly, video camera data also capture features of high temporal resolutions, whether collected via by fixed sensors[13] or on mobile sources[96]. Capturing dozens of frames per second, these video data provide detailed user trajectories that enhance causation understanding at intersections[13,20], on selected highway segments[104,58], and across entire networks[26].

Fourthly, emerging big data sources utilize automatic data collection methods, significantly reducing human errors and enhancing accuracy. Traditional datasets often rely on human annotation and may suffer from measurement errors. For example, crash reports could miss environmental details and inaccurately locate incidents; AADT estimates might derive from inaccurate traffic counts in non-representative scenarios of the traffic flow; and errors in roadway geometry measurements could result from incorrect sensor usage. Collecting the *real* ground truth with traditional datasets is nearly impossible. In contrast, by automating data collection, emerging big data sources address this issue by capturing data with minimal bias consistently. While they may still be skewed due to penetration rates and sensor errors, these sources comprehensively record all data for which the sensors are designed, enabling robust downstream data mining for various purposes. A notable example is travel surveys. Traditionally, these involve sampling individuals or households across a study region, who then record demographic information, vehicles, and member activities; members ages 5+ are asked to report their travel for a 24-hour travel period, with these periods evenly distributed throughout the study time frame. Despite extensive quality controls, how-

ever, this survey procedure is still prone to issues such as nonresponse[61] and misreporting[185], which can significantly impact the summarized insights critical for policy-making, such as during the COVID-19 pandemic for assessing lockdown compliance. Therefore, LBS data have been used to track mobility patterns and curfew compliance[126], offering more accurate, timely, and scalable survey data.

Finally, emerging big data sources bring a critical advantage to safety analysis by capturing longitudinal behaviors of individual road users, a dimension traditional methods fail to track. Traditional datasets are limited to isolated snapshots without continuity: (1) a single crash record includes the vehicles involved in the crash and these records are aggregated per segment or intersection in crash databases for analysis purposes; (2) a single vehicle passage triggers a disruption at an inductive loop, which then increments the count by one and records the speed, and loop detector database aggregates these information to every 20s or 30s to present volume, occupancy, as average speed; (3) a pedestrian pushes the push-button to cross a street, and the dataset records this event with its associated timestamp. When the pedestrian is impatient or unsure if the previous push is successful, multiple attempts may be made and recorded[111]. These isolated data points cannot trace users' movements before or after incidents, hindering the extraction of deeper knowledge, e.g., who are driving on the freeway network and if different demographic groups are represented equally and how a vehicle gets caught in a crash and if this conflict can be predicted from the trajectory before the event[167]. Emerging big data sources can capture user-specific trajectories, enabling the construction of comprehensive driving profiles and movement trajectories[15,6]. This rich information supports diverse applications such as commercial fleet management[100,62,160], traffic violation analysis[152], near-miss detection[80], and behavioral studies like jaywalking on high-

ways[123], work commute[138], travel mode choices[203]. Video cameras also records longitudinal data by capturing continuous user behaviors within their field of view, which facilitates sophisticated analyses like object detection[87] and movement tracking[192]. This combination of technologies offers unprecedented insights into road safety dynamics, significantly enhancing the predictive capabilities of safety models and interventions.

Compared to traditional datasets, emerging big data sources offer significant advantages: (1) they display great variety enabling many to complement each other effectively; (2) they provide extensive coverage across all road functional classes, over larger geographical areas, various time periods, and diverse population activities; (3) they are collected at finer spatiotemporal granularity describing dynamic traffic factors at another level; (4) they leverage automatic data collection procedures that enhance accuracy; (5) they capture longitudinal behaviors of individual road users, facilitating the generation of in-depth knowledge. Consequently, these capabilities enable the proactive identification and management of risks in line with the Safe System Approach, fostering safer traffic environments.

1.2.3 Road environment perception

Safe Roads is the foundation for other Safe System elements. For *safe speeds*, road design directly influences driver speed selection through visual cues, geometric constraints, and traffic calming features. Well-designed roadways encourage appropriate speeds without relying solely on enforcement. For *safe road users*, properly designed infrastructure reduces decision complexity, provides clear guidance, and creates forgiving environments that accommodate human error. Signs, markings, and geometric design all work together to make correct behavior intuitive. For *safe vehicles*, the road environment determines the effectiveness of vehicle

safety systems. Advanced driver assistance systems and crash avoidance technologies can only function optimally when operating within well-designed roadway environments with clear delineation and predictable geometries. For *post-crash care*, road design impacts emergency response through access provisions, shoulder width, and clear zones that facilitate rapid arrival of first responders and safe extraction of crash victims. When done correctly, *safe roads* establishes an environment where human errors are less likely to occur and less likely to result in serious injuries or fatalities when they do occur, making it the fundamental element upon which all other aspects of the Safe System Approach depend.

To achieve *safe roads* and automated safety assessment for iRAP, the technical challenge of road environment perception must be addressed, which involves identifying and categorizing physical infrastructure components in the driving environment using visual data sources. This task involves the systematic detection, localization, and classification of critical roadway assets across diverse urban and rural environments. It encompasses the identification of diverse elements including traffic signs, signals, lane markings, guardrails, pedestrian facilities, roadside barriers, drainage structures, and other components that collectively constitute the built environment of roadways. The recognition translates raw imagery into structured data catalogs that ideally can capture the type, location, condition, and context of each infrastructure element, such as required by iRAP[79]. The resulting digital inventory provides a comprehensive baseline for monitoring infrastructure changes over time and supporting predictive analytics for lifecycle management. By formulating infrastructure recognition as a computer vision problem, we establish the technical foundation for developing scalable, automated approaches to infrastructure inventory and condition assessment that can replace labor-intensive manual surveys.

To serve safety assessment protocols like iRAP, Street-level images (SLIs) have become the standard data sources adopted by transportation authorities and recommended by established rating practices; however, these visual assets have significant potential beyond their current use in manual annotation. These images refer to systematically collected geo-referenced photographs taken along road networks that capture the view of roadway environments from a perspective similar to human vision [208]. Unlike video-based data sources, SLI data consists of systematically collected discrete images that offer several distinct advantages for infrastructure analysis: (1) accessibility and cost-effectiveness: The unit costs of collecting and storing image data continue to decrease, democratizing access to this valuable resource. Many SLI sources are now available through public platforms or transportation agency archives without requiring specialized collection equipment. For example, Washington State's State Route Image Collection provides roadway imagery for the entire state highway network [187]. (2) comprehensive geographic coverage: SLI datasets span entire jurisdictions or regions, providing complete network visibility beyond high-priority corridors. This extensive coverage enables standardized analysis across diverse urban and rural contexts, eliminating the sampling biases inherent in selective data collection approaches. (3) temporal freshness: Many locations are captured repeatedly over time, creating a multi-temporal database that enables analysis of infrastructure condition trajectories and environmental changes. This temporal dimension allows for monitoring deterioration, evaluating maintenance interventions, and establishing baseline conditions for before-and-after comparisons. (4) human perspective view: SLIs capture roadway elements from a perspective similar to how road users experience them, making it valuable for understanding how infrastructure is perceived in its operational context. This perspective alignment facilitates more intuitive interpretation of safety concerns

and user experience factors than would be possible with overhead imagery alone. (5) standardized acquisition: Collection protocols ensure consistent image quality, positioning, and metadata across large geographic areas, facilitating automated analysis at scale. This standardization reduces preprocessing requirements and improves the generality of recognition models across different regions. (6) contextual richness: Images capture the broader roadway context, including surrounding land use and operational characteristics that influence safety performance. This contextual information provides valuable indicators for socioeconomic status and human dynamics within cities, enabling multidisciplinary applications beyond pure infrastructure management[209]. (7) multi-source integration potential: SLIs can be combined with other data sources like LiDAR and aerial imagery to create more complex models. Integrating SLIs with aerial photographs creates complementary perspectives that, when further enhanced with LiDAR point clouds and 3D geometry models, produces a more complete representation of infrastructure elements in their spatial context.

Recognizing roadway infrastructure elements from SLIs has evolved significantly with recent advances in computer vision and deep learning. Traditional computer vision techniques that preceded deep learning models relied heavily on handcrafted features and classical machine learning classifiers for infrastructure recognition. The cornerstone of traditional approaches was the manual design of feature descriptors tailored to capture visual characteristics of roadway infrastructure. For example, Histogram of Oriented Gradients (HOG) feature descriptor countes occurrences of gradient orientations in localized portions of an image. For infrastructure applications, HOG excels at capturing shape-based elements like traffic signs by quantifying edge directions and intensities. However, HOG struggles with complex backgrounds and lighting variations common in roadside environments[14]. Scale-Invariant

Feature Transform (SIFT) generates distinctive local features invariant to image scale, rotation, and partially invariant to illumination changes. This makes SIFT valuable for identifying consistent infrastructure elements across different viewpoints and conditions. SIFT keypoints are particularly useful for recognizing textured infrastructure with distinctive patterns, though the computational complexity limited real-time applications. These extracted features were then fed into traditional machine learning algorithms for classification or prediction. Support Vector Machines (SVMs) are widely used to classify infrastructure elements based on extracted features, finding optimal hyperplanes to separate different classes (e.g., different types of traffic signs). While effective for well-defined categories, SVMs struggled with the high intra-class variance present in aged or non-standard infrastructure[29]. Despite limited success with carefully calibrated systems for specific use cases, these traditional methods faced significant challenges during application, making them inadequate for large-scale, automated infrastructure recognition across diverse transportation networks, particularly when inventorying aging or non-standard elements. For example, deep domain expertise is typically needed to design features that would capture the essential visual characteristics of different infrastructure elements. This process was laborious and often required iterative refinement for each specific type of infrastructure. What is worse, this process is required for every combination of environmental factors due to poor generalization and sensitivity to visual variability: even within the same county, Seattle's complex interchange bridge structure is drastically different from North Bend's mountain roads (as in Figure 1.1. Similarly, when weather drifts away from controlled conditions similar to these models' training examples or when real-world infrastructure exhibits enormous visual variability due to aging, damage, maintenance practices, regional design variations, and environmental factors, performances

(a) SLI on I-5 freeway in Seattle

(b) SLI on SR202 in North Bend

Figure 1.1: Amount of details and level of complexities in SLIs can be drastically different for pre-DL models: (a) Seattle's I-5 freeway exhibiting a complex underpass with height restriction signage and curving roadway, (b) North Bend's SR202 displaying a simpler suburban road with clear lane markings and mountain backdrop.

of traditional models can degrade dramatically due to these novel situations, such as unusual lighting, weather conditions, or non-standard infrastructure variants. It is impossible to plan for all potential combinations and thus handcraft features for all of them beforehand. Another issue with these traditional methods is modular processing, where feature extraction and classification are treated as separate stages. This prevents end-to-end optimization and limits the ability to learn complex relationships between visual attributes and infrastructure categories.

Deep learning (DL) models overcome the drawbacks of traditional approaches by automatically learning hierarchical representations directly from data without requiring handcrafted features. This paradigm shift provides several critical advantages for roadway infrastructure recognition: (1) end-to-end learning: DL models jointly optimize feature extraction and classification, eliminating the need for manual feature engineering while discovering more discriminative representations[34]. (2) hierarchical feature learning: Networks automatically learn representations at multiple levels of abstraction, from low-level edges and textures to high-level semantic concepts, capturing the complex visual patterns of diverse infrastruc-

ture elements[67,55]. (3) contextual understanding: Deep architectures incorporate spatial context across different receptive fields, enabling more robust recognition even when infrastructure elements are partially occluded or appear in challenging environments[69]. (4) invariance to visual variations: Through data augmentation and architectural innovations like convolutional operations, DL models develop robust features that maintain performance across viewpoint changes, lighting conditions, and scale variations[135]. (5) transferrability: Pre-trained models can transfer knowledge from large-scale datasets to specific infrastructure recognition tasks, significantly reducing the amount of domain-specific labeled data required. Within our context of roadway infrastructure recognition from SLIs, DL models can be categorized into two groups with distinct capabilities and applications. Object detection models identify and localize discrete infrastructure elements within images using bounding boxes paired with class predictions. They have been used widely for (1) detecting and classifying countable infrastructure elements such as traffic signs and cars, (2) providing spatial localization through coordinate-based bounding boxes. Region-based CNNs (R-CNN family) features two-stage detectors that generate region proposals first and then classify each region[66,147]. They typically offer higher accuracy for infrastructure elements with variable sizes and aspect ratios, such as traffic signs at different distances[68]. Single Shot Detectors (SSD) features one-stage architectures like SSD[117] and YOLO (You Only Look Once)[146] prioritize speed by making predictions in a single forward pass, enabling efficient processing of street-level imagery for rapid infrastructure assessment. In particular, YOLO revolutionized object detection by reframing it as a regression problem, simultaneously predicting bounding boxes and class probabilities in a single network pass. Classic iterations progressively improved the architecture with feature pyramid networks, advanced backbones, and anchor refinements.

They excel at real-time processing of SLIs, enabling mobile or vehicle-mounted systems to identify traffic signs, signals, and roadside safety elements at highway speeds[146]. More recent versions incorporate transformer elements, enhanced feature aggregation, and advanced training techniques. They demonstrate superior performance in detecting small objects. Feature Pyramid Networks (FPN): These architectures explicitly handle multi-scale feature representation, improving detection of infrastructure elements that appear at vastly different scales in street-level imagery. Despite their success, object detection models do not apply well to our goal because they (1) only provide approximate localization through rectangular bounding boxes; (2) cannot distinguish between touching or overlapping infrastructure elements; (3) present limited ability to analyze the shape, condition, or precise boundaries of elements; (4) do not support area-based measurements or continuous feature analysis. Semantic segmentation models, on the other hand, classify every pixel in an image, providing detailed delineation of infrastructure element boundaries. Because of their nature, this group of models fit better for roadway infrastructure element recognition, because they precisely map the spatial extent and shape of both discrete and continuous infrastructure elements. This facilitates comprehensive scene understanding for complex roadway environments because many infrastructure objects appear in non-traditional shapes in SLIs. Fully Convolutional Networks (FCNs) pioneered end-to-end segmentation by replacing fully connected layers with convolutional layers, enabling pixel-wise prediction across entire images[229]. This family of models introduced the critical concept of skip connections that combine deep semantic information with shallow spatial information, allowing the recovery of fine details for precise boundary delineation. Their architecture transformed image segmentation by demonstrating that classification networks could be adapted for dense prediction tasks while maintaining

spatial information throughout the network[120]. Encoder-Decoder Networks (U-Nets) employ symmetric pathways with skip connections to preserve spatial details, which is crucial for precisely delineating object boundaries in complex scenes. The encoder path progressively reduces spatial resolution while capturing context, while the decoder path recovers the original resolution through upsampling operations, creating a U-shaped architecture that balances feature extraction with spatial precision. Initially developed for biomedical image segmentation, U-Net's balanced design has proven remarkably versatile across domains where boundary precision directly impacts downstream applications[71]. Deeplab family is another series of models that progressively addressed fundamental challenges in dense prediction through innovative techniques for multi-scale processing and contextual reasoning. DeepLabv1 introduced atrous (dilated) convolutions to expand receptive fields without downsampling, while DeepLabv2 added atrous spatial pyramid pooling (ASPP) to capture objects at multiple scales with parallel dilated convolutions[37]. DeepLabv3 further refined the ASPP module with batch normalization and image-level features, while DeepLabv3+ incorporated an effective decoder module for recovering object boundaries with higher precision[39]. Deeplabv3+ is widely used as benchmark for newer models. Transformer-based models incorporate attention mechanisms to focus on the most relevant parts of an image, improving recognition of fine details and modeling complex spatial relationships. SegFormer combines efficient transformer encoders with lightweight multilayer perceptron decoders, eliminating the need for positional encodings while achieving excellent performance with significantly reduced computational complexity[195]. MaskFormer introduced a mask classification approach that unifies semantic and instance segmentation by predicting a set of binary masks alongside corresponding class predictions[42]. Mask2Former further advanced this paradigm with a

Transformer decoder containing masked attention, enabling precise mask prediction across semantic, instance, and panoptic segmentation tasks within a single architecture[41]. These transformer-based approaches represent a fundamental shift from pixel-level classification to mask-level prediction, demonstrating superior performance in capturing long-range dependencies and handling complex scene compositions. Semantic segmentation models generally have higher computational demands than object detection approaches because probabilities across all candidate classes need to be calculated for each pixel; require pixel-level annotations, which are more labor-intensive to create; do not distinguish between individual instances of the same class. To summarize, the major differences between object detection models and semantic segmentation models include (1) their different output format, with object detection producing bounding boxes and class labels for discrete elements and semantic segmentation generating dense pixel-wise classifications for the entire scene; (2) localization precision of objects, with object detection providing approximate localization while semantic segmentation offering pixel-perfect boundary delineation in the ideal case; (3) annotation requirements differ, with object detection requiring annotation per bounding box, more efficient to create than the pixel-level masks needed for segmentation.

Foundation models represent the next generation, characterized by their unprecedented scale and diversity of training data. Unlike conventional deep learning models trained on domain-specific datasets, foundation models are trained on vast, diverse collections of datasets spanning multiple domains to develop general-purpose representations that capture universal visual concepts and relationships. This paradigm shift from narrowly focused supervised learning to broad self-supervised or weakly-supervised pretraining enables these models to develop emergent capabilities that transfer remarkably well across tasks with minimal adap-

tation. Models like CLIP (Contrastive Language-Image Pre-training)[143] learn aligned representations across vision and language inputs through contrastive learning on over 400 million image-text pairs, enabling zero-shot recognition capabilities through natural language supervision. SAM (Segment Anything Model)[102], trained on over 11 million annotated images, demonstrates remarkable generalization to unseen segmentation tasks through its prompt-based architecture, allowing users to specify objects of interest using points, boxes, or text. DINOv2[134] and other self-supervised models learn rich visual representations without explicit labels by solving pretext tasks across vast image collections, developing emergent capabilities that transfer robustly across visual domains. A trend of foundation models is to incorporate multi-modal information, designing multiple perception channels to simultaneously process and align representations across modalities like vision, language, audio, and even 3D data. For example, vision-language models like BLIP-2[109] combines image understanding with language comprehension, allowing for more nuanced interpretation of visual content through natural language reasoning. This cross-modal fusion enables capabilities such as visual question answering, image-grounded dialogue, and detailed image captioning that captures semantic nuance beyond what either modality could achieve independently. One biggest benefit of multi-modality is transferability across domains and tasks. Models trained on diverse data collections develop representations that generalize effectively to downstream applications with minimal finetuning. For instance, GPT-4V[5] extend large language model capabilities to visual inputs, enabling systems to reason across modalities and perform complex tasks like visual programming, document understanding, and multi-step reasoning about visual scenarios.

Recognizing roadway infrastructure via automated methods and through the lens of emerg-

ing data sources, especially SLIs, has its unique challenges. First of all, SLIs depict complex scenes of urban streets or highways, which could include varying architectural styles, multitudes of vehicular and pedestrian traffic, as well as roadside elements like signage and landscaping. The lighting in SLIs can vary dramatically within the same scene due to natural factors such as the time of day and weather conditions. This variation affects how objects are illuminated and leads to unclear boundaries of them. Shadows, glare, and reflections can obscure the object edges, making it challenging to distinguish one object from another, especially in crowded urban settings. In other domains such as medical research and remote sensing, this is of a lesser problem as images are either taken in controlled environments or the position of camera eliminates the issue[21,211]. Secondly, infrastructure elements are typically positioned in the background within SLIs, while dynamic road users occupy the foreground. This is concerning as mainstream research and applications in traffic sensing and autonomous driving prioritize the immediate surroundings of the ego-vehicle. This includes objects that directly affect the vehicle's path and immediate decisions, such as pedestrians crossing the street or vehicles merging into the lane and signaling a turn. As a result, critical infrastructure elements in the background has received less attention in the model development, as demonstrated by the taxonomy of roadway infrastructure being far simpler and undeveloped than road user classes in open-sourced datasets[182,26,45]. This spatial arrangement might also lead to a depth perception challenge where distant but significant infrastructure objects are underrepresented. Thirdly, unlike road users whose shape and appearance follow patterns, infrastructure elements can exhibit much more diverse and challenging patterns in shapes, orientations, colors, textures, sizes, among others. One critical difference worth mentioning is road users typically are countable objects, i.e., *things*, that have an expected size

within SLIs; but infrastructure elements could be *things*, such as traffic signs or trash bins, or *stuff* that can occupy large portions of the scene, such as sidewalk and lane markings. A guardrail for example could be at a similar position within SLIs as a bench and they even have similar structural patterns, but they belong to entirely different categories and have entirely different safety implications. Objects belonging to the same taxonomy class could also exhibit entirely different patterns by geography. For example, some freeway streets lights are mounted on huge towers rather than short poles, trying to light a larger area and appear in disc shapes. Positioned in the background of complicated scenes and varied in shapes and appearances, infrastructure elements often appear fragmented and thus likely exhibit multiple objectives of the target class, disqualifying many DL models that are designed for instance and panoptic segmentation. Lastly, taxonomy of infrastructure elements in open-sourced datasets (if present) can confuse models, which may necessitate the use of advanced global contextual modeling. For example, several pieces of white lane marking could indicate broken white line that separates lanes moving in the same direction, or they could also indicate crosswalks at intersections. Deciding if they belong to the same class and accurately producing their classes might refer to the detection of other elements like traffic signals and might need global understanding of the directions of these white lane marking pieces.

In summary, our discussion of infrastructure element recognition has presented its role for safety assessment, data, models, and challenges. First of all, it determines the success of automated safety assessment by translating unstructured visual data into structured catalog detection results. SLIs emerge as an optimal data source, featuring comprehensive coverage, human-perspective views, and high spatial-temporal resolution. DL models have been trained and tested on domain datasets but still suffer from genelizability, while foundation

models offer a promising frontier to transfer knowledge acquired from massive-scale pretraining to this application domain. The challenges of recognizing infrastructure elements in SLIs is unique, particularly the presence and equal importance of both countable "things", e.g., *traffic signs*, and uncountable "stuff", e.g., *sidewalk*. Based on these challenges, semantic segmentation fits our task better given its design to produce pixel-level labels.

1.3 Research gap

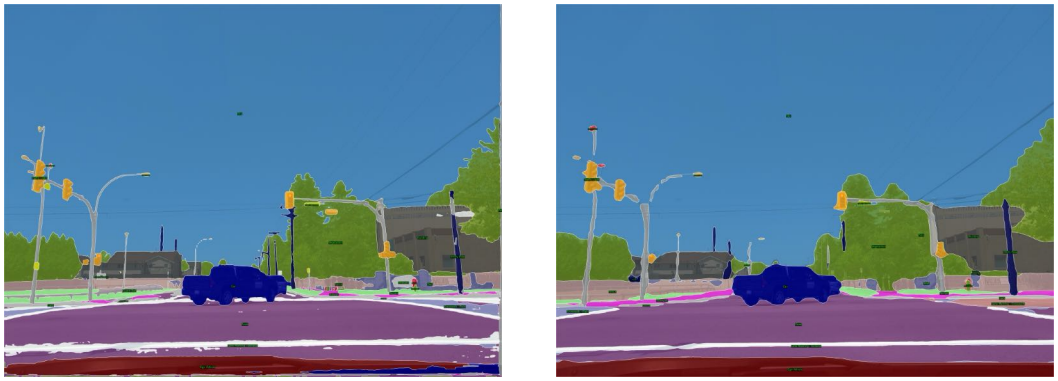
The promise of applying computer vision to facilitate automated road safety assessment has long been hindered by several limitations. First, object detection models are more widely understood and easier to implement but face inherent constraints—they can only detect countable objects like signal lights, traffic signs, and vehicles. This approach fails to capture crucial continuous elements such as road surfaces, lane markings, and roadside sidewalks that significantly impact safety ratings in iRAP assessments. Second, semantic segmentation models offer comprehensive analysis but remain unsatisfying for deployment. Despite improving performance metrics in controlled research environments, these domain-specific models reveal alarming fragility when confronted with the infinite variability of global road infrastructure. This limited capacity to generalize has kept semantic segmentation technologies from being applied to practical road safety frameworks like iRAP. For example, as shown in Figure 1.2 comparing outputs of two domain segmentation models DeepLabv3+ and Mask2Former, they produce inconsistent segmentation of critical safety elements—traffic lights appear with varying shapes and boundaries, road surfaces lack distinction between safe and hazardous areas, and even vehicle boundaries are imprecisely defined. The most serious shortcoming lies in their ambiguous segmentation of infrastructure elements such as guard rail on the side of

street, lane markings for crosswalks, small traffic signs, and poles in the distance. This is why these domain segmentation models have not be trusted for safety-critical applications where better identification of infrastructure elements are expected as they contribute directly to safety risk assessment.

The following unique challenges have hindered domain models from reaching satisfying performance levels. First of all, SLIs depict complex scenes of urban streets or highways, which could include varying architectural styles, multitudes of vehicular and pedestrian traffic, as well as roadside elements like signage and landscaping. The lighting in SLIs can vary dramatically within the same scene due to natural factors such as the time of day and weather conditions. This variation affects how objects are illuminated and leads to unclear boundaries of them. Shadows, glare, and reflections can obscure the object edges, making it challenging to distinguish one object from another, especially in crowded urban settings. In other domains such as medical research and remote sensing, this is of a lesser problem as images are either taken in controlled environments or the position of camera eliminates the issue[21,73,211]. Secondly, infrastructure elements are typically positioned in the background within SLIs, while dynamic road users occupy the foreground. This is concerning as mainstream research and applications in traffic sensing and autonomous driving prioritize the immediate surroundings of the ego vehicle. This includes objects that directly affect the vehicle's path and immediate decisions, such as pedestrians crossing the street or vehicles merging into the lane and signaling a turn. As a result, critical infrastructure elements in the background has received less attention in the model development, as demonstrated by the taxonomy of roadway infrastructure being far simpler and undeveloped than road user classes in open-sourced datasets[182,26,45]. This spatial arrangement might also lead to a depth perception challenge

where distant but significant infrastructure objects are underrepresented. Thirdly, unlike road users for which prior knowledge of shape and appearance typically exists[115], infrastructure elements can exhibit much more diverse and challenging patterns in shapes, orientations, colors, textures, sizes, among others. One critical difference worth mentioning is road users typically are countable objects, i.e., *things*, that have an expected size within SLIs; but infrastructure elements could be *things*, such as traffic signs or trash bins, or *stuff* that can occupy large porpotions of the scene, such as sidewalk and lane markings. A guardrail for example could be at a similar position within SLIs as a bench and they even have similar structural patterns, but they belong to entirely different categories and have entirely different safety implications. Objects belonging to the same taxonomy class could also exhibit entirely different patterns by geography. For example, some freeway street lights are mounted on huge towers rather than short poles, trying to light a larger area and appear in disc shapes. Positioned in the background of complicated scenes and varied in shapes and appearances, infrastructure elements often appear fragmented and thus likely exhibit multiple objectives of the target class, disqualifying many DL models that are designed for classical instance and panoptic segmentation. Lastly, taxonomy of infrastructure elements in open-sourced datasets (if present) can confuse AI/ML, which may necessitate the use of advanced global contextual modeling. For example, several pieces of white lane marking could indicate broken white line that separates lanes moving in the same direction, or they could also indicate crosswalks at intersections. Deciding if they belong to the same class and accurately producing their classes might refer to the detection of other elements like traffic signals and might need global understanding of the directions of these white lane marking pieces.

The emergence of foundation segmentation models like Segment Anything Model (SAM)[101]



(a) Deeplabv3+ **(b)** Mask2former

Figure 1.2: Domain models prioritize different aspects of semantic segmentation quality: (a) Deeplabv3+ extracts better boundaries for objects while (b) Mask2former produce more coherent masks. They agree on most objects in the scene but disagree on road infrastructure elements, e.g., guardrail, pole, etc. This is a major limitation preventing them from being applied to automated safety assessment.

has opened a revolutionary frontier in computer vision, offering unprecedented generality across visual domains that traditional models cannot match. These models demonstrate remarkable capability to identify and segment any object in an image with precision that adapts to diverse environments—functioning effectively regardless of geographic region, lighting conditions, or infrastructure styles. For road safety professionals long frustrated by the limited performance of domain-specific models, these foundation models represent not merely an incremental improvement but potentially a paradigm shift in assessment capabilities. However, two gaps limit the direct application of these foundation models to road safety assessment contexts. Firstly, while SAM demonstrates extraordinary ability at producing precise segmentation masks, these masks remain class-agnostic, i.e., the model only delineates object boundaries without possessing any domain knowledge or understanding their significance to road safety frameworks. For example, Figure 1.3 shows how SAM produces masks in a zero-shot fashion for six SLIs in a widely researched dataset[131]. While SAM does not possess any

knowledge of the domain or this dataset, the masks it produces are of high quality, capturing each individual object that appears in the scene: tire, window, door, lane marking, traffic sign, street light, pole, manhole, building, guardrail, curb cut, and grass lawn. Boundaries of SAM masks are in general better than those of the domain model. This creates a critical disconnect between segmentation and interpretation: a guardrail, a pedestrian crossing, and roadside vegetation might all be perfectly segmented with pixel-perfect precision, but without classification into domain-specific taxonomies that are connected to iRAP assessment protocols, these masks remain essentially unusable for practical safety evaluation. The segmentation itself, while technically impressive, lacks the semantic meaning required for safety rating protocols. Secondly, despite their remarkable capabilities, foundation models operate on a prompt-driven paradigm that requires strategic guidance to focus their attention effectively, particularly when applied to specialized and complicated domains like infrastructure element recognition in SLIs. This prompting dependency creates a new research challenge that remains largely unexplored: what forms of prompts are the most effective? what are necessary components of a successful template for optimal prompts? The absence of established prompting templates for infrastructure assessment creates a gap between foundation model potential and practical deployment for road safety applications.

1.4 Outline

The goal of this dissertation is to customize vision foundation models for road environment perception in the context of roadway safety assessment by automating the recognition of road infrastructure elements. The broad context of road safety assessment and Road Assessment Programme (RAP) protocols is discussed and their challenges are comprehensively summa-

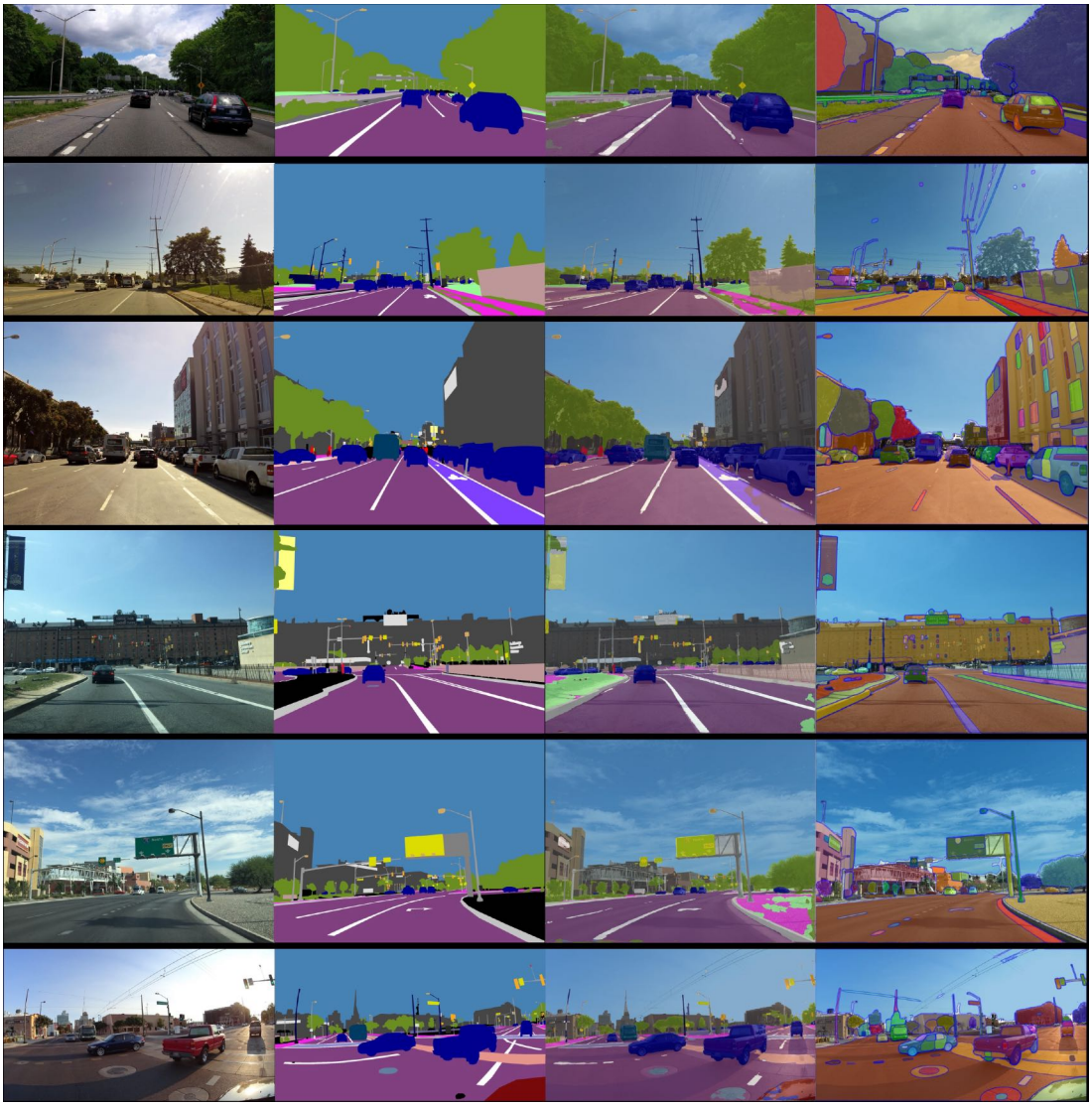


Figure 1.3: Examples of domain segmentation model (Deeplabv3+) vs. SAM. The four columns from left to right are (a) original SLIs, (b) ground truth labels, (c) domain model results, and (d) SAM results. Notice SAM masks are of high quality even though it has not seen this dataset in its training. The domain model has seen these data during training.

rized. Two approaches adapting vision foundation models to this context are proposed, experimented on surrogate research datasetes, and thoroughly analyzed. This dissertation further demonstrates the practical value of proposed pipelines by integrating them with external

emerging data sources to create more robust analytical frameworks for elements in Safe System Approach. While this work does not fully bridge the gap between these vision foundation models and fully automated iRAP coding, it proves a significant step forward by demonstrating that vision foundation models can be effectively customized to this specialized task. The contributions and and organization of following chapters are as follows.

Chapter 1 describes the United States' road-safety crisis and shows that current frameworks such as the iRAP Star Rating still rely on labor-intensive, image-by-image annotation of roadway elements to support the Safe System approach. Because this manual process is slow, costly, and hard to scale, the chapter argues that applying vision foundation models to large-scale street level imagery dataset offers a paradigm shift toward fully automated infrastructure inventories. It pinpoints two open research gaps of this adaptation to road environment perception domain that must be solved before these models can serve road-environment perception, thereby setting the agenda for the rest of the dissertation.

Chapter 2 reviews literature of existing efforts to automate iRAP Star Ratings, summarizes automated road environment perception from SLIs, and presents classic DL and recent vision foundation models along with their evolution. Gaps are identified in the current approaches for automating roadway assessment with computer vision models, for addressing unique challenges of infra element recognition from SLIs, and for tailoring vision foundation models to road safety applications.

Chapter 3 integrates foundation segmentation models with domain priors in grid sampling mode, which corresponds to SAM's "everything mode" and is designed for novice annotators unfamiliar with the dataset. In this configuration, the system automatically generates masks for all visible road infrastructure elements without requiring specific prompts, en-

abling annotators to quickly familiarize themselves with the full range of detectable features. Results demonstrate that incorporating domain-specific knowledge priors can significantly enhance classification accuracy by up to 22% for masks produced by foundation models of certain infrastructure element classes. Additionally, iterative sampling techniques, which refine initial detections through successive passes, are found to be effective in further performance improvement, particularly for small or partially occluded infrastructure elements critical to safety assessments.

Chapter 4 introduces an efficient way to incorporate multi-modal input for the pipeline to boost adaptive sampling, enabling to SAM's "anything mode" and is designed for experienced annotators focusing on specific categories that are important yet easy to overlook. This approach simulates a future scenario where annotators could generate text descriptions of infrastructure elements through typing or speech and feed these natural language prompts directly into the proposed pipeline. Experimental results reveal that introducing this second foundation model to process multi-modal input and thus the second modality overcomes two key limitations of grid sampling with domain priors: first, it significantly improves mask coverage which can be a limiting factor for small and distant objects in "everything mode"; second, it effectively lifts the dependency on domain priors and thus avoids their false confidence from training. Further analysis identifies the most efficient form of prompts for SAM, showing interaction between various modes of input. The last section of this chapter presents a mini case study, where stakeholders requested the identification of a specific subset of road infrastructure elements along its highway network. The proposed pipeline demonstrated both effectiveness and efficiency in meeting the intended objectives, while maintaining extensibility for similar applications, without retraining/finetuning the DL models included.

Chapter 5 develops a framework that fuses connected vehicle data (CVD) with automatically parsed SLIs to create a probabilistic, link-level picture of speed compliance. A processing pipeline cleans and map-matches CVD trajectories, filters for free-flow conditions, and bootstraps percentile speeds while a computer vision module detects speed limit signs (or variable speed states) and assigns confidence scores. These two streams are wrapped in a Bayesian Monte-Carlo layer that propagates fleet penetration, sampling rate, and sign detection uncertainties to produce full posterior distributions of the compliance gap and the exceedance probability. The resulting risk layer supports context-sensitive speed limit setting, feeds into iRAP Star Rating protocols, and flags segments where fresh imagery or field audits would most improve decision confidence. By coupling dynamic operating speeds with visually derived roadway context under an explicit uncertainty model, the chapter advances Safe System analytics from static point estimates to continuously updating, credibility-banded safety metrics.

Chapter 6 synthesizes the dissertation's key findings, showing how vision foundation models can be customized without task-specific finetuning for road environment perception, a domain task of safety assessment. It highlights the zero-shot capabilities of the proposed pipelines, maps each capability to iRAP use cases, and situates the contributions within current Safe System practice. The chapter concludes by identifying research frontiers such as uncertainty-aware multi-sensor fusion, continual model adaptation, and policy feedback loops that integrate real-time infrastructure and behavioral data.

2

Literature review

2.1 Overview

This chapter reviews literature in three aspects, roughly following the components and their order in Section 1.2. The first section reviews iRAP Star Rating protocol by introducing its variables related to *safe roads* and thus infrastructure elements, as well as the state-of-the-art research and practices attempting to automate this procedure. The second section reviews

SLIs as an emerging transportation big data source and general research looking into semantic segmentation using SLIs. Distinction between semantic segmentation and object detection, as well as distinction between semantic segmentation and other segmentation schemes are more formally discussed. Unique challenges introduced in Section 1.3 will reveal themselves during the review of relevant SLI datasets and applications. The last section of the chapter emphasizes the review of semantic segmentation models, from pre-DL models, to Deep Learning (DL) domain models, and finally vision foundation models. In particular, Segment Anything Model (SAM)[101] is described in greater detail, as well as existing works that adapt SAM to their corresponding domains, even though they are not safety-, infrastructure-, or even civil-related.

2.2 iRAP Star Ratings

The International Road Assessment Programme (iRAP) Star Rating protocol has emerged as a globally recognized methodology for evaluating road infrastructure safety. This section presents a comprehensive list of the protocol's variables before narrowing down on the variables related to *safe roads* and infrastructure elements, which form the foundation of systematic road safety assessment. Due to the amount of literature contributing to road infrastructure object segmentation/detection, e.g., traffic signs, bridges, and barriers, with various kinds of emerging transportation big data sources, e.g., LiDAR, SLIs, GPS, and satellite images, we limit the discussion to research directly mentioning iRAP process.

2.2.1 Current rating practices

iRAP Star Rating[199] involves evaluating approximately 50 different road attributes across six major categories: roadside severity, mid-block design, intersection configuration, vulnerable road user (VRU) facilities and land use, speed management, and flow environment. Each attribute is coded according to specific criteria to objectively assess safety risks and identify potential countermeasures. This systematic approach to road safety assessment enables the identification of high-risk road sections and the prioritization of cost-effective countermeasures to reduce fatalities and serious injuries. Annotation in iRAP, referred to as "coding", is a manual process where trained personnel assess road safety by analyzing geo-referenced images for each 100-meter road segment. This process requires coders to analyze detailed imager, record specific attributes for each segment, apply standardized criteria with defined risk categories, identify worst case scenarios within segments, and make precise measurements of features such as lane width and roadside hazard distances. Human biases and errors requires rigorous quality standards including multiple levels of review (peer, supervisor, independent) and systematic data validation before final acceptance.

Several factors make large-scale implementation of manual iRAP coding difficult. The resource intensity is significant, with a single experienced coder only processing 15-25km per day, meaning a 3,000km road network requires a team of 4-6 coders working for at least one month. The expertise and training requirements are substantial, with coders needing road engineering backgrounds, specialized training in the iRAP procedure, and consistent interpretation of standards across coders requiring ongoing supervision. Operational constraints further limit scalability, with fatigue management requiring coding shifts, the need for dedicated workspace with specialized equipments, and additional time required for collaborative

review sessions. This resource intensive nature of the process makes it expensive to deploy worldwide, particularly in developing countries which often have the highest road fatality rates but may lack resources for full implementation. The need for regular reassessment of road networks adds to the long-term cost, making it difficult to achieve comprehensive coverage in many regions. These scalability challenges highlight the need for more efficient approaches to road safety assessment that maintain the comprehensiveness and objectivity of the iRAP methodology while reducing the resource intensity of manual annotation.

The Roadside category focuses on the severity of potential hazards adjacent to the roadway, considering both the type of roadside object and its distance from the lanes of travel. Objects are ranked from highest risk, e.g., cliffs, trees, and rigid structures, to lowest risk, e.g., safety barriers, with distance measured in four bands, i.e., 0-1m, 1-5m, 5-10m, and ≥10m. This systematic approach recognizes that the risk of roadside objects diminishes with increased distance from the travel lane.

Mid-block attributes constitute the largest category, mainly describing the physical characteristics of road segments between intersections. Key variables include carriageway type, median type, road condition, delineation quality, sight distance, lane configurations, and curvature. For example, the quality of a curve is assessed based on whether signage and markings effectively help drivers judge curvature and adjust their speed accordingly. Adequate delineation is critical at night, as it enables drivers to anticipate changes in road alignment and conditions.

Intersection configuration involves the type, quality, channelization, and volume of intersecting roads. The methodology acknowledges the complexity of intersections, particularly in urban environments, by evaluating how many points of potential conflict exist between

traffic flows. Intersection quality is determined by how well drivers can anticipate intersection conditions, with factors such as advance warning, signing, markings, and sight distance all being considered.

The Vulnerable Road User (VRU) facilities and Land Use category addresses infrastructure provisions for pedestrians, cyclists, and motorcyclists, recognizing these road users' heightened vulnerability. This includes specialized facilities such as pedestrian crossings, sidewalks, bicycle lanes, and motorcycle paths. The coding also captures surrounding land use patterns (educational, commercial, residential, etc.) as these influence the likely presence and behavior of vulnerable road users, especially pedestrians.

Speed management attributes record both posted speed limits and estimated operating speeds, acknowledging that the two often differ. This category also considers speed differences between vehicle types and the presence of speed management features that might reduce operating speeds below posted limits. The focus on speed reflects its fundamental role in crash outcomes and injury severity.

Lastly, the Flow category quantifies the presence of various road users, including motorcyclists, bicyclists, and pedestrians, providing essential data for predicting crash likelihood. In addition to common measures adopted for vehicular traffic like annual average daily traffic (AADT), this category requires estimate of motorcycle percentage, as well as bicycle peak hour flow and pedestrian flow crossing the street and along street on both driver and passenger side of traffic.

Contribution of these variables to safety Star Ratings is characterized by several important interactions that extend beyond simple presence or absence. Roadside elements demonstrate this complexity clearly: the same object, e.g., a tree, presents dramatically different risk

based on its distance from the traveled lane. This creates a hierarchical risk matrix with 62 distinct roadside severity combinations (see page 50 in the manual[79]), where a cliff at any distance represents higher risk than a tree at 0-1 meters away from traffic, which in turn exceeds the risk of a rigid pole at 5-10 meters away from traffic. This distance-object interaction enables nuanced safety assessment of diverse roadside environments. Infrastructure elements also exhibit quality-based interactions. For curve assessment, the physical curvature (very sharp to gentle) interacts with delineation quality—a sharp curve with poor signing and marking presents significantly higher risk than the same curve with adequate delineation. Similarly, pedestrian crossing facilities are evaluated not just on design type (marked/signalized/grade-separated) but on their quality of implementation, refuge presence, and visibility. The methodology further captures synergistic interactions between complementary elements. For example, median barriers' effectiveness interacts with their design (concrete/metal/wire-rope), width of separation, and motorcycle-friendliness. Sidewalk safety increases with both separation distance from traffic (0-1m, 1-3m, >3m) and the presence of physical barriers, creating a spectrum of pedestrian protection. Speed management infrastructure demonstrates the most complex interactions, as its effectiveness is assessed through impact on operating speeds relative to posted limits, creating a dynamic relationship between physical elements and driver behavior. This approach recognizes that road environment influences actual road user behavior rather than existing in isolation. The iRAP star rating methodology ultimately evaluates road safety through these nuanced interactions, recognizing that infrastructure elements work as systems rather than isolated features, with their protective value highly dependent on specific implementation characteristics and their relationship to surrounding elements.

Table 2.1: Variables defined in iRAP Star Rating protocol[79] organized into six categories.

Category	Variable	Classes	Comments/Explanations
Roadside	Roadside severity - driver/passenger side distance	0-1m, 1-5m, 5-10m, >=10m	Roadside objects and their distances to the road together define the risk. Only record the roadside object of the highest risk in the 100-meter segment. Non-standard median barriers might be coded as roadside object.
	Roadside severity - driver/passenger side object	Cliff, Tree, Non-frangible sign/post/pole, Unprotected safety barrier end, Aggressive vertical face, Upwards slope (roll over/no roll over), Deep drainage ditch, Downwards slope, Large boulders, Semi-rigid/Rigid structure/bridge/bldg, Safety barrier (concrete/metal/wire rope/motorcycle friendly), no object	
	Shoulder rumble strips	Present, Not Present	Any textured markings running along a road which function to warn drivers leaving the lane on the passenger-side of the roadway. Only be found on paved shoulders and should be a continuous, unbroken lin.
	Paved shoulder - driver/passenger side	None, Narrow (0-1m), Medium (1-2.4m), Wide (>=2.4m)	Width of the safe and drivable section of road from the edge line to the edge of the paving. No edge line, no paved shoulder should be recorded. Wherever an edge line is present, record the paved shoulder.
Mid-block	Roadway/Carriageway type	Carriageway A/B of a divided road, undivided road, motorcycle facility	Divided carriageway: (1) either a barrier or wide physical median consistently; (2) for a distance of 400m or more. One-way, bus/transit lane
	Upgrade cost	High, Medium, Low	
	Median type	Center Line, Wide Center Line (0.3-1m, >1m), Continuous central turning lane, Flexible posts, Physical median width (0-1m, 1-5m, 5-10m, 10-20m, >20m), Safety barrier (concrete, metal, motorcycle friendly, wire rope), One-way	Road infrastructure feature that separates the two opposing traffic flows for both divided and undivided carriageways. Do not include defective safety barriers. Safety barrier - motorcycle friendly has supporting post or legs.

Continued on next page

Table 2.1 – continued from previous page

Category	Variable	Classes	Comments/Explanations
	Road condition	Poor, Medium, Good	Ability to provide level or running surface free from defects that may adversely affect vehicle path. 'Poor' or 'medium' if defects are present for 10m or more at any point on a coding segment.
	Delineation	Poor, Adequate	Adequacy of road lines and markings (center lines, lane markers, edge lines; guideposts and delineators, road studs and hazard markers; signage on road and posted). "Poor" for generally absent or in poor condition.
	Sight distance	Poor, Adequate	Ability of a driver to see and/or anticipate road conditions and other road users ahead. "Poor" if SD<50m for speed<70km/h or SD<80m for speed>70km/h
	Number of lanes	Four or more lanes, Three lanes, Three and two lanes, Two lanes, Two and one lanes, One lane	Number of traffic lanes in the direction of travel. Do not code changes over short lengths of road less than 400m or turning lanes at intersections.
	Lane width	Narrow <2.75m, Medium 2.75m to <3.25m, Wide ≥3.25m	Distance from the edge line to the adjacent lane marking. If no edge line, measure from pavement edge to adjacent lane marking.
	Curvature	Very sharp, Sharp, Moderate, Straight or gently curving	Horizontal alignment of the road. Very sharp: can only be driven at <40km/h, radius <200m. Sharp: 40-70km/h, radius 200-400m. Moderate: 70-100km/h, radius 400-800m. Straight: >100km/h, radius >800m.
	Quality of curve	Poor, Adequate, Not applicable	How easy it is to judge how sharp a curve is and if it can be driven safely. Quality reflects extent to which signs and markings help the driver judge curvature.
	Upgrade cost	High, Medium, Low	Influence that the surrounding land-use, environment and topography will have on the cost of major works. Based on the relative cost of acquiring additional land and complexity of working on that land.
	Skid resistance	Unsealed-poor, Unsealed-adequate, Sealed-poor, Sealed-medium, Sealed-adequate	Skidding resistance and texture depth of the road surface.

Continued on next page

Table 2.1 – continued from previous page

Category	Variable	Classes	Comments/Explanations
	Centre line rumble strips	Not Present, Present	Any textured markings running along the centre of a road which function to warn drivers crossing the median.
	Shoulder rumble strips	Not Present, Present	Any textured markings running along a road which function to warn drivers leaving the lane on the passenger-side of the roadway.
	Paved shoulder - driver/passenger side	None, Narrow om to <1m, Medium 1m to <2.4m, Wide ≥2.4m	Width of the safe and drivable section of road from the edge line to the edge of the paving. Code each side of the road separately.
	Presence of service road (frontage road)	Not present, Present	Presence of a service road running parallel with the main carriageway.
	Vehicle parking	Two sides, One side, None	Record the extent of vehicle parking along the side of the road.
	Street lighting	Not Present, Present	Record the presence of street lighting that is sufficient to illuminate pedestrians and bicyclists.
Intersection	Intersection type	3/4-leg - unsignalized/signalized - with/without protected turn lane, Mini roundabout, Roundabout, Merge lane, Median crossing point (informal/formal), Railway Crossing active/passive, None	Presence and type of intersections. If intersection is very complex, consider how many points of potential conflict exist between flows and the highest risk points.
	Quality of intersection	Poor, Adequate, N/A	Quality of the intersection design, advance warning, signing, and markings. "Poor" could be very short merge lanes, poor deflection angles, lack of advance signing and marking, lack of facilities for VRUs, limited or obstructed sight distance.
	Intersection channelization	Not present, Present	If there are raised islands or coloured hatching present at an intersection that designate intended vehicle paths.
	Intersecting road volume	≥15,000 vehicles, 10,000 to 15,000 vehicles, 5,000 to 10,000 vehicles, 1,000 to 5,000 vehicles, 100 to 1,000 vehicles, 1 to 100 vehicles, Not applicable	An estimate of the AADT of the intersecting road. For railway crossings an estimate of the number of trains running through the crossing per day is required.

Continued on next page

Table 2.1 – continued from previous page

Category	Variable	Classes	Comments/Explanations
	Property access points	Commercial Access 1+, Residential Access 3+, Residential Access 1 to 2, None	Number of commercial and residential driveways and minor access lanes.
VRU facilities and Land Use	Pedestrian crossing facility - major/side road	None, Unsignalized (raised) unmarked/marked crossing with/without refuge, signalized with/without refuge, grade separated facility	Presence of purpose-built pedestrian crossing facilities on the inspected road and on the side (intersecting) road. Only record crossing facilities which provide some safety benefit to pedestrians.
	Pedestrian crossing quality	Poor, Adequate, Not applicable	Effectiveness of the pedestrian crossing on the inspected road or side road.
	Pedestrian fencing	Not present, Present	Presence of pedestrian fencing or other barriers which effectively control pedestrian crossing flow.
	Sidewalks - driver/passenger side	None, Informal path (0-1m, >=1m), Non-physical separation (0-1m, 1-3m, >=3m), physical barrier	Do not code a sidewalk that is not continuous over the entire length of the coding segment. Distance is measured from nearest driving lane.
	Facilities for motorcycles	None, Motorcycle lane on roadway, Motorcycle path - one-way, Motorcycle path - two-way, Motorcycle path with barrier (one-way/two-way)	Presence of purpose-built facilities for motorcycles, mopeds and other light motorised vehicles capable of speeds of 30km/h or more.
	Bicycle facility	None, Extra wide outside (>=4.2m), On-road lane, Off-road path, Off-road path with barrier, Signed shared roadway, Shared use path	Purpose-built facilities for bicyclists. Only record facilities specifically for bicycles. Do not code bike facilities of poor standard. If shared with pedestrians, code it as shared use path.
	Land use - driver/passenger side	Educational, Commercial, Industrial and manufacturing, Residential, Farming and agricultural, Undeveloped areas	Type of roadside development that is observed on both the driver-side and passenger-side of the road.
	Area type	Urban, Rural	Type of area surrounding the road. Area type is coded as being either 'urban' or 'rural' depending on the level surrounding development.
	School zone warning	No school zone warning (school present), School zone – static signs or road markings, School zone – flashing beacons, Not applicable	Presence of a school zone.

Continued on next page

Table 2.1 – continued from previous page

Category	Variable	Classes	Comments/Explanations
	School zone crossing guard	Not present, Present, Not applicable	Presence of a crossing supervisor or warden.
Speeds	Speed limit	From 20 to 150 km/h (Codes 1-25) or 20 to 90 mph (Codes 31-45)	Actual posted numerical speed limit for general traffic and for motorcycles and trucks.
	Operating speed - 85th percentile/mean	20, 25, 30, ..., 85, 90, >=90	km/h
	Speed differential	Present, Not Present	If the difference in speed (operating or speed limit) between cars and trucks and/or cars and motorcycles exceeds 20km/h or 12mph.
	Speed management	Not present, Present	Presence of road infrastructure features that will typically reduce the operating speed by 5km/h to 10km/h below the speed limit.
Flow	AADT		
	Motorcycle percentage	None, 0, 5, 10, 20, 40, ..., 80, 99, 100	
	Bicycle peak hour flow	None, 1, 3, 5, 7, >=8	Observed number of bicycles per coding segment.
	Pedestrian peak hour flow across/along road driver/passenger side	0, 5, 25, 50, 100, 200, ..., 500, 900, >=900	Observed flows are not directly used in Star Ratings. They form one part of the evidence used during the Star Rating analysis in estimating flows on each road segment.

2.2.2 Existing automation efforts

Acknowledging the inherent limitations of iRAP's manual annotation approach, pioneering work has emerged to develop computational methods that automate the infrastructure rating workflow.

The first group of studies focuses on the extraction of subsets of iRAP-related variables using emerging big data sources such as video, LiDAR, and satellite imagery. Brkić *et al.*[22] developed a fully automated framework that integrates mobile LiDAR point clouds and YOLO-based DL models to detect and classify roadside features, enabling the assessment of two iRAP attributes: roadside severity-object (RSS–O) and roadside severity-distance (RSS–D). They transformed point clouds collected on Croatian highways into road cross-section images at 10-meter intervals, which were then used to detect road boundaries and classify 13 iRAP-defined object types and their distances using labeled training data and YOLO object detection. Their work demonstrated that LiDAR data significantly facilitates roadside object distance measurement at large scale and provides input data of good quality to downstream modules. Brkić *et al.*[24] proposed a novel framework that used high-resolution satellite imagery combined with the YOLO v5 object detector to identify iRAP-defined road attributes such as school zones, divided carriageways, and four types of pedestrian crossings. Their method involved vectorizing road center lines, segmenting road images, transforming them into road-oriented views, and training YOLO models on annotated segments to detect the relevant infrastructure elements across 83.5 km of roads in Split, Croatia. The model achieved strong performance with high class-wise accuracy ranging as well as high precision/recall, outperforming earlier approaches while enabling detailed class-level differentiation, e.g., inspected vs. side-road pedestrian crossings, with or without refugee islands. Wijnands *et al.*[191]

presented the first systematic analysis of all intersection designs in Australia using satellite imagery and deep learning to identify safer road infrastructure. The researchers processed approximately 900,000 satellite images of Australian intersections using customized computer vision techniques and a deep autoencoder to extract high-level features like intersection type, size, shape, lane markings, and complexity, which were then clustered to identify similar designs. They linked these intersection designs to driving behaviors captured in telematics data from 66 million kilometers of driving, revealing that four-way intersections had more frequent hard acceleration events than three-way intersections, T-intersections had relatively low hard deceleration frequencies, and roundabouts consistently maintained low average speeds. This domain-specific feature extraction approach enabled the identification of specific infrastructure improvements that could result in safer driving behaviors, potentially reducing road trauma. Tardy *et al.* [171] developed an automated approach for road inventory using a deep learning model and 3D point cloud data acquired by a low-cost mobile mapping system (MMS). The methodology extracts key road parameters including road width, number of lanes, individual lane widths, superelevation, and safety barrier height. The research also used a deep learning model based on Point Transformer architecture to perform semantic segmentation of road environments into five classes: asphalt, road markings, road signs, barriers, and other elements. Results showed an overall intersection-over-union score of 84% for point cloud segmentation and centimetric errors for road inventory parameters, with the number of lanes correctly estimated in 81% of cases. This approach offers a safer and more automated alternative to traditional manual road inventory methods and has potential for building digital models from as-built infrastructure acquired by mobile mapping systems.

The second group aims for full end-to-end automation of the entire safety rating process,

directly predicting star ratings or complete attribute sets from raw input data. The standard iRAP process involves three transparent steps: identifying the presence of specific infrastructure attributes for each road segment, assigning risk ratings according to a defined protocol, and aggregating these ratings through a weighted scoring system. However, the following studies bypassed this structure entirely, collapsing the pipeline into a single prediction step and turning the process into a black box—undermining transparency and interpretability, as well as obscuring alignment with the established rating framework. Song *et al.* [165] proposed a deep convolutional neural network architecture that directly predicts roadway star ratings from street-level panoramas using multi-task learning, attention mechanisms, and unsupervised training. The proposed model took panoramas as input and used a modified VGG-16 backbone with spatial attention layers to estimate star ratings and 23 auxiliary road attributes, e.g., median type, road condition, number of lanes. The training incorporated supervised, multitask, and unsupervised loss terms, including a loss encouraging consistency across nearby panoramas. Despite reasonable results reported, the shortcoming of this research is clear. Although each variable has a dedicated channel and the loss functions are customized for multi-tasking, it still suffers from tremendous lack of domain knowledge, e.g., classifying upgrade cost and intersecting road volume from SLIs. It was not surprising to see the results in terms of accuracy were not convincing either. Overlooking the physical definitions and the relationships between variables resulted in the bad prediction power and interpretability. The most concerning issue was the poor performance on 1-star rated roads (the most concerning category), which should be the highest priority for a safety assessment system. Additionally, the authors failed to address how their system would handle temporal variations in road conditions (like construction or seasonal changes) that could significantly

impact safety ratings. Furthermore, the safety-aware routing application lacked comprehensive evaluation against established transportation safety benchmarks or metrics such as potential crash reduction factors, travel time reliability, or actual historical crash data correlation: instead relying solely on visual examples of alternative routes with subjective annotations of hazardous features. Finally, the modest improvement over baseline methods (just 3.2% higher accuracy) raises questions about the value of such a complex system for practical deployment in transportation infrastructure planning. Kačan *et al.*[93] explored multi-task deep learning framework for large-scale iRAP attribute coding. By proposing an end-to-end multi-task learning model that classifies 33 iRAP safety attributes simultaneously from monocular dashboard video, the authors sought to reduce the need for semantic segmentation. Their architecture used shared CNN backbones (ResNet-18 and DenseNet-121), spatial pyramid pooling, and per-attribute attention modules, trained with balanced cross-entropy loss and optionally extended to handle multi-frame inputs for temporal context. Trained on a large 1850 km dataset from Bosnia and Herzegovina, their best model achieved a macro-F1 score of 61.5% and showed high accuracy for well-represented attributes like "delineation" and "divided carriageway", while struggling with rare or visually ambiguous features like "roadside severity." While the authors proposed an interesting approach for automating road safety assessment, their methodology has several limitations that diminish its practical utility. The approach relies heavily on a dataset from a single country with limited geographical diversity, raising serious concerns about generalizability to different road systems, designs, and environmental conditions worldwide. Additionally, the model's poor performance on several safety-critical attributes (particularly roadside severity with only 39.7-47.9% macro-F1 scores) makes it unsuitable for real-world deployment where these features significantly im-

pact safety ratings. The paper also lacks proper comparison with traditional computer vision techniques like object detection and semantic segmentation pipelines that might offer more interpretable and verifiable results for stakeholders. Finally, although the authors acknowledged their work was a "preliminary feasibility study", they failed to address how the substantial performance gap between well-performing and poorly-performing attributes could be bridged in future work.

2.3 Automated recognition with SLIs

SLIs such as those provided by mapping platforms, have emerged as a prominent data source for road infrastructure analysis and safety assessment. Its widespread adoption is largely attributed to extensive geographic coverage, frequent updates in urban areas, and relatively low cost compared to other data collection methods. Furthermore, SLI provides rich visual context at human-eye level, making it particularly suitable for identifying roadside features relevant to frameworks like iRAP. However, the use of SLIs also introduces several challenges. Variability in lighting conditions, occlusions from vehicles or vegetation, differences in camera angles, and inconsistencies in image resolution can significantly impact model performance and limit the reliability of automated feature extraction. These limitations necessitate careful pre-processing and robust modeling to ensure accurate interpretation of SLIs for domain tasks.

2.3.1 General applications

SLIs have been generally applied to transportation engineering and urban planning. For example, one major application is understanding urban environments. DL models extract high-

level semantic information from SLIs and deliver them to downstream urban planning tasks. Wang, Li, and Rajagopal [186] proposed a unsupervised multi-modal framework to understand intrinsic patterns of urban communities. By jointly encoding visual, textual, and geospatial information into the neighborhood representation vector, their model can compare to fully-supervised methods in downstream prediction tasks. Extensive experiments on three U.S. metropolitan areas from the study demonstrated the model's interpretability, generalization capability, and its value in neighborhood similarity analysis like house pricing prediction. Alhasoun and González [11] proposed an approach to collect and label imagery data then deploy advancements in computer vision to automate street category labeling, e.g., alley, commercial throughway, downtown residential, downtown commercial, highway, highway ramp, industrial, neighborhood commercial, neighborhood residential, park, residential throughway. convolutional networks extracted feature embeddings of the SLIs and a classification model used these embeddings to produce the street category. Research in [85,17,17,95,65] focused on urban environmental sensing, where SLIs are fed to segmentation models which guesses the class each pixel in the SLI belongs to. Then the ratio of pixels belonging to different classes are aggregated into numeric indices for walkability, biking, and street safety. Cai et al. [27] went one step further, as they not only performed semantic segmentation, visual environment extraction, but also depth calculation [59] to quantify drivers' visual environment in real driving situations. The derived visual measures in this research then served as features for a frequency-based traffic crash model.

A group of studies focus on the detection and understanding of roadway infrastructure elements. Aykac et al. [16] used Roadway Information Database (RID) of SHRP2 program [163], as a surrogate for a video log to design and test algorithms to detect rumble strips in the road-

way images. The image correction algorithm along with automated feature extraction and convolutional neural network model was successful in detecting rumble strip locations and measuring their length and pitch. Abou Chacra and Zelek[4] identified distressed pavement regions and pinpointed cracks with them by using a combination of support vector machines (SVM) and image processing techniques. Their pipeline first segmented *road* pixels from the image using texture information[3] and then extract built another set of descriptors of the pixels using which the SVM classifies in to distressed areas of road. The segmentation step involves patch extraction at multiple scales. Rahman et al.[144] applied a classic DL model[68] that can capture both convolutional and recurrent information to map road safety barriers. Weld et al.[189] assessed sidewalk accessibility with SLIs. Their model use crops of SLIs, positional feature of the crop, as well as geographic feature of the SLI within the broader context of a city's geography as inputs and a ResNet architecture plus fully connected layers to process these information[69]. The final output of the DL model is a score vector of five classes: curb ramps, missing curb ramps, obstructions, surface problems, and null crop. Discussions were also conducted on the quality of the model, effect of contextual input features on performance, cross-city generality, among others. Another contribution of the study was its crowdsourcing tool allowing remote contribution of SLIs and labels via an online platform. Ning et al. proposed a DL pipeline to automatically collect sidewalk inventory at scale[133]. They used a sliding window strategy to read the input images in patches and then apply a finetuned segmentation model for the task to predict pixel-level sidewalk label, before merging the inference results back to the original image size. If there is shadow or occlusion by canopies or buildings that result in gaps in the extracted sidewalk, SLIs are queried at locations with gaps to fill them.

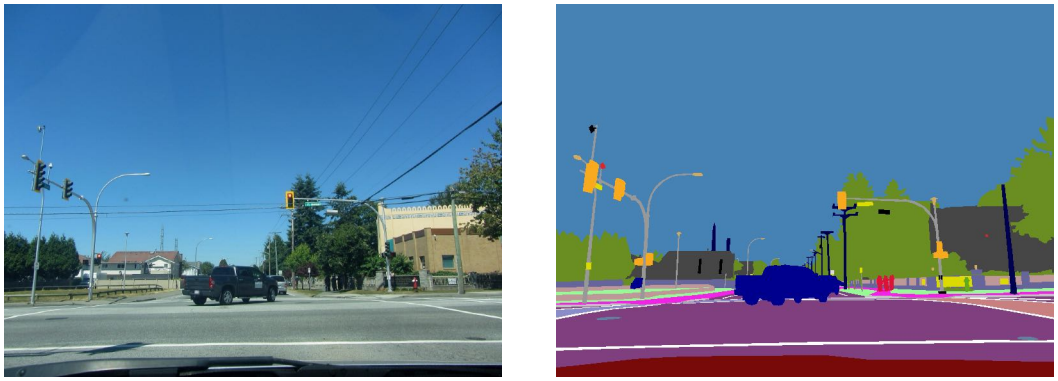
2.3.2 Semantic segmentation for infrastructure elements

Semantic segmentation and object detection are two fundamental tasks in computer vision, each serving distinct purposes in visual understanding. Object detection focuses on identifying and localizing instances of predefined object categories by predicting bounding boxes and associated class labels. While effective for tasks requiring rough spatial localization, bounding boxes do not provide precise information about object shape or boundaries. In contrast, segmentation—particularly semantic segmentation—assigns a class label to each individual pixel in an image, enabling fine-grained delineation of objects and surfaces. This pixel-level classification is essential in applications like road infrastructure assessment, where understanding the exact extent and boundaries of features such as sidewalks, medians, or bike lanes is critical. Although more computationally demanding, segmentation provides a richer and more detailed representation of scene geometry compared to bounding-box-based detection.

While semantic segmentation assigns a class label to each pixel, it does not distinguish between multiple instances of the same class—treating all pixels labeled as "car" as belonging to a single, undifferentiated group. Instance segmentation addresses this limitation by not only classifying each pixel but also assigning it to a specific object instance, effectively combining object detection and segmentation. This enables more detailed analysis in scenarios involving crowded scenes or overlapping objects. Panoptic segmentation further unifies semantic and instance segmentation by producing a single coherent output that includes both per-pixel class labels for "stuff", e.g., road, sky, sidewalk, and instance-level labels for "things", e.g., vehicles, pedestrians, traffic signs. However, due to the lack of consistent instance-level annotations and the focus on infrastructure classes that are typically amorphous and non-discrete, e.g., sidewalks, medians, shoulders, this research is limited to semantic segmentation.

Several open-sourced segmentation datasets are based on SLIs. CityScapes[45] dataset were collected in 50 European cities using vehicle-mounted sensors spanning good and medium weather conditions during daytime. It contains 5K images with fine annotations plus 20K images with coarse annotations. There are 30 object classes, grouped into eight major categories, three of which are related to roadway infrastructure: (1) *flat* includes road, sidewalk, parking, rail track; (2) *construction* includes building, wall, fence, guardrail, bridge, and tunnel; (3) *object* includes pole, pole group, traffic sign, traffic light. Images in PASCAL VOC[51] were collected from a variety of scenes. It contains 11K images including 7K segmentations. The 20 object classes besides *background* in its taxonomy include *vehicles* (from aeroplane to bus and boat), *households* (bottle, chair, dining table, potted plant, sofa, TV/-monitor), *animals* (from bird to sheep), and *other* (person). ADE20K[225] covers a wide range of scenes and environments, including both indoor and outdoor settings, such as streets, parks, rooms, offices, and public spaces. It contains 22K training images and 2K validation images. Notably, ADE20K contains detailed pixel-wise segmentation label for more than 3K concepts, organized in their WordNet definition and hierarchy[226] and about 200K parts of objects and parts of parts are annotated. For concepts related to driving/roadway/traffic, the *stuff* group includes non-countable materials or elements that form the continuous aspects of road environments and the *objects* group are countable and discrete items. For example, *stuff* group include roads (asphalt, concrete, dirt roads), sidewalks (paved, gravel), markings (lane markings, crosswalks), surfaces (pavements, kerbs), while the *object* categories include vehicles, street furniture (traffic lights, street lamps, traffic signs, benches), road barriers (guardrails, bollards), miscellaneous (potholes, manhole covers, drainage grates), and vegetation. ADE20K-outdoor is a subset of ADE20K that only contains outdoor scenes, as

used in Qi et al.[142]. Lastly, Mapillary Vistas[131] is specifically targeting SLIs related to roads and traffic, with 25K high-resolution images crowd-sourced from worldwide and 65 classes included. Concepts are organized in seven root groups include object, construction, human, marking, nature, void, and animal, where each of them is further broken down intro macro-groups and then specific classes. For example, *object-support* (root group and macro-group) include pole, utility pole, and traffic frame, while *construction-barrier* include curb, fence, wall, guardrail, and other barrier. Figure 2.1 shows an example from the dataset. The label map demonstrates high-quality semantic segmentation with precise object boundaries, particularly evident in the vehicle silhouette and traffic infrastructure elements that maintain their structural integrity despite simplification. The segmentation exhibits sophisticated class differentiation between closely related roadway elements (traffic signals, poles, utility poles) that would typically challenge less robust taxonomies. Edge preservation in the segmentation is notably accurate at critical boundaries between roadway markings and asphalt surfaces, which is essential for safety-critical applications requiring precise delineation of drivable space. The visualization reveals the dataset's comprehensive coverage of both dominant classes, e.g., road surfaces and vegetation, and safety-relevant minority classes, e.g., traffic signals and pedestrians that colored red, supporting Mapillary Vistas' suitability for roadway safety assessment without additional annotation requirements. Given this connection and the labeling quality, we treat Mapillary Vistas as a surrogate dataset for real-world iRAP data and evaluate model performances on Mapillary Vistas.



(a) Original SLIs **(b)** Ground-truth label maps

Figure 2.1: Mapillary Vistas dataset include (a) high-quality SLIs and (b) their ground truth segmentation map labels. These ground truths (panel **b**) demonstrate pixel-wise segmentation annotation with precise object boundaries. The annotation exhibits sophisticated class differentiation between closely related roadway elements, e.g., traffic signals, poles, and utility poles, that are coarsely treated in other datasets. This dataset has a comprehensive coverage of both dominant classes like cars and roads, plus safety-relevant minority classes, e.g., traffic signals, curbs, guard rails, fences, etc.

2.4 Semantic segmentation models

In this section, we skip classic models that precedented DL models, e.g., superpixels, CRFs, hand-crafted features, and start with DL models directly.

2.4.1 DL models

Domain DL models are trained on specific segmentation datasets. Fully Convolutional Network (FCN)[120] is a pioneering model in semantic segmentation because it was the first to successfully apply deep neural networks to such a dense prediction task directly at the pixel level, in an end-to-end manner without the need for any fully connected layers[127]. Prior to FCNs, most DL models focused on image classification tasks like on the ImageNet[46,106], where the output is a single label for an entire image, which clearly does not address the needs of semantic segmentation. While there were methods that adapt classification networks for lo-

calization and detection, such as the R-CNN family[56,55,147,67] for object detection, FCN was the first to modify the architecture of CNNs to produce spatial maps instead of classification scores, enabling end-to-end training for segmentation. In addition to its end-to-end learning, FCN is also featured by its encoder-decoder structure, as well as the upsampling layers and skip connections, which combined merge deep, semantic information from the lower layers with shallow, appearance information from the higher layers.

The DeepLab family is a series of models that also contributed significantly to semantic segmentation task[36,37,38,39]. A fundamental contribution of this family of models is atrous convolution, also known as dilated convolution, which adjusts the sampling of input signal by using a dilation rate that expands the spacing between kernel points and thus enhances the convolution's field of view without increasing parameters or computational complexity[36]. Then the iterations in the model family (1) adds Atrous Spatial Pyramid Pooling (ASPP) which applies the atrous convolution in parallel with different dilation rates, greatly enhancing the model's capacity to capture multi-scale information[37]; (2) improves ASPP by incorporating batch normalization and adding image-level features extracted by global average pooling to the last feature map, which captures broader context[38]; (3) introduces a complete encoder-decoder structure as the previous models of the family focused on the encoder part that deepens the network to enhance feature extraction at various scales[39]; the addition of decoder enables the model to produce more precise and high-resolution segmentation maps, especially at object boundaries, because the decoder reconstructs the spatial details that are compressed in the encoding phase.

Even more recent advancements in the field have embraced Transformers, such as SegFormer[195]. SegFormer differs from the DeepLab family and previous models in several as-

pects. First, it integrates the Transformer architecture to handle the entire feature extraction process, leveraging self-attention mechanisms that naturally capture long-range dependencies and global context—beneficial for parsing complex scenes in semantic segmentation. Second, SegFormer adopts a hierarchical Transformer backbone and uses a lightweight MLP decoder to process features at multiple scales, in contrast to earlier models with heavy decoders relying on extensive convolutional operations. As a result, SegFormer achieves both scalability and operational efficiency across a range of device capacities and image resolutions. Complementary to this trend, MaskFormer[42] and Mask2Former[41] propose a unified framework that treats segmentation as a set prediction task. By predicting a fixed set of binary masks and their associated class labels, these models enable a shared architecture across semantic, instance, and panoptic segmentation. Mask2Former further improves performance by incorporating multi-scale deformable attention, making it one of the current state-of-the-art models for segmentation tasks.

2.4.2 Foundation models

Unlike DL models trained on domain datasets typically consisting of tens of thousands of images, foundation models are trained on datasets that are up to three orders of magnitude larger. This vast scale of pretraining endows them with a strong capacity to generalize across tasks and domains—including ours—in a zero-shot fashion, without the need for additional fine-tuning.

One example of a foundation model designed for segmentation tasks is the Segment Anything Model (SAM)[101]. SAM's architecture (see Figure 2.2) is built on a vision transformer (ViT) backbone and consists of three primary components: an image encoder that processes

the input image once to create an image embedding, a prompt encoder that handles various prompt types (points, boxes, text, and masks), and a mask decoder that predicts segmentation masks based on the combined image and prompt embeddings. SAM enables interactive prompting where users can provide different types of signals (including points, boxes, or masks) with positive/negative labels to guide the segmentation process, making it exceptionally flexible for identifying and isolating objects in images without category constraints. If no specific prompts are provided, SAM defaults to an evenly spaced sampling grid and prompts the model with each sampled point. This is called "everything mode" where it attempts to segment all possible objects in the image. This prompt-driven approach allows SAM to perform real-time interactive segmentation with zero-shot capability across diverse visual contexts. The training methodology involves pre-training on the extensive SA-1B dataset, followed by fine-tuning to optimize the model's performance for specific segmentation tasks. SAM's training dataset represents an unprecedented scale in segmentation data, containing over 1 billion masks derived from approximately 11 million diverse images. This dataset embraces the "anything" premise by capturing a variety of objects and scenes without being constrained to predefined categories, ensuring comprehensive coverage across different domains. The data collection process employed a semi-automated approach where initial masks were generated through computer vision techniques, followed by human verification and refinement.

A growing body of research chooses to adapt SAM to specific application domains, aiming to extend its general-purpose capabilities to domain-specific segmentation tasks. Given there is little transportation-focused research using SAM, Table 2.2 summarizes how SAM is adapted to domains of remote sensing, medical imaging, factory production, and organizes

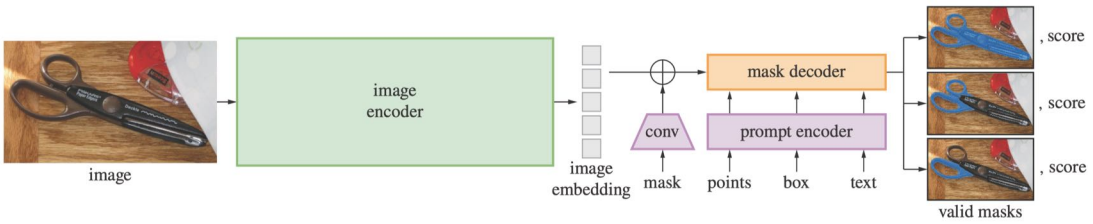


Figure 2.2: Model architecture of Segment Anything Model (SAM). It consists of three primary components: a powerful image encoder that processes the input image once to create an image embedding, a prompt encoder that handles various prompt types (points, boxes, text, and masks), and a lightweight mask decoder that predicts segmentation masks based on the combined image and prompt embeddings. Source: Kirillov *et al.* [101].

them by the type of prompts and modules/parameters finetuned. Empty cells in the image encoder column could mean (1) applied as original if no extra parameters are marked [145,64], (2) guided by strategies like LoRA [78,223,210], or (3) replaced by row-rank or compressed representations [140]. First group of research focuses on SAM's zero-shot capability, without any finetuning for the domain-specific task. These applications include evaluations of SAM on domain-specific datasets [64,70], creation of new or enhanced datasets [179,159,40,34], and other inference tasks like building change [169], product anomaly detection [110], point tracking across video frames [145]. Second group of studies aim to finetune SAM's modules. This is needed when domain target differs significantly from SAM's training data and zero-shot performances are unsatisfactory. The most common choice to finetune is mask decoder given its one of the lightweight modules in the structure [198,168]. The last group aims to improve SAM comprehensively, either by retraining it for a new domain like medical imaging [122], or by compressing its knowledge for faster and more resource-efficient performances [197,140,210,207].

Migrating the capabilities of foundation models to our application domain involves bridging the gap between general-purpose segmentation and the specific requirements of roadway safety assessment. Given that domain-specific models are typically trained on curated

SLI datasets, one promising direction is to combine their task-specific strengths with the zero-shot generalization capability of foundation models such as SAM. Several recent studies have explored this hybrid strategy, either by fusing outputs or leveraging complementary model architectures[217,40,168]. Another line of research addresses the fact that SAM produces class-agnostic masks by default, which poses challenges for multi-class semantic segmentation tasks like ours. To address this, some studies have enhanced SAM's classification capacity by repurposing its internal modules or appending additional classifiers—typically multilayer perceptrons (MLPs)—to infer semantic labels from its output masks[211,223]. Lastly, SAM's promptable architecture opens another avenue for adaptation: supplying appropriate prompts tailored to the target domain. Researchers have explored a variety of prompt generation strategies—some deriving prompts from ground truth labels[179,145,139,140], others sourcing them from domain models or upstream vision-language models[220,168,217,34,54,148,110,213], some relying on probabilistic techniques[227], and a few simulating prompts directly from test cases[207]. These strategies highlight the flexibility of prompt design as a key lever for aligning SAM with domain-specific segmentation needs.

2.5 Summary

This chapter reviews literature related to iRAP Star Ratings, semantic segmentation using SLIs, and recent advancements in both domain-specific and foundation models for segmentation tasks. iRAP's road infrastructure variables contribute to safety ratings through structured interactions that go beyond mere presence or absence; however, several state-of-the-art studies bypass this interpretability by adopting deep learning models that generate ratings in an end-to-end, black-box manner. SLIs have emerged as a valuable big data source

for transportation applications, offering wide coverage and high-resolution visual context. Prior research has explored the extraction of infrastructure elements from SLIs, though most approaches rely on object detection, which limits the ability to capture detailed boundaries and spatial extent. For tasks requiring pixel-level precision, semantic segmentation provides a more suitable solution. This chapter also highlights major SLI-driven datasets and the deep learning models trained on them. In addition, the Segment Anything Model (SAM) is introduced in detail, including its architecture, capabilities, and recent efforts to adapt it to domain-specific segmentation. Two main adaptation strategies are identified: (1) enable domain-specific classification for SAM either by adding domain classifiers to its downstream module or augmenting its structure with modules like MLPs, and (2) design effective prompt generation techniques tailored to the target task for SAM.

Table 2.2: Prompts and fine-tuning choices reported when adapting SAM to domain-specific applications.

Study	Prompt type				Fine-tuned module				Multi-class
	point	box	text	other	image encoder	mask decoder	prompt encoder	extra proposed	
213	✓	✓	✓	✓					
179		✓							
169									
159	✓								
145	✓								
9									
110	✓	✓							
70	✓	✓							
148		✓							
110	✓	✓							
219									
64									
218									
218									
211		✓						✓	✓
40									
54		✓							
34		✓							
217		✓		✓				✓	
31						✓		✓	
35						✓		✓	
198						✓			
73								✓	
168	✓		✓	✓		✓			
212						✓		✓	
188			✓			✓		✓	
204					✓	✓			
196				✓		✓		✓	
122		✓			✓	✓	✓		
197								✓	
207	✓	✓						✓	
216								✓	
140		✓						✓	
139		✓						✓	
227	✓	✓				✓	✓	✓	
210	✓	✓		✓				✓	
220	✓	✓	✓					✓	
223						✓		✓	✓

3

Grid-based sampling with domain priors

3.1 Introduction

The Segment Anything Model (SAM) offers powerful zero-shot segmentation capabilities but remains fundamentally class-agnostic, limiting its effectiveness in domain-specific applications. Previous studies have highlighted this tension between SAM's versatility and its semantic limitations when applied to specialized contexts. This chapter proposes to integrate

prior knowledge from domain models—well-trained, specialized classifiers—with SAM to address this gap (as pointed out in Section 2.5). While these domain models lack the cross-task generalizability discussed in Chapter 2, they provide critical semantic understanding within their specialized contexts. By embedding domain priors into SAM's zero-shot framework, we create a synergistic approach that combines SAM's robust segmentation capabilities with domain-aware classification for road infrastructure element recognition. The resulting hybrid methodology preserves SAM's foundational strengths while addressing the domain-specific requirements identified in the literature, offering a practical solution for applications requiring both precise segmentation boundaries and accurate semantic interpretation.

The rest of this chapter is organized as follows. Section 3.2 analyzes Mapillary Vistas[131] dataset's connection with iRAP variables, highlighting its advantages as a surrogate to real-world iRAP evaluation dataset consisting of SLIs, as well as reasons why other research datasets are not equally suitable for this road safety assessment task. Section 3.3 starts by introducing SAM's structure, types of prompts, and "everything mode", then transitions to prior knowledge collectible from domain models and ways to obtain them, and arrives at the proposed framework to integrate them. Section 3.4 describes first a list of hypotheses about SAM and the proposal, then evidence supporting them, results based on one forward pass of SAM, motivations for improvement, and finally the refinement and its results. Performances of these two variants are not only compared with baseline models, but also contrasted with an upper bound estimate. Section 3.5 briefly discusses the obtained results of this adaptation approach and shares its benefits to iRAP road safety assessment procedure.

3.2 Dataset

Mapillary Vistas dataset[131] is an appropriate surrogate for iRAP field imagery because its 66 pixel-level semantic classes align closely with the physical objects that are included in iRAP rating protocol. A systematic comparison shows that 19 of the 46 object-level variables specified in the iRAP coding manual have at least one direct proxy class in Mapillary Vistas taxonomy. Variables that are not directly covered by Mapillary Vistas taxonomy fall into three major categories, as summarized in Table 3.1. The first group comprises non-visual metrics that exist beyond the capability of static image segmentation, including temporal traffic dynamics such as flow counts across all transportation modes (pedestrians, cyclists, motorcyclists, passenger vehicles, and heavy vehicles), operating speeds for different vehicle classes, and speed differentials between these modes. These metrics are fundamentally time-dependent and cannot be derived from isolated visual snapshots regardless of resolution or processing sophistication. Such data can be recovered through a multi-layered approach: extracting and aggregating counts from permanent or temporary detectors strategically installed at intersections and mid-block locations; fusing with emerging mobility data sources such as connected vehicle data (CVD) that provides detailed telemetry on equipped vehicles; incorporating anonymized location-based services (LBS) data from mobile devices that offer broader coverage across the transportation network; and potentially supplementing with targeted manual counts for calibration and validation purposes. The second group of variables can technically be extracted from SLIs alone, but the extraction is limited by SLIs' resolutions. For example, rumble strips require high-resolution close-up views to properly identify their texture and depth characteristics; skid resistance demands specialized analysis beyond what

standard visual segmentation can reliably determine. These elements are critical for comprehensive road safety assessment but fall outside the standard Mapillary Vistas segmentation capabilities without supplementary processing techniques or higher resolution imagery focused specifically on road surface conditions. The third category encompasses higher-order geometric properties and advanced scene understanding capabilities that exceed what can be derived from isolated imagery. Roadway curvature and curvature quality require longitudinal analysis across multiple images to properly characterize the three-dimensional geometry and consistency of curves along the road path. Similarly, intersection type classification demands a comprehensive spatial understanding from various approach angles to accurately determine configuration and complexity. The determination of number of lanes requires sophisticated scene understanding that can distinguish between main carriageways, auxiliary lanes, and turning lanes while accounting for lane continuity beyond the immediate field of view. Sight distance calculation—a critical safety parameter—requires precise elevation data typically obtained through LiDAR systems to model the three-dimensional topography that affects driver visibility around curves and over crests.

Despite these gaps, adopting Mapillary Vistas in automated safety assessment research brings several advantages. First, it eliminates the prohibitive cost of manually labeling extensive iRAP video or SLI data collected by road authorities while still providing the pixel-level masks needed to compute segment-level metrics mandated by the Star Ratings protocol. Second, Mapillary Vistas's geographic and environmental diversity—spanning multiple continents, climate zones, and roadway standards—acts as a natural regularizer, improving cross-regional generalizability. Third, the public availability of Mapillary Vistas annotations enables the broader research community to reproduce and extend the coverage analy-

Table 3.1: Categorization of iRAP variables not directly covered by Mapillary Vistas and mitigation strategies.

Variable Category	Limitations with Mapillary Vistas	Mitigation Strategy
Category 1: Non-visual metrics		
Operating speed; speed differentials	Temporal dynamics not captured in static segmentation	Fuse with CVD and LBS data for velocity profiles across vehicle classes
Traffic flow counts (all modes)	Time-dependent measures absent from static imagery	Aggregate detector data at intersections and mid-blocks; supplement with manual counts
Category 2: Resolution-limited features		
Rumble strips	Insufficient resolution for texture and depth characteristics	Deploy targeted high-resolution imagery focused on road surfaces
Skid resistance / surface friction	Beyond standard visual segmentation capabilities	Implement supplementary processing techniques or specialized surface analysis
Category 3: Higher-order properties		
Curvature & curvature quality	Requires analysis across multiple sequential images	Process longitudinal SLI sequences or derive from GIS polylines
Intersection type	Demands spatial understanding from multiple angles	Integrate aerial imagery with ground-level views for comprehensive classification
Number of lanes	Needs advanced scene understanding beyond single frame	Apply sequential image analysis with lane continuity algorithms
Sight distance	Requires 3D topographical modeling	Incorporate LiDAR data to model elevation changes affecting visibility

sis presented in this dissertation. The mapping between Mapillary Vistas and iRAP survey data constitutes a reusable research artifact. By precisely documenting how each iRAP variable is derived from one or more Mapillary Vistas classes, the mapping provides a transparent connection between deep learning model outputs and the actuarial logic underpinning iRAP's risk models. Standardizing this vocabulary ensures that model predictions can be translated directly into inputs for Star Rating, facilitating auditability, reproducibility, and future methodological improvements. Table 3.2 presents a complete correspondence.

Table 3.2: Correspondence between iRAP Star Rating variables and Mapillary Vistas semantic classes.

iRAP category	iRAP variables	Corresponding items	Mapillary Vistas classes
Roadside	Roadside severity - driver/passenger side object	Cliff; Upward Slope	Mountain, Terrain, Snow, Vegetation
		Tree	Vegetation
		Non-frangible sign / post / pole	Pole, Utility Pole
		Unprotected safety barrier end	Barrier, Guardrail
		Aggressive vertical face	Wall, Barrier, Building, Tunnel
		Downwards slope	Terrain, Mountain
		Large boulders	Terrain, Mountain
		Rigid structure/ bridge/ building	Barrier, Building, Bridge, Tunnel
		Semi-rigid structure / bridge / building	Bridge, Building, Tunnel
		Safety barrier (concrete / metal / wire-rope)	Barrier, Guard Rail
	Paved shoulder - driver/passenger side	Paved shoulder	Service lane
Mid-block	Median type	Center Line, Wide Center Line, Continuous central turning lane	Lane Marking - General
		Flexible posts (median)	Pole; Fence
		Physical median	Vegetation, Curb, Road
	Number of through traffic lanes	Lane marking	Lane Marking - General
	Road condition	Holes	Pothole, Manhole
	Delineation	Markings and posts	Lane Marking - General, Lane Marking - Crosswalk, Traffic Sign Frame, Traffic Sign (Front), Traffic Sign (Back), Crosswalk - Plain
	Street lighting	Pole, light fixture	Street Light; Utility Pole
	Vehicle Parking	Parking area	Parking
Intersection	Intersection type	Traffic control signals	Traffic Light, Pole, Traffic Sign (Front)
		Protected turn lane marking	Lane Marking - General
		Roundabout, Mini-roundabout	Traffic Sign (Front)
		Median crossing point	Traffic Sign (Front), Lane Marking - Crosswalk, Crosswalk - Plain
	Intersection channelization	Islands and hatching to guide vehicle path	Curb, Sidewalk, Pedestrian Area, Lane Marking - Crosswalk, Crosswalk - Plain
	Railway-crossing structure		Fence, On Rails

Continued on next page

Table 3.2 – continued

iRAP category	iRAP variables	Corresponding items	Mapillary Vistas classes
	Property access points (driveways)		Curb Cut
VRU facilities	Pedestrian crossing facility - major/side road		Crosswalk – Plain, Pedestrian Area, Lane Marking - Crosswalk, Bridge
Land Use	Pedestrian fencing		Fence, Guard Rail
	Sidewalks - driver/passenger side	Sidewalk; Pedestrian Area; Curb; Curb Cut	
	Bicycle facility		Bike Lane; Bicycle
	School zone warning		Traffic Sign (Front)
Speeds	Speed limit - general/motorcycle/truck, Differential speed limits	Speed limit signs	Traffic Sign (Front)
	Speed management / Traffic calming		Curb

Unlike Mapillary Vistas, Cityscapes, Pascal VOC, ADE20K and similar segmentation benchmark datasets do not fit iRAP assessment comfortably. First, as these datasets were designed for generic scene understanding, their label taxonomies emphasize road users, e.g., cars, riders, pedestrians, or broad architectural categories, while omitting many of the static roadside objects that drive iRAP scores, as shown by Table 3.3. Second, iRAP surveys cover highways, rural arterials and peri-urban links worldwide. In contrast, Cityscapes frames are confined to German streets in urban areas; Pascal VOC mixes Flickr angles; ADE20K merges indoor and outdoor scenes. None of these sources replicate the driver-level, forward-looking perspective across multiple road classes that iRAP needs for uniform coding. SLIs of Mapillary Vistas are captured with varied camera rigs globally, covering both urban and rural carriageways under diverse curbside conditions. Thirdly, iRAP variables often hinge on dimensions or proximity, e.g. "safety-barrier within 0–1 m of edge". Accordingly, Mapillary

Vistas provides dense, pixel-level masks for small, safety-critical structures such as lane studs, guard-rail posts, drainage inlets and potholes, which potentially support detailed calculation. Cityscapes masks merge fine fixtures into a coarse "pole" or "fence" class; ADE20K labels many cues only at instance-level polygons; Pascal VOC has no pixel masks for infrastructure at all. Consequently, post-processing pipelines cannot derive distance bands or object widths with the same fidelity.

Table 3.3: Coverage of iRAP-relevant roadside assets in common semantic segmentation benchmarks.

Dataset	# classes	Roadside	Lane marking	VRU facilities	Midblock
Mapillary Vistas	65	Curb, Guard rail, Fence, Barrier, Wall, Pole, Traffic light/sign, Fire hydrant, Catch basin, ...	Lane Marking – General; Lane Marking – Crosswalk	Sidewalk, Bike lane, Pedestrian area, Parking, ...	Curb, Guard rail, Barrier, Fence, Street light, Pothole, ...
Cityscapes	30	Guard rail, Pole, Wall, Fence, Traffic sign/light	None (no lane-marking labels)	Sidewalk only	Guard rail, Fence, Parking
ADE20K (outdoor images)	150	Fence, Wall, Pole, Traffic light/sign, Street lamp, Fire hydrant, Guard rail†, ...	No lane-specific labels	Sidewalk; no bike-/crosswalk classes	none
PASCAL VOC	21	None (object set is person/animals/vehicles/furniture)	none	none	none

3.3 Methods

3.3.1 Segment Anything Model (SAM)

The Segment Anything Model (SAM) functions through three components. The image encoder, built on a Vision Transformer (ViT) backbone, transforms raw pixels into rich visual representations by splitting the image into patches and processing them with self-attention mechanisms. SAM's internal preprocessing will resize input images to 1024×1024 pixels as its defaults, before generating segmentation masks. The resized image is then converted into

a dense feature map that captures hierarchical visual information from edges and textures to semantic object properties. In a parallel thread, the prompt encoder translates user inputs into embeddings that guide segmentation, as shown in Figure 3.1. Point prompts represent user clicks on certain items that they want to include or exclude; in other words, positive points can be specified for target objects and negative points can be specified for background. Sultan *et al.* [168] explored combinations of positive and negative point prompts when segmenting road pavement and pedestrian crosswalk and sidewalk from satellite images. Number of point prompts does not directly translate to performances and SAM tends to produce coherent masks for batches or groups of point prompts. For example, if two positive point prompts are accidentally placed on a car and a person respectively, SAM will not produce two masks each representing the car and the person, but rather will produce a mask to cover both objects. Box prompts are rectangular boundaries that indicate the region of interest for segmentation. When a user draws a bounding box around an object, SAM focuses its attention within these constraints to generate a precise mask. Unlike point prompts which might segment multiple objects if they form a coherent group, box prompts reliably isolate individual objects even in cluttered scenes. For instance, in a crowded street scene, a box drawn around a specific pedestrian directs SAM to segment only that person, ignoring others nearby. Box prompts are particularly effective for initiating segmentation of objects with complex shapes or when objects are partially occluded, as they provide stronger spatial guidance than single points. Yuan *et al.* [204] used box prompts extensively in recognizing objects. Mask prompts allow users to provide rough initial segmentations that SAM can refine with greater precision. These coarse masks might come from quick manual sketches, outputs of simpler segmentation algorithms, or previous iterations of SAM itself. For example, a med-

ical professional might draw an approximate outline around a tumor in a scan, and SAM would then refine this into a precise boundary that captures subtle anatomical details. Mask prompts are especially valuable in interactive editing workflows, where users iteratively refine segmentations—starting with a preliminary mask and progressively improving it through additional prompts until achieving the desired accuracy. Text prompts enable natural language guidance for segmentation tasks. Specifically, SAM's original paper[101] claimed they used CLIP's text embedding to prompt SAM. However, such implementation was not included in SAM's source code, leaving room for researchers to define their own way of incorporating text prompts. The third component of SAM, i.e., mask decoder, combines embeddings produced by the image encoder and prompt encoder through a transformer architecture with specialized attention mechanisms. It processes token sequences representing both image features and prompts, generating binary masks that precisely delineate object boundaries. The decoder can output multiple mask predictions with confidence scores, allowing for ambiguity resolution in complex scenes where multiple valid interpretations exist, e.g., segmenting an entire car versus just its visible portion. SAM's class-agnostic design means it does not rely on predefined categories of the domain but instead understands fundamental visual concepts of "objects" and their boundaries. This general-purpose approach allows it to segment novel objects never seen during training, making it effective across diverse domains from natural scenes to medical imaging without domain-specific fine-tuning.

SAM also functions without any prompts. In the context of iRAP Star Ratings, this could correspond to situations where an annotator is starting safety assessment job fresh or when an annotator looks at data collected from an unfamiliar location, and thus the no guidance is provided. SAM then turns on its "everything mode" as an automatic object discovery system.

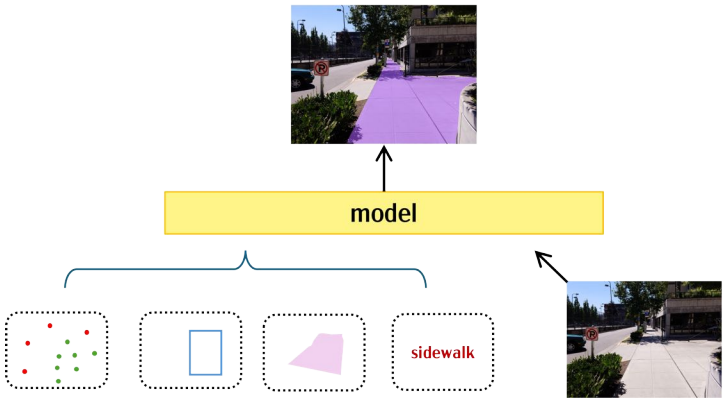


Figure 3.1: SAM can be prompted by four types of inputs: point, box, mask, and text. Two most common forms in the literature of prompts are point and box. In this dissertation, we focus on point prompts implicitly, i.e., all types of input are transformed to points to feed to SAM.

Behind the scene, the system overlays a regular equally-distanced grid of points across the image depending on desired level of detail. For a standard 1024×1024 image with 64-pixel grid spacing, this creates approximately 256 sampling points across the image surface. At each grid point, SAM performs foreground prediction: "what object exists at this point" and each point prompt is processed individually. This would generate a large number of masks over all prompted points and these masks are then processed through non-maximum suppression (NMS) algorithms to remove redundancies, as YOLO models do[146]. NMS retains only the highest-quality mask for each distinct object. The combined output of "everything mode" resembles having every possible object clicked and individually processed in the image.

Our core hypothesis in this chapter is SAM can be effectively applied to the task of semantic segmentation for road infrastructure element recognition. In particular, its strong generalization capabilities may eliminate the need for domain-specific fine-tuning, while its ability to produce high-quality, clean-cut masks is expected to enhance segmentation accuracy and boundary precision.

3.3.2 Domain priors

Gal and Ghahramani[53] proved that applying dropout during inference time can be interpreted as an approximation to a Bayesian Neural Network. This connection established dropout as more than just a regularization technique—it offers a mathematically proved approach to uncertainty quantification in deep learning models. By keeping dropout layers active during inference time, deterministic neural networks become approximate Bayesian models, capable of providing uncertainty estimates alongside predictions, without the computational burden of traditional Bayesian methods. Multiple stochastic forward passes with active dropout produce a distribution of predictions for each pixel. This distribution captures both aleatoric uncertainty (inherent randomness in the data) and epistemic uncertainty (model uncertainty due to limited training data). Each pass produces slightly different segmentation masks for the same input image, forming a probability distribution over possible class assignments for each pixel. For DeepLab architectures[39], implementing such Monte Carlo (MC) dropout requires strategic placement of dropout layers, typically after major feature extraction blocks. The most effective placement is after the Atrous Spatial Pyramid Pooling (ASPP) module, which aggregates multi-scale contextual features while preserving spatial resolution. Introducing dropout at this stage allows stochasticity to be injected into high-level semantic representations, enabling the model to capture epistemic uncertainty without significantly degrading its representational capacity or segmentation performance. For Transformer-based models like Mask2Former[41] with Swin backbones, the SwinDropPath mechanism serves as an ideal uncertainty estimator. Unlike standard dropout that randomly zeroes activations, SwinDropPath stochastically drops entire paths within the computational graph during inference, providing structured forms of uncertainty that better align with the

hierarchical representations learned by these models.

While segmentation models do produce probability distributions via softmax, these probabilities only capture aleatoric uncertainty (data uncertainty) but fail to represent epistemic uncertainty (model uncertainty). Softmax outputs only represent *relative preferences* between classes based on a single deterministic forward pass, often resulting in overconfident predictions even when the model encounters unfamiliar inputs. The fundamental limitation is that standard softmax probabilities are derived from point estimates of network weights rather than distributions over weights. This means a model might assign high confidence, e.g., 0.99 probability, to entirely wrong predicted classes when facing out-of-distribution data, providing misleading confidence in areas where the model has limited knowledge. MC dropout addresses this limitation by approximating a Bayesian treatment of neural networks through multiple stochastic forward passes with dropout enabled. Each pass samples from an implicit distribution over network weights, producing a distribution of predictions that captures both aleatoric and epistemic uncertainty—high variance across MC samples indicates regions where the model's knowledge is limited due to training data sparsity or domain shifts. Therefore, the most valuable prior information domain models can provide is pixelwise uncertainty, which serves as proxy for prediction confidence at this pixel, offering a principled way to identify regions where predictions are less reliable. This prior knowledge can be quantified using mutual information [105] based on MC runs.

If we denote the probability from the i-th MC run of class c at pixel (h, w) as $p_i(c, h, w)$, we would have $\sum_c p_i(c, h, w) = 1$ given it is a distribution over all candidate classes c. The

expected probability for class c at pixel (h, w) can be obtained as an average from all MC runs

$$\bar{p}(c, h, w) = \mathbb{E}_i\big[p_i(c, h, w)\big] = \frac{1}{N}\sum_{i=1}^{N} p_i(c, h, w) \tag{3.1}$$

Predictive entropy, i.e., entropy of the average prediction, can be computed as

$$H\big(\bar{p}(h, w)\big) = -\sum_{c=1}^{C} \bar{p}(c, h, w) \cdot \log \bar{p}(c, h, w) \tag{3.2}$$

And individual entropy from MC run i is

$$H\big(p_i(h, w)\big) = -\sum_{c=1}^{C} p_i(c, h, w) \cdot \log p_i(c, h, w) \tag{3.3}$$

Therefore, mutual information at pixel (h, w) from these MC runs

$$\mathrm{MI}(h, w) = H\big(\bar{p}(h, w)\big) - \frac{1}{N}\sum_{i=1}^{N} H\big(p_i(h, w)\big) \tag{3.4}$$

and predicted class at pixel (h, w) is

$$\hat{y}(h, w) = \arg\max_{c} \bar{p}(c, h, w) \tag{3.5}$$

We hypothesize that pixel-level mutual information serves as a reliable proxy for model confidence, and that selecting pixels based on this measure can effectively guide the classification of each mask generated by SAM.

3.3.3 Integrate SAM with domain priors

Figure 3.2 illustrates proposed zero-shot inference pipeline that integrates SAM with domain-specific prior knowledge discussed earlier.

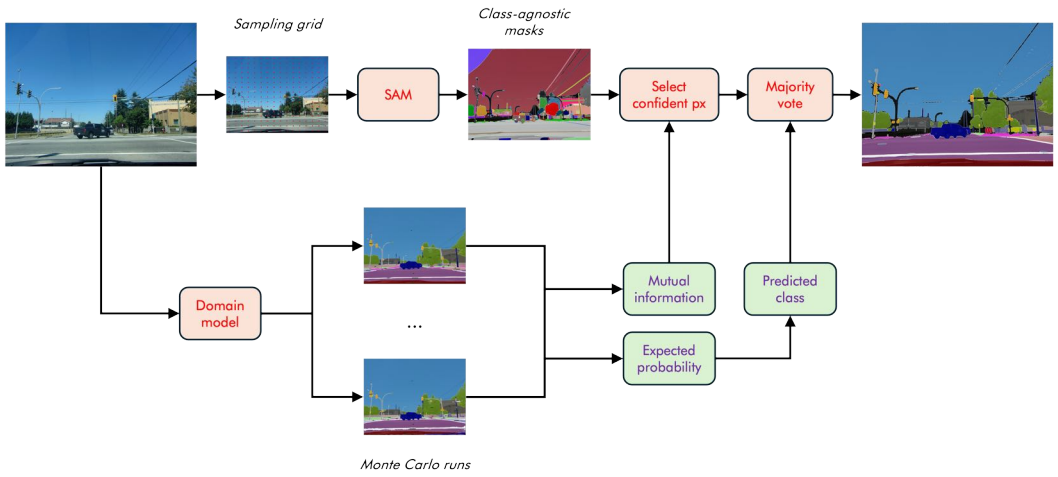


Figure 3.2: Domain priors enable context-aware classification of SAM's masks. SAM generates equally distanced grid and produces masks, i.e., in "everything mode". Prior knowledge from domain models are estimated from Monte Carlo (MC) dropout forward passes per pixel: mutual information and class probability distribution. Mutual information is used to select confident pixels per mask with pre-defined local percentiles and most likely class of these pixels is used in majority voting for the mask.

The process begins with a SLI where a uniform sampling grid is applied, creating equally distanced points. SAM treats each sampled point on the grid as an individual point prompt and generates initial segmentation masks for them. Produced masks are post-processed with hyperparameters, including (1) `pred_iou` $= 0.88$: a threshold on the model's predicted Intersection over Union (IoU) score for each mask, hoping to retain only the high-confidence masks—those that the model thinks are likely to be accurate; (2) `stability_score` $= 0.95$: a measure of stability of the predicted mask under small perturbations to the input prompt, e.g., slightly changing the point or bounding box prompt, in order to prioritize reliable and

repeatable masks, filtering out those that may fluctuate due to prompt sensitivity. Concurrently, the input image is processed through a domain model that performs multiple MC runs with Dropout or DropPath enabled (depending on the model family). This generates multiple forward passes that capture model uncertainty, from which two key metrics are computed: mutual information (quantifying prediction uncertainty per pixel) and probability expectation (estimating class probability distributions) as Equations 3.4 and 3.5 describe. Mutual information is then used to select confident pixels within each mask based on predefined local percentile band that are believed to include confident pixels. For each mask, the selected confident pixels participate in majority voting where the most frequently predicted class among these pixels becomes the mask's label. The final output is a class-aware segmentation map. Selecting confident pixels based on mutual information could be be ineffective or complex, but we hypothesize that a carefully selected range based on local percentiles is sufficient to produce good voting result. Figure 3.3 shows outputs of the proposed method step by step: (1) SAM produces class-agnostic masks from grid sampling, where colors are based on order of the masks; (2) majority voting based on domain model predictions decides a class per mask; (3) for pixels not covered by any mask, the framework would use domain model predictions as fallback.

3.4 Results

Models are benchmarked on the Mapillary Vistas dataset, which consists of 25,000 high-resolution images collected from diverse geographic locations around the world and captured under various weather and lighting conditions[131]. These images are annotated with semantic labels per pixel covering 66 object categories, 24 of which are related to road infrastructure el-

(a) Masks produced by SAM

(b) Segmentation of pixels covered by masks after domain voting

(c) Segmentation of all pixels with uncovered pixels defaulted to domain model

Figure 3.3: Proposed pipeline includes three steps: (a) SAM generates class-agnostic masks; (b) majority voting based on domain priors assigns labels per mask; (c) pixels not covered by any SAM mask fall back to domain model prediction. The pipeline enables flexible integration of strong boundaries and domain-aware classification.

ements in iRAP. The dataset is divided into 18,000 training images, 2,000 validation images, and 5,000 test images, with image resolutions ranging from 1024×768 to 4000×6000 pixels. For fair comparison, we only look at the validation set as domain benchmarks are extensively trained on the training set. The 25 classes related to road infrastructure elements are picked from the taxonomy and performances on this cohort is reported with overall metrics: Curb, Fence, Guard Rail, Barrier, Wall, Bike Lane, Crosswalk - Plain, Parking, Rail Track, Road, Service Lane, Sidewalk, Lane Markings - Crosswalk, Lane Markings - General, Catch Basin, Fire Hydrant, Junction Box, Manhole, Pothole, Street Light, Pole, Sign Frame, Utility Pole, Traffic Light, Traffic Sign - Front.

Standard metrics from semantic segmentation tasks are used in this study. Intersection over Union (IoU), also known as the Jaccard index, measures the overlap area between the predicted segmentation mask and the ground truth divided by their union. It ranges from 0 to 1, where 1 indicates perfect overlap. IoU is widely used in segmentation tasks because it penalizes both false positives and false negatives equally, making it sensitive to segmentation quality regardless of class imbalance. It is measured per class and then averaged across classes

to provide an overall performance measure: mean IoU or mIoU.

$$\texttt{IoU}_c = \frac{\texttt{TP}_c}{\texttt{TP}_c + \texttt{FP}_c + \texttt{FN}_c} \tag{3.6}$$

$$\texttt{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \texttt{IoU}_c \tag{3.7}$$

Accuracy for semantic segmentation is measured over pixels. It is the ratio of correctly classified pixels to the total number of pixels of the class in the image. Despite being straightforward, this metric can be biased when dealing with unbalanced classes among pixels, as high accuracy might be achieved by simply predicting the dominant class. This is particularly true in road infrastructure recognition in SLIs because background object classes like *sky* and *road* could take up a huge portion of the view. Accuracy can be averaged in two ways: mean accuracy mAcc is the arithmetic mean of per-class accuracies, telling how well each class is recognized by treating all classes equally, regardless of their frequency in the dataset; average accuracy aAcc is the overall pixel-level accuracy, telling overall correctness across all pixels or examples, weighted by class frequency and thus dominated by common classes.

$$\texttt{Acc}_c = \frac{\texttt{TP}_c}{\texttt{TP}_c + \texttt{FN}_c} \tag{3.8}$$

$$\texttt{mAcc} = \frac{1}{C}\sum_{c=1}^{C} \texttt{Acc}_c \tag{3.9}$$

$$\texttt{aAcc} = \frac{\sum_{c=1}^{C} \texttt{TP}_c}{\sum_{c=1}^{C} \left(\texttt{TP}_c + \texttt{FP}_c + \texttt{FN}_c\right)} \tag{3.10}$$

The Dice coefficient measures the overlap between the predicted segmentation and ground truth. Similar to IoU, value of Dice coefficient ranges from 0 to 1, with 1 indicating perfect

overlap. In fact, it is not hard to see Dice and IoU are related via $\texttt{Dice} = \frac{2 \times \texttt{IoU}}{1+\texttt{IoU}}$. This means that Dice always gives a higher numerical value than IoU for the same segmentation result (except at the extreme values of 0 and 1, where they are equal). This also means that while IoU penalizes both false positives and false negatives equally, Dice coefficient gives more weight to true positives relative to false positives and false negatives. Or in other words, Dice ignores the overwhelming number of true negatives and concentrates on how well the model captures the rare positive class, making it better suited for imbalanced data. Hence Dice is especially useful for evaluating segmentation of small structures or when classes are imbalanced.

$$\texttt{Dice}_c = \frac{2 \cdot \texttt{TP}_c}{2 \cdot \texttt{TP}_c + \texttt{FP}_c + \texttt{FN}c} \tag{3.11}$$

$$\texttt{mDice} = \frac{1}{C} \sum_{c=1}^{C} \texttt{Dice}_c \tag{3.12}$$

Precision, recall, and F score are common standard metrics used in classification problems. Precision measures the proportion of correct positive pixels among all pixels predicted as positive. High precision indicates low false positive rates, meaning when the model predicts a pixel belongs to a class, it is likely correct. Recall measures the proportion of actual positive pixels that are correctly identified. It focuses on the model's ability to find all positive instances. High recall indicates low false negative rates, meaning the model successfully identifies most pixels that belong to a class. F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics, and equals Dice coefficient in the segmentation context.

$$\texttt{Precision}_c = \frac{\texttt{TP}_c}{\texttt{TP}_c + \texttt{FP}_c} \tag{3.13}$$

$$\texttt{mPrecision} = \frac{1}{C}\sum_{c=1}^{C} \texttt{Precision}_c \tag{3.14}$$

$$\texttt{Recall}_c = \frac{\texttt{TP}_c}{\texttt{TP}_c + \texttt{FN}_c} \tag{3.15}$$

$$\texttt{mRecall} = \frac{1}{C}\sum_{c=1}^{C} \texttt{Recall}_c \tag{3.16}$$

$$\texttt{Fscore}_c = 2 \cdot \frac{\texttt{Precision}_c \cdot \texttt{Recall}_c}{\texttt{Precision}_c + \texttt{Recall}_c} \tag{3.17}$$

$$\texttt{mFscore} = \frac{1}{C}\sum_{c=1}^{C} \texttt{Fscore}_c \tag{3.18}$$

3.4.1 Single pass

The hypothesis that SAM is effective for road infrastructure element recognition using SLIs is validated by two key observations. Firstly, we demonstrate that the proposed pipeline using SAM (see Figure 3.2) achieves a higher upper bound than baseline domain model predictions. Instead of voting with domain model predictions for each mask, most frequent class is selected based on ground truth labels from confident pixels. If this combination of SAM masks with ground truth voting underperforms relative to baseline domain models, it implies that using domain models cannot surpass this ceiling—rendering further exploration of SAM-based approaches ineffective—because predictions for pixels not covered by SAM segmentation masks are defaulted to domain models. Table 3.4 presents this comparison on the validation set of Mapillary Vistas, with the two domain models trained on this dataset

(the training set). For the comparison to be fair, metrics are calculated on pixels covered by SAM's masks. Results clearly show that **S1 + GT** outperforms both domain models by large margin in all metrics. This superior performance also extends to the selected classes related to road infrastructure elements. This proves the proposed pipeline is taking advantage of SAM's superior mask quality and making the right selection of confident pixels. Table 3.5 provides per-class comparisons to support the validated upper bound.

Table 3.4: Overall metrics comparing **SAM + vote by ground truth labels** (S1 + GT) vs. baseline domain models on **covered pixels**.

Metric	S1 + GT	Mask2Former	Deeplabv3+
aAcc	0.970	0.928	0.918
mAcc	0.906	0.668	0.579
mIoU	0.862	0.516	0.468
mDice	0.920	0.642	0.593
mPrecision	0.939	0.636	0.652
mRecall	0.906	0.668	0.579
mFscore	0.920	0.642	0.593
aAcc_select	0.939	0.869	0.847
mIoU_select	0.812	0.524	0.486
mDice_select	0.891	0.670	0.631
mPrecision_select	0.920	0.678	0.690
mRecall_select	0.869	0.669	0.614
mFscore_select	0.891	0.670	0.631

Note. "S1" – SAM (one pass), "GT" – vote by ground truth label.

Quality of mask edges also supports SAM's effectiveness for our task. Figure 3.4 reveals qualitatively that SAM produces finer and more continuous boundaries than the two domain segmentation models. While the ground truth boundaries are focusing on the outer boundary of objects, SAM captures additional details such as fine structures on vehicles, traffic lights, and power lines. This enhanced delineation holds SAM great potential in capturing elements at fine scales. The upper bound performance in Table 3.4 also tells about quality of the segmentation masks: if the mask edges have good precision (the pixels of predicted mask edges are correct) and recall (ground truth mask edges are identified), then the vote result by

Table 3.5: Class-wise metrics comparing **SAM + vote by ground-truth labels** (S1 + GT) vs. baseline domain models on **covered pixels**.

Metric	Model	Curb	Fence	Guard Rail	Barrier	Wall	Bike Lane	Crosswalk P.	Parking	Rail Track	Road	Service Lane	Sidewalk	LM Crosswalk	LM General	Catch Basin	Fire Hydrant	Junction Box	Manhole	Pothole	Street Light	Pole	Sign Frame	Utility Pole	Traffic Light	Sign Front
IoU	S1 + GT	0.740	0.907	0.913	0.931	0.942	0.759	0.536	0.639	0.619	0.926	0.749	0.919	0.591	0.668	0.815	0.911	0.948	0.853	0.823	0.709	0.815	0.918	0.849	0.879	0.948
	Mask2Former	0.578	0.579	0.646	0.550	0.520	0.361	0.422	0.180	0.496	0.882	0.347	0.692	0.732	0.604	0.451	0.499	0.483	0.492	0.095	0.443	0.502	0.651	0.529	0.659	0.705
	Deeplabv3+	0.560	0.579	0.646	0.494	0.461	0.303	0.275	0.126	0.474	0.853	0.309	0.647	0.597	0.579	0.401	0.469	0.430	0.454	0.020	0.530	0.411	0.528	0.546	0.730	0.727
Acc	S1 + GT	0.822	0.938	0.944	0.950	0.971	0.874	0.628	0.676	0.828	0.980	0.773	0.966	0.658	0.710	0.858	0.958	0.974	0.901	0.853	0.769	0.896	0.957	0.920	0.942	0.975
	Mask2Former	0.775	0.765	0.779	0.720	0.675	0.590	0.526	0.312	0.719	0.940	0.453	0.868	0.821	0.766	0.552	0.673	0.710	0.618	0.135	0.560	0.689	0.799	0.698	0.769	0.813
	Deeplabv3+	0.786	0.718	0.718	0.738	0.597	0.436	0.339	0.175	0.652	0.938	0.434	0.791	0.649	0.748	0.501	0.559	0.550	0.595	0.020	0.653	0.690	0.616	0.754	0.827	0.864
Dice	S1 + GT	0.850	0.951	0.954	0.964	0.970	0.863	0.698	0.780	0.765	0.961	0.857	0.958	0.743	0.801	0.898	0.954	0.973	0.920	0.903	0.829	0.898	0.957	0.918	0.936	0.973
	Mask2Former	0.733	0.733	0.785	0.709	0.684	0.531	0.593	0.305	0.663	0.937	0.515	0.818	0.845	0.753	0.621	0.666	0.651	0.660	0.173	0.614	0.668	0.789	0.692	0.795	0.827
	Deeplabv3+	0.718	0.734	0.785	0.662	0.631	0.465	0.432	0.224	0.643	0.920	0.472	0.786	0.748	0.733	0.572	0.638	0.601	0.624	0.040	0.692	0.583	0.691	0.706	0.844	0.842
Fscore	S1 + GT	0.850	0.951	0.954	0.964	0.970	0.863	0.698	0.780	0.765	0.961	0.857	0.958	0.743	0.801	0.898	0.954	0.973	0.920	0.903	0.829	0.898	0.957	0.918	0.936	0.973
	Mask2Former	0.733	0.733	0.785	0.709	0.684	0.531	0.593	0.305	0.663	0.937	0.515	0.818	0.845	0.753	0.621	0.666	0.651	0.660	0.173	0.614	0.668	0.789	0.692	0.795	0.827
	Deeplabv3+	0.718	0.734	0.785	0.662	0.631	0.465	0.432	0.224	0.643	0.920	0.472	0.786	0.748	0.733	0.572	0.638	0.601	0.624	0.040	0.692	0.583	0.691	0.706	0.844	0.842
Precision	S1 + GT	0.881	0.964	0.965	0.979	0.970	0.853	0.784	0.921	0.710	0.944	0.961	0.950	0.854	0.919	0.941	0.949	0.973	0.940	0.960	0.900	0.900	0.958	0.916	0.929	0.972
	Mask2Former	0.695	0.704	0.791	0.699	0.693	0.482	0.681	0.297	0.615	0.935	0.596	0.774	0.870	0.740	0.711	0.658	0.601	0.708	0.239	0.680	0.648	0.778	0.686	0.822	0.842
	Deeplabv3+	0.661	0.750	0.866	0.600	0.670	0.498	0.593	0.310	0.635	0.904	0.517	0.780	0.883	0.719	0.668	0.744	0.664	0.657	0.757	0.737	0.504	0.787	0.664	0.862	0.821
Recall	S1 + GT	0.822	0.938	0.944	0.950	0.971	0.874	0.628	0.676	0.828	0.980	0.773	0.966	0.658	0.710	0.858	0.958	0.974	0.901	0.853	0.769	0.896	0.957	0.920	0.942	0.975
	Mask2Former	0.775	0.765	0.779	0.720	0.675	0.590	0.526	0.312	0.719	0.940	0.453	0.868	0.821	0.766	0.552	0.673	0.710	0.618	0.135	0.560	0.689	0.799	0.698	0.769	0.813
	Deeplabv3+	0.786	0.718	0.718	0.738	0.597	0.436	0.339	0.175	0.652	0.938	0.434	0.791	0.649	0.748	0.501	0.559	0.550	0.595	0.020	0.653	0.690	0.616	0.754	0.827	0.864

Note. "S1" = SAM single-pass; "GT" = voting by ground-truth label.

ground truth pixels contained in the masks is supposed to be perfect. All metrics being higher of **S1 + GT** tells that SAM generate masks significantly better than the two domain models. Table 3.6 provides another quantitative measurement of this quality. While the precision metrics might appear lower for SAM at first glance, recall is the more important metric in this task given the inconsistent level of details between needs of this task and SAM's output. Ground truth labels annotate pixels by 65 object classes, unable to distinguish window and tire of a car. SAM, on the other hand, identifies "objectness" and produces class-agnostic masks for all of them, many of which could be penalized as "false positives" simply because these objects are not distinguished in labels, artificially lowering SAM's precision scores. On the contrary, recall for edge map tells the percentage of pixels in the ground truth edge map being captured. At three-pixel dilations, SAM's substantially higher recall indicates its more comprehensive boundary detection capabilities. This comparison in edge map recall is visualized in Figure 3.5, where edge map of the ground truth label is colored per pixel. Green

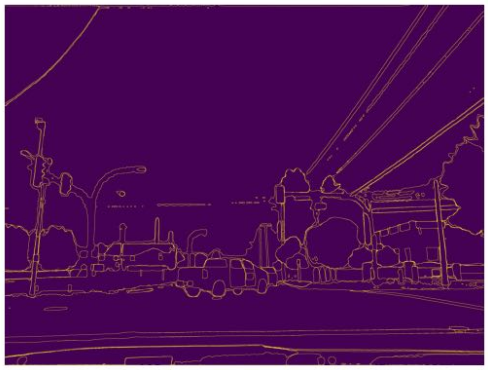
indicates the existence of pixel in the edge map of the corresponding model's output within the dilation neighborhood and thus successfully identified; while red means there is not any pixel in model's output edge map close enough and thus the pixel not identified. SAM captures more boundaries of major objects in the SLI, including buildings, trees, poles, traffic lights, persons, cars, roadside objects, and even lane markings and manholes, but also fails to capture details in the guardrails and poles of street lights, compared to Deeplabv3+. This performance differential suggests that SAM's architecture enables it to capture a more exhaustive and accurate representation of scene boundaries, making it particularly valuable for applications requiring precise object delineation in complex urban environments.

Table 3.6: Boundary quality metrics (3-pixel and 10-pixel dilation radius) of SAM (one pass) vs. baseline domain models.

Model (dilation)	Precision	Recall	F1
SAM (3 pixel)	0.296	0.494	0.370
SAM (10 pixel)	0.524	0.767	0.622
Mask2Former (3 pixel)	0.423	0.292	0.345
DeepLabv3+ (3 pixel)	0.380	0.425	0.402

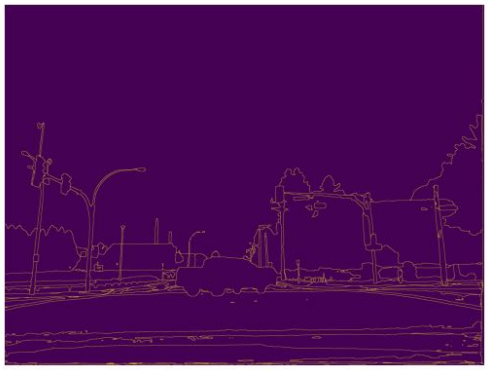
After establishing a rough upper bound for the proposed framework and observing the high-quality masks produced by SAM, the next step is to investigate why pixel-level uncertainty is a meaningful signal. Figure 3.6 presents pixel-level mutual information (measure of prediction uncertainty) and misclassification rate. As mutual information (x-axis) increases beyond approximately 10^{-6}, there is a corresponding increase in the misclassification rate (blue line with diamond markers). This trend becomes particularly pronounced at mutual information values above 10^{-5}, where the misclassification rate rises sharply from about 10% to over 50% at the highest values. This plot illustrates a clear relationship between mutual information and misclassification rates, demonstrating that our uncertainty metric effectively identifies problematic regions in the segmentation process, providing a valuable signal for

(a) Ground Truth



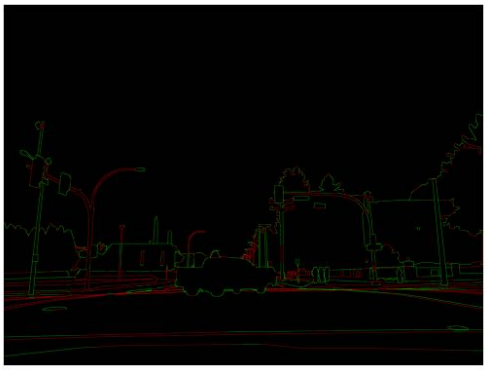
(b) SAM

(c) Mask2Former

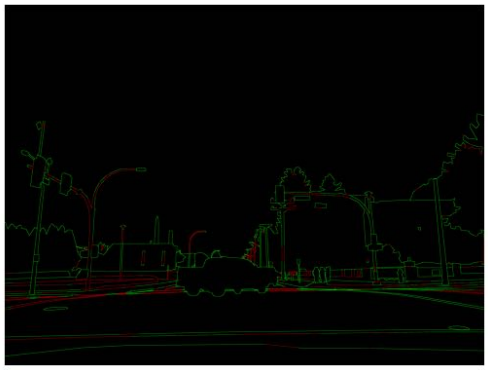


(d) Deeplabv3+

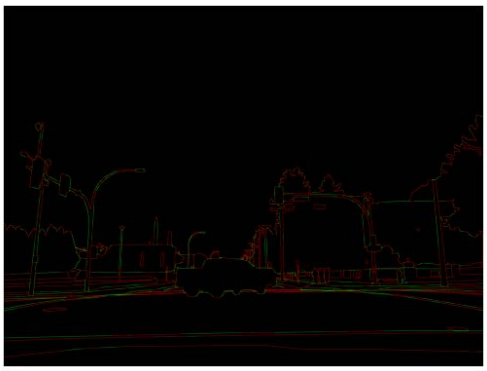
Figure 3.4: Edge map of SAM's masks in a representaive street scene. SAM (panel **b**) produces finer and more continuous boundaries than the two domain segmentation models (panels **c-d**). While the ground truth boundaries (panel **a**) are relatively sparse and coarse, SAM (panel **b**) captures additional details such as fine structures on vehicles, traffic lights, and power lines. This enhanced delineation contributes to its superior boundary recall performance.



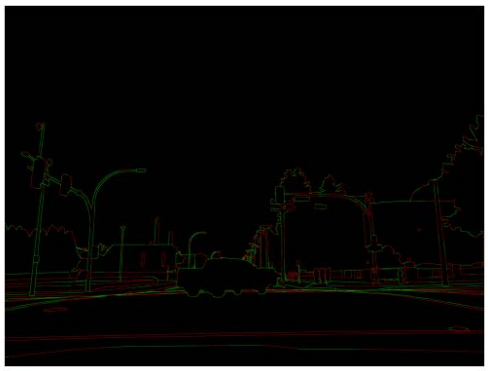
(a) SAM (3 pixel dilation)



(b) SAM (10 pixel dilation)



(c) Mask2Former (3 pixel dilation)



(d) Deeplabv3+ (3 pixel dilation)

Figure 3.5: Edge maps of SAM's masks compare with two baseline domain models. At three-pixel dilation, edges of SAM's masks (panel **a**) achieve 0.49 recall, while Mask2Former (panel **c**) achieves 0.29 and Deeplabv3+ (panel **d**) achieves 0.43. Edges of SAM's masks achieve 0.77 recall at ten-pixel dilation (panel **b**).

detecting potential errors in the model's predictions.

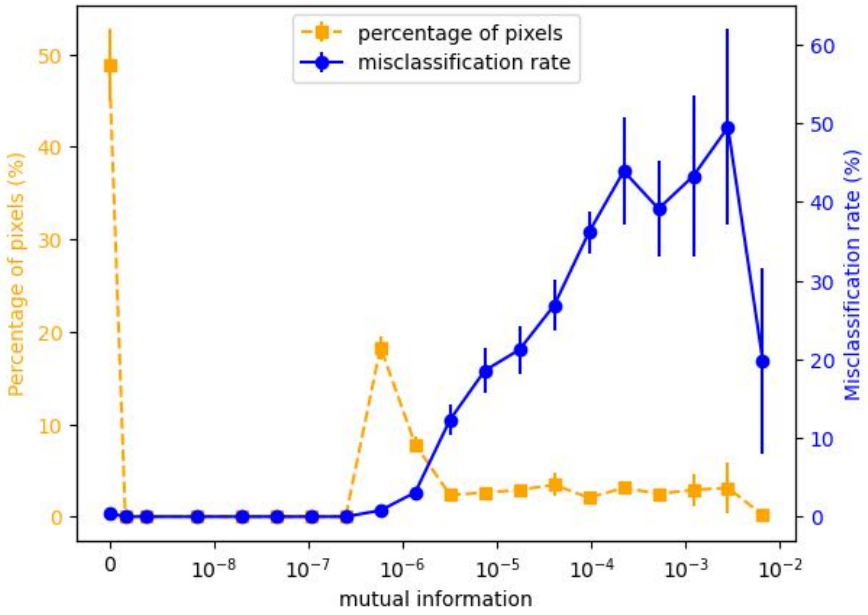


Figure 3.6: Pixel-wise mutual information is positively related to misclassification rate. As mutual information increases for pixels, misclassification rates also rise for them, indicating that pixels with higher uncertainty across MC runs are more prone to error. This trend highlights mutual information as a meaningful indicator of prediction confidence, making it a useful form of prior knowledge from domain models for identifying unconfident pixels.

With SAM's utility established and mutual information validated as an effective indicator of pixel-level confidence, the next key step is to determine an appropriate voting mechanism to assign class labels to SAM-generated masks. Figure 3.7 provides insights into optimizing this voting mechanism by plotting the relationship between local percentiles and misclassification rates, where local percentile of a pixel is calculated by ranking mutual information within the mask this pixel belongs. Looking at the blue line representing misclassification rates, we observe a U-shaped pattern with the lowest error rates occurring between the 30th and 50th local percentiles (approximately 4.5-5%). As the local percentile increases beyond 60th percentile, misclassification rates begin to rise gradually, then increase dramatically after

the 80th percentile, ultimately exceeding 17% at the highest examples. This stability allows for direct comparison of performance differences. Given the error rate of pixels at 0th - 75th percentiles do not deviate much, a upper threshold of 70th local percentile can be used. An even better approach is to use interval 30th to 50th percentiles as this range is associated with the lowest pixel-wise misclassification rates. This finding establishes a data-driven approach to uncertainty-based voting that maximizes segmentation accuracy while maintaining computational efficiency by utilizing only the most confident and stable subset of pixels.

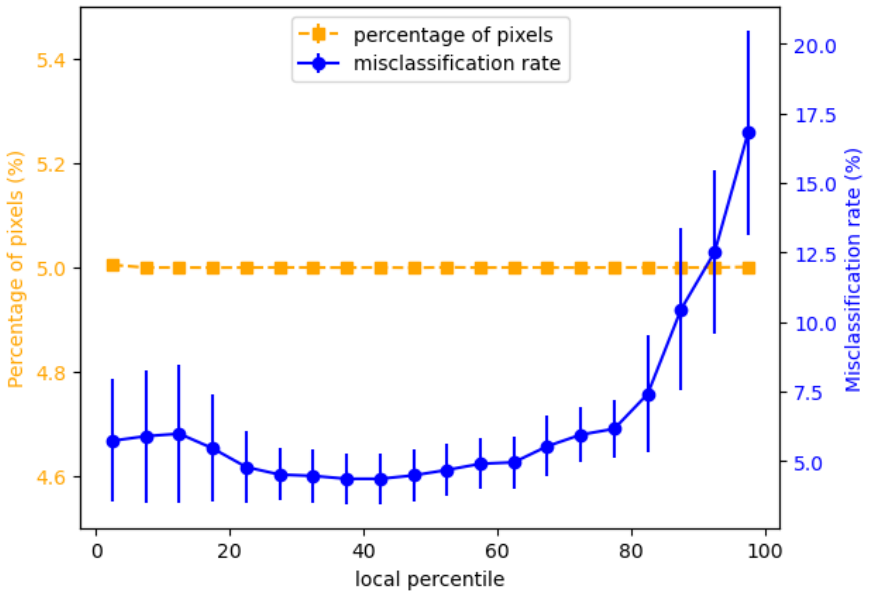


Figure 3.7: Local percentiles of pixel-wise mutual information are related to misclassification rate. This relationship exhibits a clear and monotonic pattern, with the highest misclassification rates concentrated in the top percentiles. In contrast, pixels within the middle percentile range (e.g., 20th–60th) show the lowest error rates. This simplicity makes percentile-based thresholding a practical strategy for identifying reliable predictions. By retaining only pixels within a selected confidence interval, such as the 20–60% range, downstream voting steps can prioritize more trustworthy pixels.

Leveraging these validated components, the proposed model achieves solid performance compared to baseline models. Table 3.7 shows S1 + M (combining SAM with Mask2Former votes) does not reach the upper bound of S1 + GT, but it does reach the range of domain

model performances. It is particularly noticeable that S1 + M outperforms Deeplabv3+ at `mIoU` and `mDice` metrics considering overall object classes in the dataset and among the selected infrastructure class cohort. This means the proposed approach demonstrates specialized effectiveness for infrastructure segmentation tasks. Table 3.8 and Table 3.9 describe a similar story for all pixels regardless of whether they are covered or not by SAM's masks.

Table 3.7: Overall metrics comparing **SAM + vote by ground truth labels** (S1 + GT) vs. baseline domain models vs. **S1 + M** on **covered pixels**.

Metric	S1 + GT (upper bound)	Mask2Former (baseline)	Deeplabv3+ (baseline)	S1 + M (proposed)
aAcc	0.970	0.928	0.918	0.912
mAcc	0.906	0.668	0.579	0.616
mIoU	0.862	0.516	0.468	0.473
mDice	0.920	0.642	0.593	0.617
mPrecision	0.939	0.636	0.652	0.624
mRecall	0.906	0.668	0.579	0.616
mFscore	0.920	0.642	0.593	0.617
aAcc_select	0.939	0.869	0.847	0.843
mIoU_select	0.812	0.524	0.486	0.490
mDice_select	0.891	0.670	0.631	0.634
mPrecision_select	0.920	0.678	0.690	0.657
mRecall_select	0.869	0.669	0.614	0.614
mFscore_select	0.891	0.670	0.631	0.634

Note. "S1" – SAM (one pass), "GT" – vote by ground truth label.

3.4.2 Iterative sampling

Table 3.10 presents the drop of performance from pixels covered by SAM masks to all pixels in the SLIs. Performance on all pixels is consistently lower than performances on pixels that are covered by SAM masks. This suggests that the quality of prediction is higher in areas where SAM provides coverage, which means our proposed mechanism that domain votes on SAM masks is better than domain models themselves. Two strategies could narrow this gap by improving SAM's pixel-level coverage. The first idea focuses on increasing the sampling density of the grid. Given the dimensions of SLIs in Mapillary Vistas, the sampling grid can

Table 3.8: Overall metrics comparing S1 + GT vs. baseline domain models vs. **S1 + M** on **all pixels**.

Metric	S1 + GT (upper bound)	Mask2Former (baseline)	Deeplabv3+ (baseline)	S1 + M (proposed)
aAcc	0.913	0.914	0.908	0.873
mAcc	0.825	0.646	0.562	0.588
mIoU	0.785	0.488	0.448	0.469
mDice	0.872	0.618	0.576	0.601
mPrecision	0.932	0.609	0.634	0.644
mRecall	0.825	0.646	0.562	0.588
mFscore	0.872	0.618	0.576	0.601
aAcc_select	0.887	0.852	0.834	0.815
mIoU_select	0.739	0.494	0.466	0.476
mDice_select	0.844	0.645	0.614	0.627
mPrecision_select	0.913	0.649	0.674	0.689
mRecall_select	0.791	0.647	0.596	0.582
mFscore_select	0.844	0.645	0.614	0.627

Note. "S1" – SAM (one pass), "GT" – vote by ground truth label, "M" – vote by Mask2Former predictions.

start from 32×32 and grow to 64×64, or 128×128. That would translate to roughly a sampling point every 24 pixels per side of the SLI. The hypothesis is that this dense sampling grid would capture all meaningful objects in the SLI in one pass. The second idea also seeks to cover all meaningful objects in the SLI with the sampling grid; however instead of completing in one pass with an extremely dense grid, it proposes a two-pass iterative approach. This idea considers the post-processing mechanism of SAM masks, in particular the NMS step, where overlapping masks are merged to preserve the best one. Given this mechanism, an extremely dense sampling grid hoping to capture everything in one pass would be inefficient because all of its sampled points will be processed individually but most of the generated masks will be merged. This fact motivates an adaptive approach, where a coarse grid can capture major objects in the SLI from the first pass. In the second pass a finer grid can avoid areas covered by masks generated via the coarse grid and focus on previously uncovered pixels. This iterative approach builds on an greedy assumption that SAM masks are coherent and masks produced in later passes are not bigger than masks produced in previous passes, otherwise they would

Table 3.9: Class-wise metrics comparing S1 + GT, Mask2Former, Deeplabv3+, and **S1 + M** on **all pixels**.

Metric	Model	Curb	Fence	Guard Rail	Barrier	Wall	Bike Lane	Crosswalk P.	Parking	Rail Track	Road	Service Lane	Sidewalk	LM Crosswalk	LM General	Catch Basin	Fire Hydrant	Junction Box	Manhole	Pothole	Street Light	Pole	Sign Frame	Utility Pole	Traffic Light	Sign Front
IoU	S1 + GT	0.588	0.762	0.758	0.858	0.888	0.733	0.512	0.554	0.571	0.897	0.683	0.849	0.563	0.629	0.763	0.811	0.908	0.837	0.807	0.574	0.679	0.824	0.716	0.809	0.906
	Mask2Former	0.537	0.542	0.598	0.521	0.492	0.358	0.415	0.179	0.480	0.869	0.339	0.672	0.719	0.591	0.432	0.445	0.457	0.485	0.093	0.384	0.425	0.599	0.453	0.597	0.662
	Deeplabv3+	0.530	0.563	0.596	0.473	0.444	0.298	0.271	0.134	0.454	0.842	0.308	0.630	0.590	0.569	0.386	0.428	0.415	0.448	0.020	0.482	0.385	0.507	0.493	0.683	0.694
	S1 + M	0.464	0.519	0.589	0.513	0.515	0.354	0.257	0.137	0.414	0.843	0.316	0.645	0.467	0.535	0.444	0.514	0.480	0.501	0.065	0.387	0.479	0.625	0.498	0.647	0.700
Acc	S1 + GT	0.653	0.789	0.784	0.875	0.915	0.843	0.601	0.586	0.764	0.950	0.704	0.892	0.626	0.668	0.804	0.853	0.933	0.885	0.836	0.623	0.747	0.859	0.776	0.867	0.932
	Mask2Former	0.739	0.736	0.756	0.701	0.652	0.583	0.517	0.312	0.696	0.931	0.441	0.847	0.810	0.754	0.537	0.629	0.689	0.611	0.135	0.508	0.635	0.772	0.657	0.733	0.791
	Deeplabv3+	0.760	0.701	0.682	0.722	0.578	0.428	0.334	0.187	0.629	0.932	0.430	0.774	0.641	0.737	0.483	0.519	0.533	0.588	0.020	0.609	0.650	0.599	0.718	0.794	0.844
	S1 + M	0.579	0.646	0.652	0.662	0.657	0.525	0.321	0.220	0.606	0.923	0.383	0.801	0.536	0.605	0.514	0.620	0.687	0.605	0.102	0.452	0.607	0.735	0.610	0.727	0.785
Dice	S1 + GT	0.740	0.865	0.862	0.924	0.941	0.846	0.678	0.713	0.727	0.946	0.812	0.918	0.720	0.772	0.865	0.896	0.952	0.911	0.893	0.729	0.809	0.904	0.834	0.894	0.951
	Mask2Former	0.698	0.703	0.748	0.685	0.660	0.527	0.587	0.304	0.648	0.930	0.506	0.804	0.836	0.743	0.603	0.616	0.627	0.653	0.171	0.555	0.597	0.749	0.624	0.748	0.797
	Deeplabv3+	0.692	0.720	0.747	0.642	0.615	0.459	0.426	0.236	0.624	0.914	0.471	0.773	0.742	0.725	0.557	0.599	0.586	0.619	0.039	0.651	0.555	0.673	0.661	0.812	0.820
	S1 + M	0.634	0.683	0.742	0.678	0.680	0.523	0.409	0.241	0.585	0.915	0.481	0.784	0.637	0.697	0.615	0.679	0.649	0.667	0.123	0.558	0.648	0.769	0.665	0.786	0.823
Fscore	S1 + GT	0.740	0.865	0.862	0.924	0.941	0.846	0.678	0.713	0.727	0.946	0.812	0.918	0.720	0.772	0.865	0.896	0.952	0.911	0.893	0.729	0.809	0.904	0.834	0.894	0.951
	Mask2Former	0.698	0.703	0.748	0.685	0.660	0.527	0.587	0.304	0.648	0.930	0.506	0.804	0.836	0.743	0.603	0.616	0.627	0.653	0.171	0.555	0.597	0.749	0.624	0.748	0.797
	Deeplabv3+	0.692	0.720	0.747	0.642	0.615	0.459	0.426	0.236	0.624	0.914	0.471	0.773	0.742	0.725	0.557	0.599	0.586	0.619	0.039	0.651	0.555	0.673	0.661	0.812	0.820
	S1 + M	0.634	0.683	0.742	0.678	0.680	0.523	0.409	0.241	0.585	0.915	0.481	0.784	0.637	0.697	0.615	0.679	0.649	0.667	0.123	0.558	0.648	0.769	0.665	0.786	0.823
Precision	S1 + GT	0.855	0.958	0.958	0.977	0.968	0.848	0.777	0.909	0.693	0.942	0.958	0.946	0.848	0.915	0.938	0.943	0.972	0.939	0.959	0.880	0.883	0.953	0.902	0.924	0.971
	Mask2Former	0.662	0.673	0.741	0.670	0.668	0.481	0.678	0.296	0.607	0.929	0.595	0.765	0.865	0.733	0.688	0.604	0.575	0.701	0.233	0.612	0.563	0.727	0.593	0.763	0.802
	Deeplabv3+	0.636	0.741	0.825	0.579	0.658	0.494	0.589	0.320	0.620	0.897	0.520	0.772	0.881	0.713	0.657	0.709	0.651	0.653	0.750	0.698	0.485	0.768	0.611	0.831	0.797
	S1 + M	0.701	0.725	0.860	0.694	0.704	0.520	0.565	0.268	0.566	0.906	0.645	0.768	0.784	0.822	0.764	0.750	0.614	0.744	0.155	0.729	0.694	0.807	0.731	0.854	0.866
Recall	S1 + GT	0.653	0.789	0.784	0.875	0.915	0.843	0.601	0.586	0.764	0.950	0.704	0.892	0.626	0.668	0.804	0.853	0.933	0.885	0.836	0.623	0.747	0.859	0.776	0.867	0.932
	Mask2Former	0.739	0.736	0.756	0.701	0.652	0.583	0.517	0.312	0.696	0.931	0.441	0.847	0.810	0.754	0.537	0.629	0.689	0.611	0.135	0.508	0.635	0.772	0.657	0.733	0.791
	Deeplabv3+	0.760	0.701	0.682	0.722	0.578	0.428	0.334	0.187	0.629	0.932	0.430	0.774	0.641	0.737	0.483	0.519	0.533	0.588	0.020	0.609	0.650	0.599	0.718	0.794	0.844
	S1 + M	0.579	0.646	0.652	0.662	0.657	0.525	0.321	0.220	0.606	0.923	0.383	0.801	0.536	0.605	0.514	0.620	0.687	0.605	0.102	0.452	0.607	0.735	0.610	0.693	0.785

Note. "S1" – SAM (one-pass); "GT" – vote by ground-truth label; "M" – vote by Mask2Former predictions.

have been included in early masks already.

A critical argument for iterative sampling is the combination of diminishing return and exponentially growing compute costs of denser grids in the one heavy pass. As intuitively shown in Figure 3.8, when the uniform sampling grid becomes denser, pixel coverage ratio improves but the increase diminishes. Meanwhile, computation cost increases exponentially because each sampled point must be processed individually and is thus $\mathcal{O}(n^2)$ on the resolution.

More formally, SAM generates high-quality segmentation masks across diverse visual domains with remarkable zero-shot transferability, but its prompt-based architecture inherently limits comprehensive pixel coverage across entire scenes, particularly in complex, feature-dense environments with overlapping objects or ambiguous boundaries. Therefore, SAM's effectiveness in infrastructure element recognition within SLIs is fundamentally tied to cov-

Table 3.10: Overall metrics comparing covered pixels vs. all pixels for S1 + M.

Metric	Covered pixels	All pixels
aAcc	0.912	0.873
mAcc	0.616	0.588
mIoU	0.473	0.469
mDice	0.617	0.601
mPrecision	0.624	0.644
mRecall	0.616	0.588
mFscore	0.617	0.601
aAcc_select	0.843	0.815
mIoU_select	0.490	0.476
mDice_select	0.634	0.627
mPrecision_select	0.657	0.689
mRecall_select	0.614	0.582
mFscore_select	0.634	0.627

Note. "S1" – SAM (one pass), "M" – vote by Mask2Former predictions.

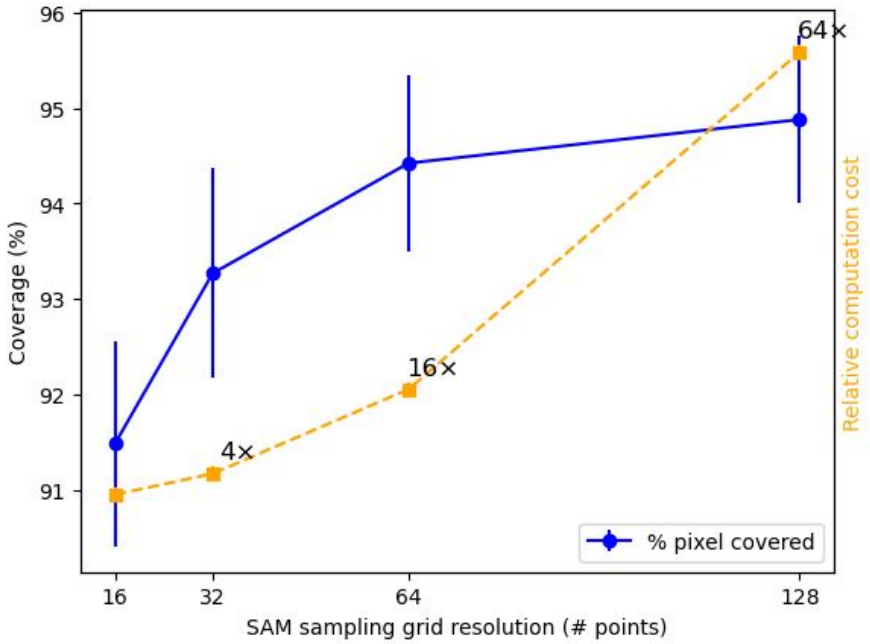


Figure 3.8: Cost of coverage by denser grid sampling increases exponentially, while the benefit shows diminishing returns. As the number of sampling points used by SAM increases, pixel coverage improves from 91% to 95%. However, the relative computational cost grows rapidly (e.g., 64x at 128 points), whereas the marginal gain in coverage becomes increasingly small. This trade-off highlights the need for a smarter iterative sampling approach if high level of coverage must be achieved.

erage of masks it produces. This relationship between mask coverage and detection performance becomes critical when dealing with infrastructure elements that exhibit unconventional morphologies, atypical visual characteristics, and variable scale representations within SLIs. While increasing the density of sampling grid in "everything mode" represents a straightforward approach to enhancing mask coverage within our application, this strategy requires an inherent balance: denser sampling improves coverage and thus assuming masks are of high quality, performance will enhance as well; however, this improvement comes at the expense of substantially increased computational overhead. The computational burden increases exponentially with sampling grid density, which can significantly hinder practical adoption by researchers and practitioners.

To address this tension between coverage and computational efficiency, we propose an iterative sampling methodology based on the idea previously described. This approach adds a loop to SAM module as shown in Figure 3.9. This approach begins with a relatively sparse initial sampling grid to establish reasonably good basic mask coverage. Subsequently, the system samples additional prompt points from uncovered regions, progressively refining the segmentation with minimal addition of sampled points. This adaptive approach maintains high-quality mask coverage critical for infrastructure element detection while simultaneously optimizing the computational efficiency of the SAM implementation in SLI datasets. Contrasting "S1", which stands for masks produced by one forward pass of SAM, we denote masks produced by two forward passes of SAM as "S2".

The sampling grid from the first and second passes look very different (see Figure 3.10) because masks from the first iteration already cover significant number of pixels in the scene (see Figure 3.11. In particular, the second pass adds poles, fences, street lights to captured

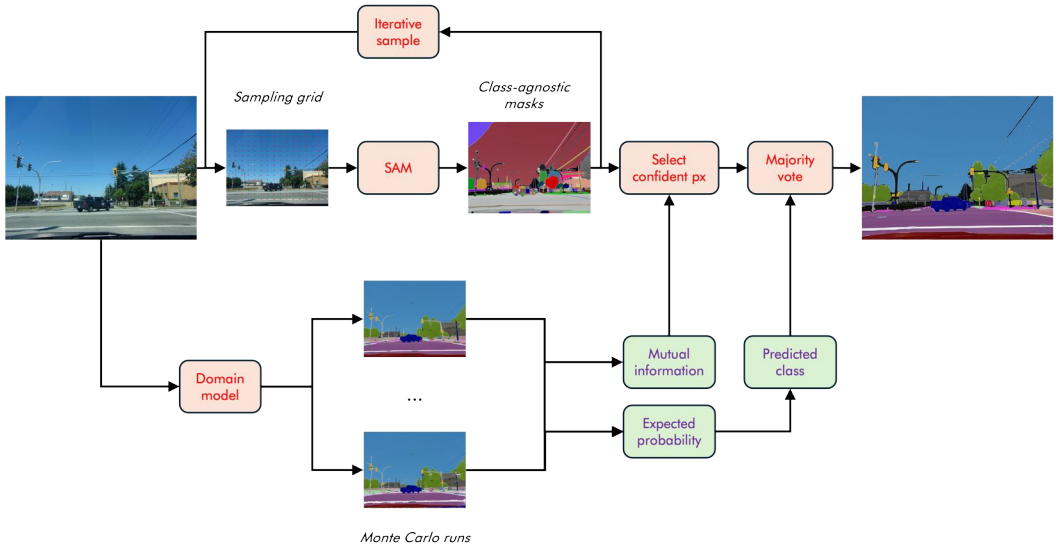


Figure 3.9: Iterative sampling adds a feedback loop to enhance refine spatial prompt selection for improved mask coverage. Building on the previous pipeline Figure 3.2, this design introduces a loop that re-samples additional points in uncovered or uncertain regions based on previous outputs. This iterative process allows Segment Anything Model (SAM) to incrementally improve its coverage with minimal redundancy.

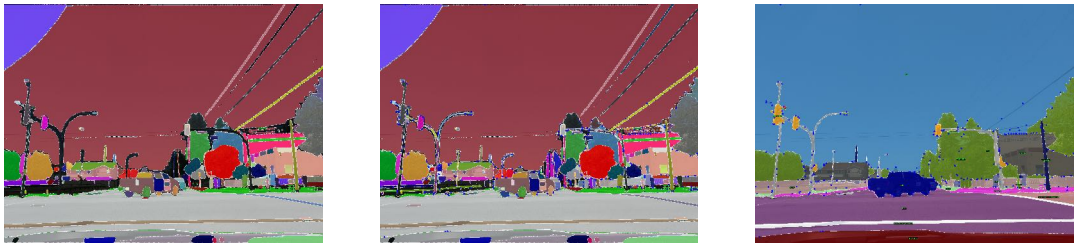
objects and many of them are at very small scale in the SLI.

(a) Sampling grid from 1st pass.

(b) Sampling grid from 2nd pss.

Figure 3.10: Iterative sampling generates adaptive grid patterns across different forward passes of SAM: (a) the first pass builds a standard equally distanced grid to capture coarse objects; (b) the second pass samples on the previously uncovered pixels with an adaptive grid.

This is also reflected in performance metrics. Table 3.11 shows that **S2** + **M** (masks gener-



(a) Masks from the first pass

(b) Final masks and sampling points from the second pass

(c) Final results and sampling points from the second pass

Figure 3.11: Iterative sampling improves mask coverage efficiently by sampling additional points on top of first pass (panel **a**). Sampling of the second pass (panel **b**) targets uncovered or under-segmented regions, producing refined masks and improved coverage (panel **b**). Final segmentation results with combined masks and sampling points from both passes (panel **c**) exhibits enhanced completeness and object delineation with minimal additional cost.

ated by two SAM forward passes + domain vote by Mask2Former) outerforms **S1 + M** significantly in all aspects. Specifically, across metrics that account for all classes in the Mapillary Vistas dataset, **S2 + M** shows consistent improvement of 20% absolute value on metrics covering all object classes. An even more positive trend holds for metrics accounting for selected infrastructure element classes, where **S2 + M** boosts performance past **S1 + M** by almost 30% absolute value: `mIoU_select` from 0.476 to 0.507, `mDice_select` from 0.627 to 0.658, and `mRecall_select` from 0.582 to 0.626. These results suggest that the two-pass SAM strategy followed by domain voting yields better pixel-level coverage and due to SAM's high-quality masks and their boundaries plus trustworthy domain voting, the improved coverage translates to better segmentation results. **S2 + M** also outperforms domain models, showing our proposed strategies to integrate SAM masks and domain model voting are on a promising path. Table 3.12 further provides a breakdown of this leap on selected infrastructure classes.

Table 3.11: Overall metrics comparing S1 + GT vs. baseline domain models vs. **two proposed variants** on **all pixels**.

Metric	S1 + GT (upper bound)	Mask2Former (baseline)	Deeplabv3+ (baseline)	S1 + M (variant 1)	S2 + M (variant 2)
aAcc	0.913	0.914	0.908	0.873	0.891
mAcc	0.825	0.646	0.562	0.588	0.607
mIoU	0.785	0.488	0.448	0.469	0.491
mDice	0.872	0.618	0.576	0.601	0.621
mPrecision	0.932	0.609	0.634	0.644	0.660
mRecall	0.825	0.646	0.562	0.588	0.602
mFscore	0.872	0.618	0.576	0.601	0.621
aAcc_select	0.887	0.852	0.834	0.815	0.847
mIoU_select	0.739	0.494	0.466	0.476	0.507
mDice_select	0.844	0.645	0.614	0.627	0.658
mPrecision_select	0.913	0.649	0.674	0.689	0.694
mRecall_select	0.791	0.647	0.596	0.582	0.626
mFscore_select	0.844	0.645	0.614	0.627	0.658

Note. "S1" – SAM (one pass), "S2" – SAM (two pass), "GT" – vote by ground truth label, "M" – vote by Mask2Former predictions.

3.5 Discussions

This chapter first proposes a framework to integrate SAM's zero-shot mask generation with prior knowledge from domain models trained on road infrastructure element recognition task. This framework is refined with an addition of iterative sampling step, allowing adaptive refinement over initial masks and grids based on the model's evolving confidence and coverage. Results show that while one-pass model performs similarly with baseline domain models, it offers competitive performances on selected object classes that correspond to road infrastructure elements. The two-pass model boosts the performance significantly over its predecessor, by 20% absolute value in metrics accounting for all object classes in the dataset and by 30% absolute value in metrics accounting for infrastructure elements. Other major contributions of the chapter include finding a numerical upper bound of the proposed methodology and articulating the tradeoff between the diminishing return in pixel-level coverage and exponentially increasing computing costs due to rising grid density, which leads to

Table 3.12: Class-wise metrics comparing S1 + GT vs. baseline models vs. **two proposed variants** on **all pixels**.

Metric	Model	Curb	Fence	Guard Rail	Barrier	Wall	Bike Lane	Crosswalk P.	Parking	Rail Track	Road	Service Lane	Sidewalk	LM Crosswalk	LM General	Catch Basin	Fire Hydrant	Junction Box	Manhole	Pothole	Street Light	Pole	Sign Frame	Utility Pole	Traffic Light	Sign Front
IoU	S1 + GT	0.588	0.762	0.758	0.858	0.888	0.733	0.512	0.554	0.571	0.897	0.683	0.849	0.563	0.629	0.763	0.811	0.908	0.837	0.807	0.574	0.679	0.824	0.716	0.809	0.906
	Mask2Former	0.537	0.542	0.598	0.521	0.492	0.358	0.415	0.179	0.480	0.869	0.339	0.672	0.719	0.591	0.432	0.445	0.457	0.485	0.093	0.384	0.425	0.599	0.453	0.597	0.662
	Deeplabv3+	0.530	0.563	0.596	0.473	0.444	0.298	0.271	0.134	0.454	0.842	0.308	0.630	0.590	0.569	0.386	0.428	0.415	0.448	0.020	0.482	0.385	0.507	0.493	0.683	0.694
	S1 + M	0.464	0.519	0.589	0.513	0.515	0.354	0.257	0.137	0.414	0.843	0.316	0.645	0.467	0.535	0.444	0.514	0.480	0.501	0.065	0.387	0.479	0.625	0.498	0.647	0.700
	S2 + M	0.551	0.683	0.748	0.833	0.655	0.420	0.325	0.404	0.415	0.912	0.421	0.514	0.466	0.645	0.567	0.676	0.587	0.321	0.403	0.463	0.445	0.742	0.514	0.803	0.711
Acc	S1 + GT	0.653	0.789	0.784	0.875	0.915	0.843	0.601	0.586	0.764	0.950	0.704	0.892	0.626	0.668	0.804	0.853	0.933	0.885	0.836	0.623	0.747	0.859	0.776	0.867	0.932
	Mask2Former	0.739	0.736	0.756	0.701	0.652	0.583	0.517	0.312	0.696	0.931	0.441	0.847	0.810	0.754	0.537	0.629	0.689	0.611	0.135	0.508	0.635	0.772	0.657	0.733	0.791
	Deeplabv3+	0.760	0.701	0.682	0.722	0.578	0.428	0.334	0.187	0.629	0.932	0.430	0.774	0.641	0.737	0.483	0.519	0.533	0.588	0.020	0.609	0.650	0.599	0.718	0.794	0.844
	S1 + M	0.579	0.646	0.652	0.662	0.657	0.525	0.321	0.220	0.606	0.923	0.383	0.801	0.536	0.605	0.514	0.620	0.687	0.605	0.102	0.452	0.607	0.735	0.610	0.727	0.785
	S2 + M	0.614	0.833	0.813	0.838	0.781	0.690	0.421	0.837	0.685	0.966	0.481	0.873	0.486	0.756	0.641	0.760	0.738	0.332	0.143	0.532	0.632	0.854	0.745	0.880	0.779
Dice	S1 + GT	0.740	0.865	0.862	0.924	0.941	0.846	0.678	0.713	0.727	0.946	0.812	0.918	0.720	0.772	0.865	0.896	0.952	0.911	0.893	0.729	0.809	0.904	0.834	0.894	0.951
	Mask2Former	0.698	0.703	0.748	0.685	0.660	0.527	0.587	0.304	0.648	0.930	0.506	0.804	0.836	0.743	0.603	0.616	0.627	0.653	0.171	0.555	0.597	0.749	0.624	0.748	0.797
	Deeplabv3+	0.692	0.720	0.747	0.642	0.615	0.459	0.426	0.236	0.624	0.914	0.471	0.773	0.742	0.725	0.557	0.599	0.586	0.619	0.039	0.651	0.555	0.673	0.661	0.812	0.820
	S1 + M	0.634	0.683	0.742	0.678	0.680	0.523	0.409	0.241	0.585	0.915	0.481	0.784	0.637	0.697	0.615	0.679	0.649	0.667	0.123	0.558	0.648	0.769	0.665	0.786	0.823
	S2 + M	0.710	0.812	0.856	0.909	0.792	0.735	0.491	0.576	0.730	0.954	0.605	0.679	0.636	0.784	0.724	0.807	0.741	0.486	0.171	0.631	0.616	0.851	0.679	0.891	0.831
Fscore	S1 + GT	0.740	0.865	0.862	0.924	0.941	0.846	0.678	0.713	0.727	0.946	0.812	0.918	0.720	0.772	0.865	0.896	0.952	0.911	0.893	0.729	0.809	0.904	0.834	0.894	0.951
	Mask2Former	0.698	0.703	0.748	0.685	0.660	0.527	0.587	0.304	0.648	0.930	0.506	0.804	0.836	0.743	0.603	0.616	0.627	0.653	0.171	0.555	0.597	0.749	0.624	0.748	0.797
	Deeplabv3+	0.692	0.720	0.747	0.642	0.615	0.459	0.426	0.236	0.624	0.914	0.471	0.773	0.742	0.725	0.557	0.599	0.586	0.619	0.039	0.651	0.555	0.673	0.661	0.812	0.820
	S1 + M	0.634	0.683	0.742	0.678	0.680	0.523	0.409	0.241	0.585	0.915	0.481	0.784	0.637	0.697	0.615	0.679	0.649	0.667	0.123	0.558	0.648	0.769	0.665	0.786	0.823
	S2 + M	0.710	0.812	0.856	0.909	0.792	0.735	0.491	0.576	0.730	0.954	0.605	0.679	0.636	0.784	0.724	0.807	0.741	0.486	0.171	0.631	0.616	0.851	0.679	0.891	0.831
Precision	S1 + GT	0.855	0.958	0.958	0.977	0.968	0.848	0.777	0.909	0.693	0.942	0.958	0.946	0.848	0.915	0.938	0.943	0.972	0.939	0.959	0.880	0.883	0.953	0.902	0.924	0.971
	Mask2Former	0.662	0.673	0.741	0.670	0.668	0.481	0.678	0.296	0.607	0.929	0.595	0.765	0.865	0.733	0.688	0.604	0.575	0.701	0.233	0.612	0.563	0.727	0.593	0.763	0.802
	Deeplabv3+	0.636	0.741	0.825	0.579	0.658	0.494	0.589	0.320	0.620	0.897	0.520	0.772	0.881	0.713	0.657	0.709	0.651	0.653	0.750	0.698	0.485	0.768	0.611	0.831	0.797
	S1 + M	0.701	0.725	0.860	0.694	0.704	0.520	0.565	0.268	0.566	0.906	0.645	0.768	0.784	0.822	0.764	0.750	0.614	0.744	0.155	0.729	0.694	0.807	0.731	0.854	0.866
	S2 + M	0.843	0.791	0.904	0.992	0.802	0.690	0.612	0.439	0.685	0.942	0.723	0.556	0.918	0.814	0.834	0.860	0.734	0.905	0.201	0.782	0.601	0.892	0.624	0.901	0.891
Recall	S1 + GT	0.653	0.789	0.784	0.875	0.915	0.843	0.601	0.586	0.764	0.950	0.704	0.892	0.626	0.668	0.804	0.853	0.933	0.885	0.836	0.623	0.747	0.859	0.776	0.867	0.932
	Mask2Former	0.739	0.736	0.756	0.701	0.652	0.583	0.517	0.312	0.696	0.931	0.441	0.847	0.810	0.754	0.537	0.629	0.689	0.611	0.135	0.508	0.635	0.772	0.657	0.733	0.791
	Deeplabv3+	0.760	0.701	0.682	0.722	0.578	0.428	0.334	0.187	0.629	0.932	0.430	0.774	0.641	0.737	0.483	0.519	0.533	0.588	0.020	0.609	0.650	0.599	0.718	0.794	0.844
	S1 + M	0.579	0.646	0.652	0.662	0.657	0.525	0.321	0.220	0.606	0.923	0.383	0.801	0.536	0.605	0.514	0.620	0.687	0.605	0.102	0.452	0.607	0.735	0.610	0.693	0.785
	S2 + M	0.614	0.833	0.813	0.838	0.781	0.650	0.421	0.837	0.645	0.966	0.481	0.873	0.486	0.756	0.641	0.760	0.738	0.332	0.143	0.532	0.632	0.854	0.745	0.880	0.779

Note. "S1" – SAM (one pass), "S2" – SAM (two pass), "GT" – vote by ground truth label, "M" – vote by Mask2Former predictions.

the iterative sampling proposal.

This proposed methodology advances iRAP's safety assessment by automating the recognition of road infrastructure elements, moving one step closer to reliable and objective star rating computation. The enhanced framework, incorporating iterative sampling, demonstrates strong performance on infrastructure-specific object classes by surpassing traditional domain models. Moreover, masks generated by SAM offer improved interpretability, with cleaner and more coherent boundaries compared to the often fragmented outputs of domain models. In SAM's "everything mode" applied in this chapter, human annotators can quickly surface all visually distinct regions in an SLI without providing prompts or prior setup. This significantly accelerates on-boarding, enabling new annotators to familiarize themselves with the data and begin detailed annotation or measurement.

4

Targeted sampling with multi-modal input

4.1 Introduction

The Segment Anything Model (SAM) can generate high-quality masks without domain knowledge, yet it still depends on external prompts—like points or boxes—to decide what and where to segment. Rather than interpreting its outputs post-hoc as in the previous chapter, this chapter explores SAM's adaption to road infrastructure element segmentation task by

embedding domain-specific knowledge into those prompts from the beginning. This represents a shift from previous strategies, which assign semantic labels to class-agnostic masks using downstream classifiers. Multi-modal input, e.g., text, is transformed into task-informed point prompts by a second foundation model and guides SAM to produce semantically relevant masks during inference. By injecting domain priors into the input, this approach turns on SAM's "anything mode" and maintains its generalizability while aligning its outputs more closely with specialized application needs.

4.1.1 Limitations of post-hoc voting

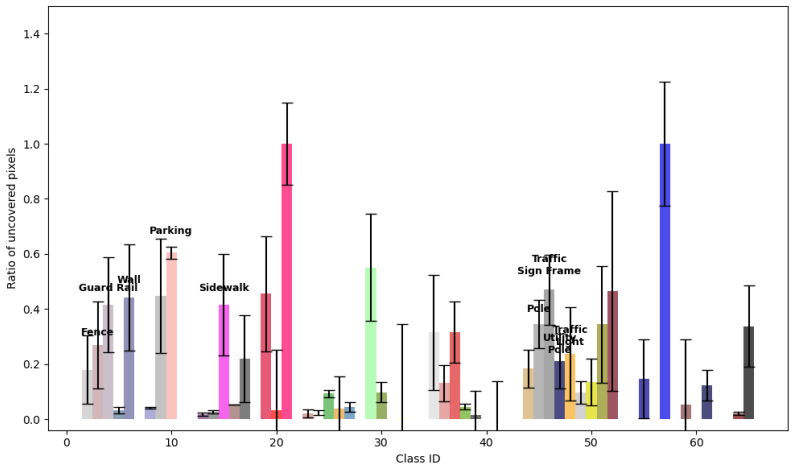
The pipeline proposed in Chapter 3 relies solely on visual input, leverages SAM's "everything mode" to generate class-agnostic masks, which are then interpreted post-hoc. Prior knowledge from domain models—trained specifically on this task and dataset—is used to assign semantic labels to each segmentation mask, effectively acting as classifiers. While this approach achieves strong and competitive performance, it carries two implicit limitations that motivate another path.

The first limitation is pixel-level coverage. While iterative sampling does improve mask coverage efficiently, zero-shot mask generation in "everything-mode" cannot provide any guidance for it. We can only hope that resampling in each iteration will benefit infrastructure classes or hard classes, given the imbalanced pixel coverage ratio. However, the truth is that *thing* classes are easier to be discovered by SAM's masks than *stuff* classes, due to their shapes, scales, and roles in the SLI context. This could lead to imbalanced coverage ratio and imbalanced speed to improve that coverage ratio among infrastructure classes. Figure 4.1 presents evidence to support this speculation. Multiple infrastructure classes suffer from poor cover-

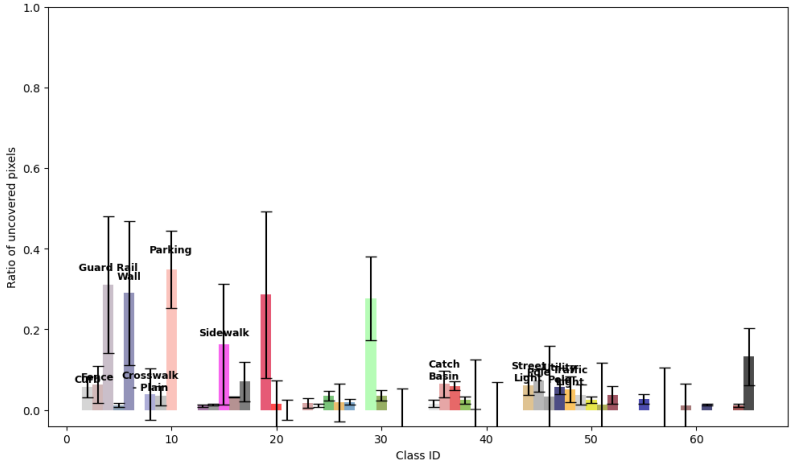
age by SAM's segmentation masks generated by the first forward pass: *parking*, *wall*, *guard rail*, *traffic sign frame*, *sidewalk*, among others. They all have uncovered ratio of at least 40% and they are all *stuff* classes. The second forward pass of SAM as part of the iterative sampling approach improves the situation and the uncovered ratio dropped. However, several infrastructure classes still stand out: *parking*, *guard rail*, *wall*, and *sidewalk*. Improvement of coverage ratio is not equally distributed clearly: while coverage ratio for pixels of *parking* class improves by more than 20% and that for *traffic sign frame* improves by more than 30%, that for *guard rail* and *wall* only improves by 10% or less.

The second limitation is the assumption that domain models trained on the dataset would produce some correctly labeled pixels in the segmentation mask with good confidence and that these confident pixels would dominate the domain voting for the mask. Figure 4.2 provides a counter-example, where two instances of *manhole* are ignored by domain models even though SAM produced accurate masks for them: they are wrongly classified with confidence. This could lead to cases where even with improved coverage, accuracy does not improve accordingly due to domain models' limitation.

These two limitations of iterative sampling motivates a more proactive approach that can potentially kill two birds with one stone: adding another foundation model that enables multi-modal input. The first benefit is more proactive sampling for targeted infrastructure classes by focusing and tailoring prompts on classes of interest in order to boost performances for these classes. Attention to other classes is no longer necessary. The second benefit is reliance on domain models disappears because they are not needed anymore to vote for a segmentation mask's class. It becomes the user's responsibility to pick a class to describe well so that the multi-modal foundation model produces meaningful prompts for segmen-



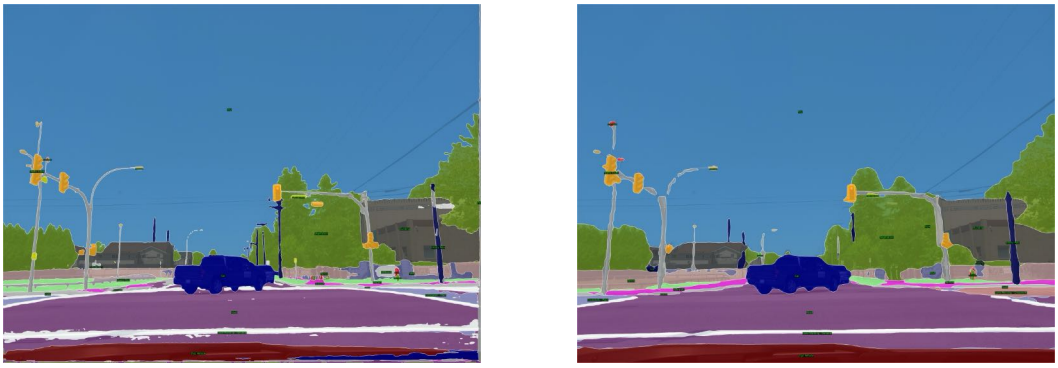
(a) S1: masks generated by one forward pass of SAM



(b) S2: masks generated by two forward passes of SAM

Figure 4.1: Ratio of pixels not covered by SAM's masks are unevenly distributed across classes. Multiple classes belonging to infrastructure elements suffer from poor coverage by masks generated by first pass of SAM (panel **a**): parking, wall, guard rail, traffic sign frame, sidewalk, among others. These classes have an uncovered rate of at least 40%. The second forward pass of SAM (panel **b**) as part of the iterative sampling approach proposed in Chapter 3 significantly improved the situation and the uncovered rate dropped. However, several infrastructure classes still suffer from low pixel-level coverage: parking, guard rail, wall, and sidewalk. Improvement of coverage ratio is not equally distributed either: on average infrastructure classes witnessed around 20% improvement.

tation. A third, bonus advantage is the added interactivity: giving analysts real-time control and feedback strengthens human-computer collaboration, encourages user engagement, and



(a) Deeplabv3+ **(b)** Mask2Former

Figure 4.2: Domain models fail confidently for certain infrastructure classes like *manhole*. Even though SAM has correctly generated mask for two manholes in the SLI as in Figure 3.11, domain models fail to classify any pixels in the region as class *manhole* through MC runs. In this case, the both Deeplabv3+ (panel **a**) and Mask2Former (panel **b**) "confidently" predicted the mask wrongly as *road*. This is a potential limiting factor for the iterative sampling pipeline proposed in Chapter 3.

ultimately accelerates adoption of the safety assessment application. For example in a future date, the annotator could just talk to their computer and the SLI gets segmented for the set of classes for their task. Section 4.2 will introduce in greater detail multi-modal foundation models that can make this happen.

4.1.2 Contribution and outline

This chapter proposes a shift from interpreting SAM's zero-shot masks post-hoc to steering the model with domain knowledge provided as multi-modal input. It first diagnoses two residual flaws of the post-hoc voting pipeline, i.e., poor mask coverage for *stuff* infrastructure classes and reliance on occasionally wrong domain-model labels, to motivate a new direction. The chapter then designs a multi-modal framework: two candidate language-vision backbones are compared, the better aligned model (Grounding DINO) is paired with SAM, and class-specific text is converted into point and box prompts so that only pixels relevant to chosen iRAP assets are sampled. A systematic description template refines each asset's word-

ing, while analytical and empirical tests show how different prompt modes affect SAM, ultimately identifying a setting that surpasses baseline domain models and the earlier "everything-mode" on hard classes by improving IoU and reducing false activations. A pilot study confirms our proposal's real-world transferability, and the discussion distills how this targeted sampling pipeline enables more flexible, automated annotation for iRAP safety assessments.

The rest of this chapter is organized as follows. Section 4.2 starts by introducing two representative text-based foundation models, then analyzes their differences to identify the one more suitable for our task, and finishes by explaining the proposed multi-modal framework. Section 4.3 describes strategies and results to engineer class descriptions as text inputs to the new pipeline, then explores modes of prompts and their combinations for SAM that work the best for our task with mathematical explanations to support observations, then compares with baseline models and "everything mode" from last chapter to demonstrate the efficacy of the new proposal, and closes with a pilot study testing our pipeline. Section 4.4 discusses results obtained from this adaptive sampling approach, points out its implications for flexible and automated asset annotation, and shares its benefits to iRAP road safety assessment procedure.

4.2 Methods

4.2.1 Multi-modal foundation models

Contrastive language-image pre-training (CLIP)[143] and Grounding DINO[116] are both foundation models that marry text and visual inputs. CLIP is a powerful vision-language model that has shown great benefits for various multi-modal tasks (see Figure 4.3). It uses a dual-encoder architecture: one transformer encodes text while another encodes images, projecting

both into a shared embedding space. Its training employs a large set of image–caption pairs collected from the Internet and CLIP learns to align matching image-text pairs while pushing apart mismatched ones using a contrastive loss. During inference, it enables zero-shot recognition by comparing an image's embedding to those of class names or descriptions, without the need for task-specific fine-tuning. Equation 4.1 calculates the cosine similarity between image embeddings and text embeddings

$$sim(v, w) = \frac{v \cdot w}{||v|| \cdot ||w||} \tag{4.1}$$

where v is image embedding vector, w is text embedding vector, $v \cdot w$ is the the dot product between the two vectors, $||v||$ is L2 norm of the image vector and $||w||$ is L2 norm of the text vector. Equation 4.2 shows CLIP's contrastive objective function in order to maximize similarity between semantically similar text-image pairs and to minimize dot product for dissimilar pairs.

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(v_i, w_i)/\tau)}{\sum_{j=1}^{N} \exp(sim(v_i, w_j)/\tau)} \tag{4.2}$$

where v_i is image embedding vector for the ith sample, w_i is text embedding vector for the ith sample, $sim(v_i, w_i)$ is cosine similarity between image and text embeddings as defined in Equation 4.1, τ is temperature parameter (hyperparameter that scales the logits), and N is batch size (number of image-text pairs).

Grounding DINO [116] is an open-vocabulary object detection model that unifies grounding and detection tasks by aligning visual features with free-form text queries. It extends the DINO (DETR with Improved DeNoising training) framework by incorporating text encoders and cross-modal fusion modules to enable grounding capabilities. The model jointly

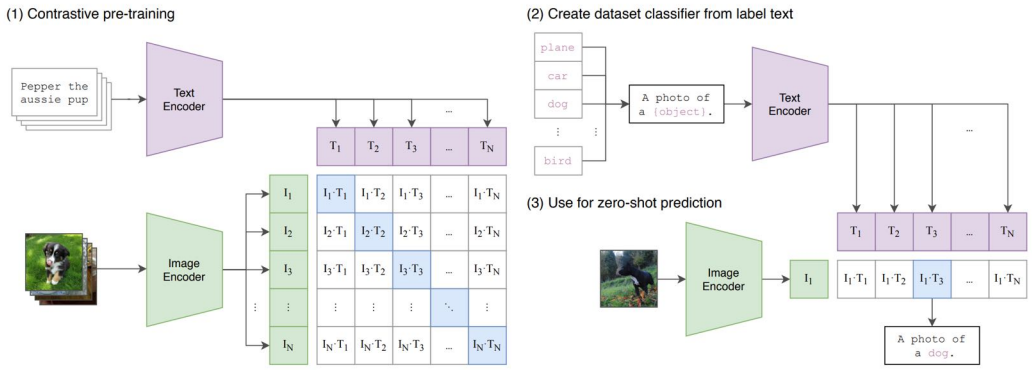


Figure 4.3: CLIP integrates text and image encoders to enable zero-shot image classification. During contrastive pre-training, the text encoder processes textual descriptions, while the image encoder transforms input images into embeddings. The model learns to align these representations in a shared embedding space, where similar text-image pairs are closer together. For dataset classification, CLIP generates text embeddings from class labels, e.g., "a photo of a dog", and compares them with image embeddings for classification. During zero-shot prediction, CLIP finds the closest matching class label for the given image by comparing the image's embedding with the pre-trained class embeddings, enabling classification without additional fine-tuning. Source: Radford *et al.* [143]

learns object localization and text alignment through contrastive and box regression losses, allowing it to predict bounding boxes conditioned on natural language. Its detection loss function (L_{det}) is used at two places within its framework. During the Hungarian matching process, this equation computes the cost matrix for bipartite matching between predicted boxes and ground truth boxes, determining which predictions correspond to which ground truths with typical weight values of $\lambda_{class} = 2.0$, $\lambda_{bbox} = 5.0$, and $\lambda_{giou} = 2.0$. After matching is established, this same form serves as the final detection loss during backpropagation, optimizing the model to produce accurate classifications and well-localized bounding boxes, although the weights may differ slightly with λ_{class} typically set to 1.0 for the final loss calculation.

$$L_{det} = \lambda_{class} \cdot L_{class} + \lambda_{bbox} \cdot L_{bbox} + \lambda_{giou} \cdot L_{giou} \tag{4.3}$$

where L_{det} is the total matching cost for bipartite matching, L_{class} is classification error cost,

L_{bbox} is L1 error of bounding box coordinates cost, L_{giou} is generalized IoU loss cost, λ_{class} is weight for classification cost, λ_{bbox} is weight for bounding box coordinate cost, and λ_{giou} is weight for generalized IoU cost. The Contrastive Loss ($L_{contrast}$) function operates in the language-vision alignment components of the architecture. In the feature enhancer module, it helps enhance image features using text features from a language encoder such as BERT, creating a shared understanding between visual and textual representations. When used in the cross-modality decoder, this loss bridges the gap between vision and language modalities, enabling the model to perform open-set detection by connecting visual features to semantic concepts. This capability is critical for the model to detect objects specified by text prompts that were not explicitly trained on, allowing generalization to novel categories and referring expressions.

$$L_{contrast} = -\log \frac{\exp(sim(f_v, f_t)/\tau)}{\sum_j \exp(sim(f_v, f_t^j)/\tau)} \tag{4.4}$$

where $L_{contrast}$ is the contrastive loss for region-text alignment, f_v is visual feature for a region, f_t is text feature for the corresponding text, f_t^j is text features for all possible texts in the batch, $sim(f_v, f_t)$ is cosine similarity between visual and text features (similar to that used in CLIP), and finally τ is the temperature parameter.

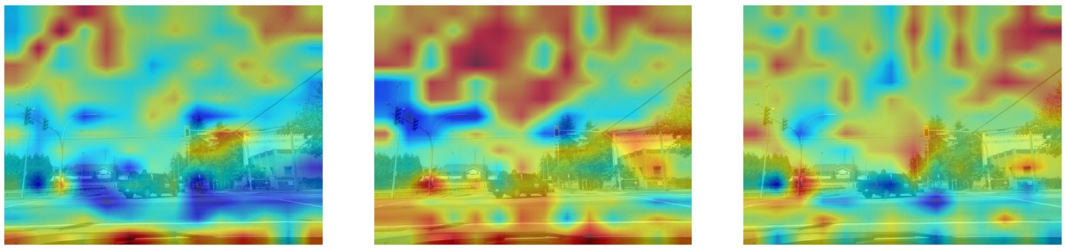
Fundamental differences exist between CLIP and Grounding DINO. Grounding DINO uses more advanced text encoders and thus is able to process longer text descriptions beyond CLIP's 77-token limitation, allowing for more informative prompts. This enables more granular and compositional understanding for Grounding DINO. Grounding DINO provides confidence scores better calibrated for practical thresholding, even though they are still relative rather than absolute, while CLIP's is purely based on cosine similarities and do not have any intrinsic absolute meaning outside of comparisons. They both leverage transformer ar-

chitectures but serve distinct purposes: CLIP focuses on global image-text matching through contrastive learning, while Grounding DINO specializes in open-vocabulary object detection with precise localization capabilities. As a results of these, CLIP tends to prioritize dominant image features while Grounding DINO effectively processes complex scenes with good focus on details. Grounding DINO demonstrates superior ability to distinguish between similar objects within an image and processes more effectively attribute-object relationships and spatial configurations. Figures 4.5, 4.6, and 4.6 present why CLIP is not a good fit for our task due to its similarity measure. Firstly, Figure 4.5 and Figure 4.6 demonstrate CLIP's raw similarity map cannot be used directly due to issues identified by Li *et al.* [112]. Specifically CLIP's similarity map exhibits opposite visualization and noisy activations for provided infrastructure classes in our context. For example, the attention map downplays true pixels that correspond to the provided classes like *sidewalk* and *street light* and chooses to highlight other irrelevant pixels in the background. And because CLIP is pre-trained on images of fixed resolutions like 224 pixels, this issue cannot be addressed by improving resolution. Li *et al.* [112] proposed a variant of CLIP to overcome this issue and their model is found able to generate heatmaps and is able to process at higher resolutions with flexibility. As resolution improves, their variant produces better heatmaps: two coherent pieces of *sidewalk* in the SLI are generally covered by highlighted regions and instances of street lights are pinpointed in the SLI. However, for classes that are not present in SLIs, the heatmap still produces highlighted pixels: regions between electricity cables are highlighted for class *tunnel*. This is due to the fact that CLIP's similarity measure is relative instead of absolute, meaning it must be compared with measures of other classes otherwise they do not represent any physical meaning and cannot be used for thresholding. One idea was to use a fixed set of classes like that defined in the

Mapillary Vistas dataset and applying a softmax layer to transforming those relative similarity measures into probability distributions. However, as Figure 4.6 presents, it still fails by generating high similarity probabilities for classes that do not exist in corresponding SLI.

$$p(y = c|x) = \frac{\exp(sim(f_I(x), f_T(c))/\tau)}{\sum_{c' \in C} \exp(sim(f_I(x), f_T(c'))/\tau)} \tag{4.5}$$

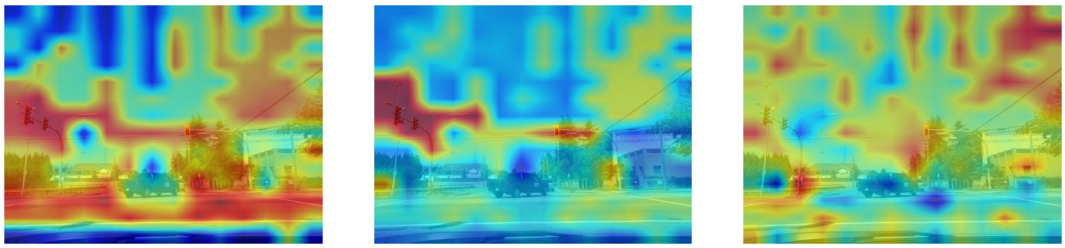
where $f_I(x)$ is the image encoder output for input image x, $f_T(c)$ is the output of text encoder for class label c, C is the set of all possible class labels, τ is temperature parameter, and $sim(\cdot)$ is the cosine similarity function as defined in Equation 4.1.



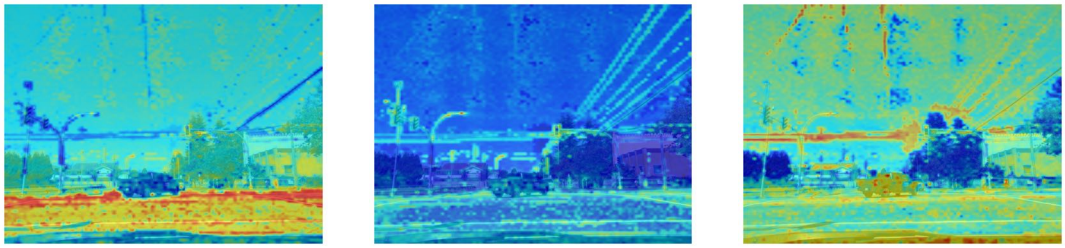
(a) Sidewalk **(b)** Streetlight **(c)** Tunnel

Figure 4.4: CLIP[143] heatmap exhibits opposite visualization and noisy activations for provided infrastructure classes. For example, CLIP's class attention map downplays true pixels that correspond to the provided classes like *sidewalk* (panel **a**) and *street light* (panel **b**) and chooses to highlight other irrelevant pixels in the background. Activations can also be noisy like for *tunnel* (panel **c**). This observation agrees with that as reported in Li *et al.*[112]. Note: CLIP is pre-trained on images of fixed resolutions; the presented heatmaps are generated for 224 px.

In summary, CLIP is not suitable for our task aiming to segment road infrastructure elements from SLIs because its similarity measure is relative and cannot be easily thresholded. Grounding DINO is favored by more than this reason: it is an object detector and its output takes the form of bounding boxes, from each of which activation for the targeted class can be computed. This avoids the complication in CLIP-like models where activations, i.e., similarities, are continuous across the whole space, making it a non-trivial problem to deduce

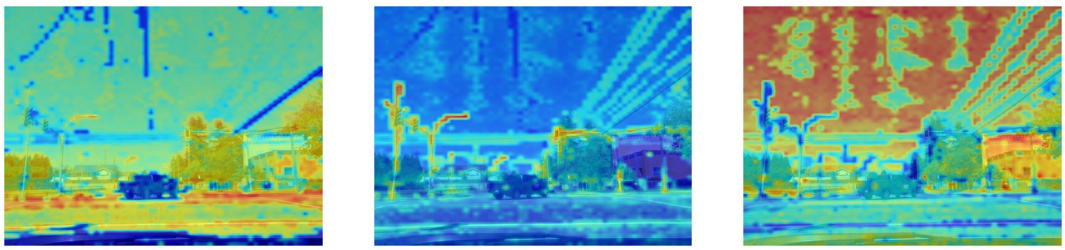


(a) Sidewalk (224 px) **(b)** Streetlight (224 px) **(c)** Tunnel (224 px)



(d) Sidewalk (1600 px) **(e)** Streetlight (1600 px) **(f)** Tunnel (1600 px)

Figure 4.5: CLIP Surgery [112] is able to generate reasonably good heatmap by overcoming issues as seen in Figure 4.4 and is able to process at higher resolutions with flexibility, although this typically requires more computation. As resolution improves, CLIP Surgery's heatmaps become more accurate: two coherent pieces of *sidewalk* (compare panel **a** to panel **d**) in the SLI are generally covered by highlighted regions and instances of street lights are pinpointed in the SLI (compare panel **b** to panel **e**). However, for classes that are not present in SLIs, the heatmap still produces highlighted pixels: CLIP Surgery highlighted regions between electricity cables at 1,600px resolution for *tunnel* (compare panel **c** to panel **f**). This may be due to CLIP's nature of looking for similar features rather than objects.



(a) Sidewalk (1024 px): softmax **(b)** Streetlight (1024 px): softmax **(c)** Tunnel (1024 px): softmax

Figure 4.6: Applying *softmax* to CLIP or CLIP Surgery's similarity map does not effectively turn the similarity measure into absolute values that can be used for selecting confident points via thresholding. While applying softmax function does help with class *sidewalk* (panel **a**), it introduces slightly more false positive (FP) pixels for *street light* (panel **b**) on the poles and significantly more FP pixels for *tunnel* (panel **c**) that are located in the sky and a building.

the number of instances for the object. Any arbitrarily set threshold risks losing or artificially adding instances. For example, there are two instances of *sidewalk* in Figure 4.5(d). On the

one hand, if threshold to apply to the activation map is set too high multiple incomplete instances may surface; on the other hand, if the threshold is set too low, then the activated pixels may not be cut and the map could degenerate to one connected region.

(a) Bounding box **(b)** Segmentation

Figure 4.7: Grounding DINO produces bounding boxes that meet the spatial and semantic alignment of image regions with the provided text, enabling precise localization of objects or concepts described in the input. Stability and confidence thresholds can be used to filter out less confident bounding box predictions, ensuring that only the most reliable and stable box locations are selected based on the consistency of their alignment with the text description. For example, Grounding DINO produces three bounding boxes for three persons in the shadow (panel **a**) and these bounding boxes lead to successful segmentation (panel **b**). Domain segmentation models fail to segment *persons* at a similar quality.

4.2.2 Extract visual activations from multi-modal models

One major motivation for extracting activation maps instead of bounding boxes per class is due to unique characteristics of our object classes, the shapes of which can be non-traditional and thus it is not rare that different infrastructure classes share bounding boxes. For example, *fence* and *sidewalk* at the lower left section in Figure 4.9 share exactly the same bounding box. If this bounding box is provided to SAM, experiments show that *fence* is always identified, but not the *sidewalk*. Therefore, bounding box alone is insufficient for infrastructure element recognition as it contains too much noise, unlike in other applications[198,213].

Grounding DINO offers a concise yet information-rich stream of instance-aware evidence that can be repurposed as high-quality point-prompts for SAM. The last cross-modal decoder layer is the optimal interception point: earlier tensors remain shared across competing hypotheses, whereas later stages, e.g., NMS, have already discarded spatial detail. The decoder, in contrast, realigns these modalities into a set of evolved query embeddings z_q, each of which has converged—through iterative deformable cross-attention—onto a single object–phrase hypothesis. Every query now carries (11) a refined bounding box, (2) a scalar objectness score, and (3) a reference point that anchors it in the image plane

$$\hat{\mathbf{p}}_q = (\hat{x}_q, \hat{y}_q) \in [0,1]^2 \tag{4.6}$$

Interrogating the model at this juncture therefore yields spatial supports that are already instance- and phrase-specific, a critical prerequisite for deriving interpretable prompts.

The decoder's deformable-attention module offers those supports in an analytically convenient form. For each query q and for every feature pyramid level l, the decoder regresses K learnable offsets $\Delta\mathbf{p}_{qli}$ together with corresponding attention weights a_{qli} ($i = 0, \ldots, K-1$). These offsets displace the reference point to yield the *sampling locations*

$$\mathbf{p}_{q\ell i} = \left(\hat{x}_q W_\ell,\ \hat{y}_q H_\ell\right) + \Delta\mathbf{p}_{q\ell i}, \tag{4.7}$$

where (H_l, W_l) denotes the spatial resolution of pyramid level l. In particular, the query's updated feature vector is computed solely from the $K \times L$ pixels accessed at the coordinates

in Equation 4.7

$$\mathbf{z}_q = \sum_{\ell=0}^{L-1} \sum_{i=0}^{K-1} a_{q\ell i}\, \mathbf{x}_\ell\big(\mathbf{p}_{q\ell i}\big), \tag{4.8}$$

with $x_l(\cdot)$ obtained via bilinear interpolation on the level-l feature map. The raw logits are soft-maxed across the $K \times L$ sampling positions so that $\sum_{\ell,i} a_{q\ell i} = 1$; we denote these normalized weights again by $a_{q\ell i}$ for brevity. The set $\mathcal{P}_q = \big\{\mathbf{p}_{qli}\big\}$ thus constitutes an exact, minimal cohort of the pixels that drive the object decision; no other image locations contribute signal to z_q. By construction, this stencil is scale-adaptive, with offset norms growing with object size and shape-aware angular dispersion reflecting object geometry.

Three steps are needed to translate $\mathcal{P}_q$ into SAM point prompts. *Relevance filter* will keep only those sampling points whose attention weights exceed a percentile threshold (e.g., top-30%). This preserves locations the detector itself deems salient, eliminating weak or noisy evidence. *Image-aware re-ranking* then re-scores each surviving point using local image statistics such as Sobel edge magnitude or Laplacian contrast, producing a composite score

$$s_{q\ell i} = a_{q\ell i} \quad \cdot f_{\text{edge}}\big(\mathbf{p}_{q\ell i}\big) \tag{4.9}$$

Empirically, this favors boundary-adjacent samples that guide SAM toward mask outlines. Lastly, in *diversity enforcement* avoids redundant prompts, we impose (1) quadrant coverage within the predicted bounding box and (2) scale coverage across feature pyramid levels. A greedy maximal-marginal relevance rule then selects the top k points $\{\tilde{\mathbf{p}}_j\}_{j=1}^{k}$ that maximize joint saliency while remaining spatially dispersed. For example, sidewalks in SLIs almost always reside in the lower half of the box, so enforcing points from each of the four quadrants may introduce wrong points unnecessarily.

Several lightweight refinements make the workflow fully reproducible. First, we keep Grounding DINO's public defaults of $K = 4$ offsets per feature-pyramid level and $L = 4$ levels overall. Secondly, the offsets $\Delta\mathbf{p}_{q\ell i}$ are predicted in feature-level pixel units and are therefore rescaled by the level's resolution (H_ℓ, W_ℓ) when mapped back to the image plane. Thirdly, the relevance filter retains only those sampling points whose attention weights exceed the 70^{th} percentile of $\{a_{q\ell i}\}$; with the default K and L this leaves approximately ten candidates per object. Fourthly, the image cue is the Sobel gradient magnitude $f_{\text{edge}}(\mathbf{p})$, min–max normalised to the range $[0, 1]$ within the predicted bounding box so that scores remain comparable across images. Finally, the greedy maximal-marginal relevance selector outputs $k \in [3, 5]$ prompts; empirically $k = 4$ improves mIoU by $\sim 1.2\%$ over $k = 3$ while adding only one additional click point.

The resulting $\{\tilde{\mathbf{p}}_j\}$ is typically of cardinality $k \in [3, 5]$, rendering it compact enough for interactive inspection yet rich enough to initialize SAM with high fidelity. Importantly, these prompts preserve the text-conditioned, multi-scale attention patterns of Grounding DINO, thereby transferring semantic focus to SAM without diluting mask-level precision. These organize into Algorithm 1, which leverages Grounding DINO's deformable-attention geometry to distill an object-specific, interpretable set of activation points. By coupling these points with SAM's segmentation functions, we obtain masks that inherit both the detector's open-vocabulary semantics and the segmenter's sharp delineation, providing an end-to-end bridge from phrase grounding to high-quality instance masks suitable for traffic safety assessments and related infrastructure applications.

When we feed $\tilde{\mathbf{p}}_j$ to SAM, the segmenter inherits two complementary strengths: (1) semantic focus, by the prompts lying exactly on pixels the detector deemed relevant to the phrase,

Algorithm 1: Optimal SAM Points Selection

Input: normalized attention weights w, point sets $\{\mathcal{P}_\ell\}_{\ell=0}^{L-1}$, bounding box (x_0, y_0, x_1, y_1), image I, desired number of points k
Output: set $\mathcal{S}$ of k selected points

Function `SelectSAMPoints()`**:**
- $(h, w) \leftarrow$ shape(I);
- $R \leftarrow I[y_0 : y_1,\ x_0 : x_1]$;
- **if** $R = \varnothing$ **then return** $\varnothing$;
- convert R to grayscale and compute Sobel edge map E;
- normalize E to $[0, 1]$;
- $(c_x, c_y) \leftarrow \left(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2}\right)$; `// box center`
- initialize empty list $\mathcal{L}$;
- **for** $\ell \leftarrow 0$ **to** $L-1$ **do**
  - $w_\ell \leftarrow$ level weight;
  - **foreach** $p = (x, y) \in P_\ell$ **do**
    - **if** $x \notin [0, w)$ ***or*** $y \notin [0, h)$ ***or*** dist$(p, box) > thresh$ **then**
      - **continue**
    - $e \leftarrow \begin{cases} E[y, x], & p \in R \\ 0, & \text{otherwise} \end{cases}$;
    - region$(p) \leftarrow$ quadrant of (x, y) w.r.t. (c_x, c_y);
    - $s \leftarrow w(x, y)\, e\, w_\ell$;
    - append $(p, s,$ region$(p), \ell)$ to $\mathcal{L}$;
- sort $\mathcal{L}$ by score s descending;
- initialize $\mathcal{S} \leftarrow \varnothing, \mathcal{R} \leftarrow \varnothing, \Lambda \leftarrow \varnothing$;
- **foreach** $(p, s, r, \ell) \in \mathcal{L}$ **do**
  - **if** $r \notin \mathcal{R}$ **then**
    - add p to $\mathcal{S}$; add r to $\mathcal{R}$; add ℓ to Λ;
    - **if** $|\mathcal{S}| = k$ **then return**;
- **foreach** $(p, s, r, \ell) \in \mathcal{L}$ **do**
  - **if** $\ell \notin \Lambda$ **then**
    - add p to $\mathcal{S}$; add ℓ to Λ;
    - **if** $|\mathcal{S}| = k$ **then return**;
- **foreach** $(p, s, r, \ell) \in \mathcal{L}$ ***while*** $|\mathcal{S}| < k$ **do**
  - **if** $p \notin \mathcal{S}$ **then**
    - add p to $\mathcal{S}$;

so SAM starts from the correct region even in cluttered scenes; (2) geometric precision, by edge-aware re-ranking and cross-scale diversity ensuring that the few selected points flank prominent boundaries at multiple resolutions, allowing SAM's mask refinement to "snap" to object borders. Across a range of experiments we find that three to five prompts suffice to match or exceed the mask quality of far more complex prompting heuristics, while remaining small enough for visual inspection or downstream auditing. The resulting pipeline therefore

bridges phrase grounding and open-set segmentation in a manner that is both *principled* (the math tells us exactly where the evidence comes from) and *practical* (no heavy hyper-parameter tuning is required). In the context of traffic safety assessments, for example, these masks allow analysts to recover curb edges, guard rails, and pedestrian crossings with little manual intervention, accelerating large-scale iRAP Star Ratings workflows.

4.2.3 Integrate SAM with multi-modal foundation model

Putting all components together, Figure 4.8 renders the full, class-targeted segmentation workflow. A concise yet informative description of the infrastructure class of interest, e.g. "Sidewalk – pedestrian path adjacent to roads; elevated concrete/brick surface, usually lighter than the carriageway", is engineered upstream. This phrase functions as an open-vocabulary referring expression that constrains subsequent detection to only those regions plausibly matching the class semantics. The SLI and the formatted description are jointly ingested by Grounding DINO. Its cross-modal decoder produces two complementary outputs: (1) a set of bounding boxes after NMS with objectness scores, and (2) multi-scale deformable-attention offsets that identify a handful of high-saliency sampling points per decoder layer. After normalizing and thresholding the attention weights, the top-K points defining the most discriminative pixels for each box are retained and re-projected to absolute image coordinates. The attention peaks are converted to point prompts $(x, y, \texttt{label})$, with the label set to "foreground". We experimented with three prompt configurations: *box only*, *point only*, and *box + point*. Counter-intuitively, supplying both simultaneously degrades class precision—the coarse box overrides the finer point cues, causing SAM to flood-fill the entire rectangle and largely ignore the prompted points. Quantitatively and qualitatively, the *point-only* variant yields the tight-

est, class-consistent masks, so the final pipeline drops the bounding-box prompt and retains only the attention-derived points when targeting individual infrastructure elements. In summary, Figure 4.8 exemplifies how language-conditioned detection (Grounding DINO) and prompt-driven segmentation (SAM) can be coupled: from text to boxes, from attention focus points to point prompts and to class-specific masks. The study further reveals that, for SAM, point seeds convey richer geometry than simultaneous box plus point prompts, and should therefore be preferred when high-fidelity infrastructure segmentation is required.

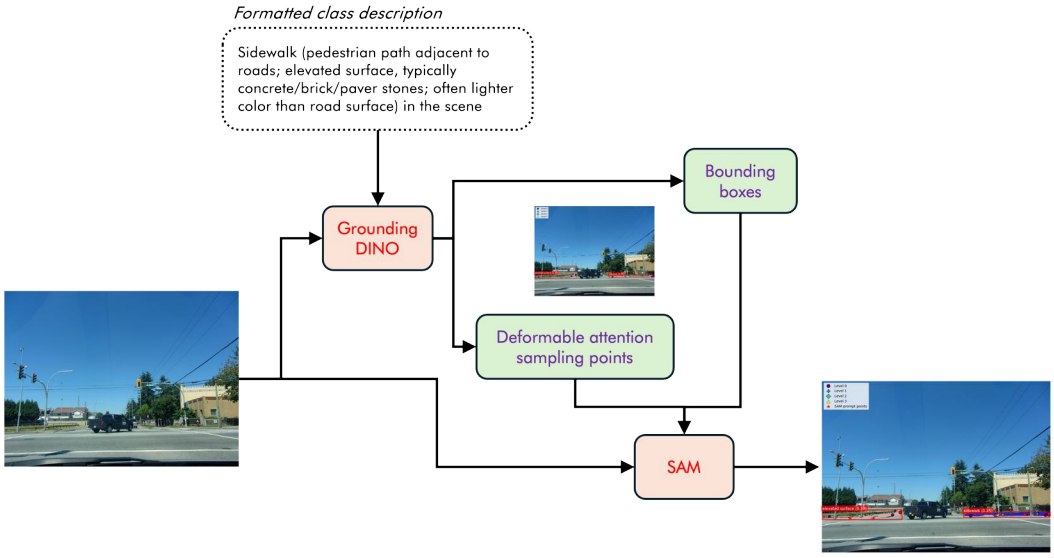


Figure 4.8: Multi-modal input facilitates targeted prompting and efficient segmentation for infrastructure classes in SLIs. The proposed pipeline starts with class description properly formatted by length, i.e., number of tokens, and by focus on visual hints of the objects. Grounding DINO ingests this text input with the SLI and produces bounding boxes aligned with parts of the description, possibly single or multiple words. Sampling points of the deformable attention mechanism is extracted from decoder layers of the model and selects the most representative points. SAM uses these selected points as well as the bounding boxes to produce segmentation masks for the target class in SLI.

4.3 Results

4.3.1 Engineered descriptions

Class descriptions must be carefully designed and formatted for three reasons. First, providing detailed and context-rich descriptions increases the likelihood that foundation models find correct associations. Class names using domain-specific terminologies may not be directly recognizable by general-purpose models. For example, *pole* and *utility pole* appear as separate classes in the Mapillary Vistas dataset, and without class-specific descriptions, foundation models may conflate them due to the shared term *pole* in their names. Second, descriptions from different sources depict the object class from different angles and thus may not be consistent. Some may contain too many nouns mentioning other concepts in the dataset, e.g., "traffic light is signal device controlling car and pedestrian flow and is typically installed at intersections on poles with street lights"; some others may be providing insufficient visual clues, e.g., "fire hydrant is water outlet typically used by fire-fighters"; yet some others are not designed for this purpose in the first place, e.g., class descriptions in Mapillary Vistas paper[131] contain many annotation instructions prepared for human annotators like "curb: raised (>3cm) formation typically between road and sidewalk or terrain; includes the brick's horizontal and vertical surface; if there is no clear border between sidewalk/terrain and curb, only the vertical surface is labeled". Lastly, certain foundation models are pre-trained with inflexible settings and thus require lengths of description be kept under number of token requirements, e.g., CLIP's hard rule is 77 tokens.

A good template prioritizes visual clues such as color, shape, placement in SLI, and materials, downplays real-world usage, and removes unnecessary texts to keep lengths within re-

quirement: `[key visual attributes] + [position info] + [distinctive features]`. Results of formatting all classes in Mapillary Vistas are presented in Table 4.1.

Table 4.1: Class names and descriptions before/after formatting.

Class	Mapillary Vistas	Proposed according to template
Curb	Raised (>3cm) formation typically between road and sidewalk or terrain. Includes the brick's horizontal and vertical surface. If there is no clear border between sidewalk/terrain and curb, only the vertical surface is labeled	Elevated boundary (>3cm) between road and sidewalk. Concrete/stone ridge with vertical face and flat top. Gray, forming continuous line along streets. Creates visible shadow line.
Fence	Freestanding structure with holes that separate two (or more) outdoor areas, sometimes temporary ones	Freestanding structure with gaps/holes separating outdoor areas. Various materials (chain-link, wood, metal). Typically linear with regular pattern.
Guard Rail	Metal structure mounted on the side of the road to prevent serious accidents. Includes the bars holding the rails	Metal protective structure along roadside. Horizontal silver/gray rails with vertical supports. Often curved or W-shaped profile.
Barrier	Fallback category for other forms of barriers, e.g. lane separator, jersey barrier	Temporary or permanent divider between areas. Includes concrete jersey barriers, plastic dividers, bollards. Often brightly colored or reflective.
Wall	Individually standing walls that separate two (or more) outdoor areas, and do not provide support for a building. Includes transparent walls	Freestanding vertical structure separating outdoor areas. Solid surface without openings. Brick, concrete, or stone. Not supporting buildings.
Bike Lane	Paved areas intended for bicycle use only. Does not include regions meant to be used by both cyclists and pedestrians. Regions of road intended for mixed cycle and car use are labeled as road. Contains only regions of ground specifically designated for cyclists where pedestrians and cars should not enter, e.g. separated by curb or lane markings (dashed or driveable)	Designated path for bicycles. Often colored green/red or marked with bicycle symbols. Separated by painted lines or physical barriers from vehicular lanes.
Crosswalk - Plain	Region of a road which is meant to cross the street but not marked with a zebra pattern, instead two parallel lines indicate the crosswalk. The road surface is marked as plain crosswalk, the two lines indicating the border are marked as general markings	Road crossing area marked by two parallel lines without zebra pattern. Connects sidewalks across streets for pedestrian crossing.
Curb Cut	Lowered curbs (≈3cm) or curb cuts (<3cm) which can be driven over, e.g. by wheelchairs	Lowered section (<3cm) of curb forming ramp. Creates smooth transition between road and sidewalk. Often at intersections and driveways.
Parking	surfaces that are intended for parking and separated from the road either via elevation or different texture/material or lane markings. Only the area which is available for parking has to be annotated. Flat areas used to get on a specific parking slot need to be labeled as road	Designated vehicle storage area. Separated from road by lines, texture differences or elevation. Often marked with P signs or painted boundaries.
Pedestrian Area	Areas solely to be used by pedestrians. E.g. in inner city areas or parks, shopping zones, etc	Paved zones exclusively for walking. Often brick/tile surfaces in plazas, malls, parks. No vehicle access. May have benches, planters.
Rail Track	Horizontal surface on which rail cars typically drive. If rail-tracks are embedded in a conventional road block, the rail track is drawn on top of a road label	Linear metal rails for trains/trams. Parallel metal strips embedded in or raised above ground. Often with wooden/concrete ties between rails.
Road	Surface on which cars usually drive on. Typically delimited by curbs, parking areas or bicycle lanes	Paved surface for vehicles. Asphalt/concrete, typically dark gray with lane markings. Bordered by curbs, sidewalks or shoulders.

Table 4.1 – continued

Class	Mapillary Vistas	Proposed according to template
Service Lane	Road shoulders, typically on highways, which offer enough space for a car to drive and stop in emergency cases. Usually on the right side of drivable lanes (viewed in driving direction, in case of right-hand traffic), but can also be on the left side	Auxiliary roadway alongside main highway. Narrower than main lanes, often separated by markings. For emergency stopping or service vehicles.
Sidewalk	Horizontal surface designated for pedestrians or cyclists. Delimited from the road by some obstacles, e.g. curbs (raised or lowered) or poles, but not only by marking. Often elevated compared to the road and often located at the side of a road. Also includes walkable parts of traffic islands, but not pedestrian areas	Pedestrian path adjacent to roads. Elevated surface, typically concrete/brick/paver stones. Often lighter color than road surface.
Bridge	Bridges (on which the ego-vehicle is not driving) including everything (fences, guard rails) permanently attached to them	Elevated structure spanning gap/waterway. Visible supports/railings. Includes attached elements like fences and guard rails.
Building	Includes structures that house/shelter humans, e.g. low-rises, skyscrapers, bus stops, carports. Translucent buildings made of glass still receive the label building. Also includes scaffolding attached to buildings	Permanent structure with walls/roof. Various sizes from sheds to skyscrapers. Includes glass structures, bus shelters, and attached scaffolding.
Tunnel	Tunnel walls and the dark space encased by the tunnel, but excluding vehicles	Enclosed passage through obstacle. Dark interior with walls/ceiling. Artificial lighting and ventilation systems visible.
Lane Marking - Crosswalk	Crosswalk indicated by a zebra pattern	Road marking with alternating white stripes (zebra pattern). Perpendicular to traffic flow, connecting sidewalks.
Lane Marking - General	Other lane markings including dashed and solid lane separation markings, stop lines, text, arrows or restricted areas	Painted road symbols including lines, arrows, text, stop lines. White or yellow, painted directly on road surface.
Catch Basin	Sewage water drainage, typically with a grid on top or inlet of a curb	Street drainage opening. Metal grate or curb inlet. Rectangular or round shape embedded in road/sidewalk.
Fire Hydrant	Water outlet typically used by fire-fighters	Water connection for firefighting. Compact metal fixture on sidewalks. Typically red/yellow, mushroom or pillar shaped with caps/outlets.
Junction Box	An electrical junction box can contain junctions of electric wires and/or cables	Utility cabinet housing electrical connections. Metal/plastic box mounted on ground or pole. Often gray/green with access panel.
Manhole	Opening in road to access canal system. Has to be big enough to fit a person and includes squared ones	Access cover to underground utilities. Round/square metal plate with patterns set into road/sidewalk. Often has text/municipal emblems.
Pothole	A depression or hollow in a road caused by wear or subsidence	Road surface depression/damage. Irregular dark hole in pavement. Often filled with water after rain. Exposes underlying road layers.
Street Light	Lights mounted on poles to illuminate the road or public places	Tall pole with mounted illumination fixture. Extends above roads/sidewalks. Light housing at top, sometimes with multiple heads/arms.
Pole	Thin and elongated, typically vertically oriented poles, e.g. sign or traffic light poles. Does not include objects mounted on the pole. Poles holding only street lights should also be labeled as pole	Slender vertical support structure. Metal/wood, cylindrical/rectangular. Used for signs, lights, without mounted objects.
Traffic Sign Frame	Frame construction holding large signs, where frame can be made of round or square poles. Can be typically found on highways	Large metal structure supporting road signs. Multiple poles supporting wide sign panels. Common on highways/intersections.
Utility Pole	Utility poles are supposed to carry electrical wires/cables, could be used to mount mobile phone network receivers, and are used in case both, cables and street lights are mounted	Tall wooden/concrete pole carrying wires. Has electrical cables, transformers, sometimes with street lights or communication equipment.

Table 4.1 – continued

Class	Mapillary Vistas	Proposed according to template
Traffic Light	Traffic lights without poles in different orientations (upright, horizontal, side, back, front) and for all types of traffic participants, e.g. regular traffic light, bus traffic light, train traffic light. If the type can be determined but does not match any of the defined categories (e.g. traffic lights for trains), then traffic light - other is used. If the orientation or type of traffic light cannot be determined, traffic light - ambiguous is used	Red/yellow/green lights in housing. Various orientations, without support pole.
Traffic Sign (Back)	Back side of traffic signs	Signal device controlling traffic flow. Rear side of road sign. Plain metal surface, usually gray/galvanized. Visible mounting brackets/supports.
Traffic Sign (Front)	Signs with the purpose of conveying information to drivers/cyclists/pedestrians, e.g. traffic signs or warning reflector posts. If the type of a frontviewed traffic sign cannot be determined, e.g. because of unknown language, traffic sign - unknown is used as fallback	Information display for road users. Distinctive shapes/colors (red octagons, yellow diamonds, etc). Contains symbols/text.
Trash Can	Recycling facility for waste collection	waste receptacle in public areas. Cylindrical/rectangular container with opening/lid. Various colors, often secured to ground/poles.

Visual evidence in Figure 4.9, Figure 4.10, and Figure 4.11 confirms that the text-engineering step, i.e., rewriting class labels into compact but semantically rich phrases, has a first-order effect on the whole pipeline from Grounding DINO to point prompts and to SAM targeted sampling. Across all three classes, the proposed descriptions consistently induce fewer but better-localized bounding boxes, denser cross-attention sampling on the object interior, and cleaner SAM masks that respect fine object boundaries. In contrast, prompting with the plain class name or with the Mapillary Vistas glossary either fails to retrieve small objects such as *manhole* or triggers large, off-target boxes that dominate the decoder's attention such as for *sidewalk* and *guard rail*, ultimately starving SAM of useful point evidence. Using only the label "sidewalk" Grounding DINO fires a single, thin box at the lower right corner, missing the wider concrete strip on the left. The Mapillary Vistas description of *sidewalk* introduces the highly salient token "road", causing the model to output a blanket box that spans the car-

riageway; the associated SAM mask consequently covers the entire pavement. Our formatted prompt "concrete pedestrian walkway parallel to the curb" suppresses competing road cues, yields two tight boxes that cover both pieces of sidewalk, and sends a well-distributed set of level-0/1/2 sampling points to SAM. The resulting mask tracks the curb line almost perfectly without leaking onto the asphalt. Manholes are small, low-contrast, and often occluded. With the raw token "manhole", the detector returns only one of the two covers in the scene. Embedding material and shape cues "circular cast-iron sewer cover" doubles the recall: Grounding DINO now isolates both discs despite strong surrounding edge clutter, and the interior-biased attention points lead SAM to segment the metallic lids while ignoring adjacent road markings. The definition provided in Mapillary Vistas, again containing the word "road", introduces an additional, spurious box that envelopes the entire lane and degrades the mask IoU by flooding SAM with irrelevant background pixels. Contrasting *sidewalk* and *manhole*, *guard rails* pose yet another challenge: they are long, thin, and specular, indicating they can be easy to detect at low resolution but hard to segment precisely. Although the plain label already returns two approximate boxes, the cross-attention points concentrate near the road centerline; SAM therefore mis-fires, producing an under-segmented mask. The Mapillary sentence generates yet another oversized *road* box and completely derails the pipeline. By contrast, the refined expression "metal safety barrier with parallel horizontal beams" steers attention onto the rail itself: bounding boxes hug the full rail length, level-2 points sprinkle the entire span, and SAM outputs a contiguous mask with only minor bleeding into adjacent vegetation. These case studies demonstrate that prompt wording is a controllable, high-leverage knob for open-vocabulary detectors: carefully crafted descriptions eliminate lexical ambiguity, align textual tokens with the visual granularity of the target object, and thereby

maximize the quality of both the detector's bounding hypotheses and the downstream SAM mask quality. Quantitative gains are thus underpinned by clear qualitative improvements: better instance recall for small objects, tighter spatial precision for elongated structures, and a dramatic reduction of false positives stemming from scene-dominant but irrelevant context such as the road surface.

4.3.2 Mode of prompt

Figure 4.12 shows point prompts alone can produce segmentation masks of good quality, and adding box prompts risks overriding the visual cues that point prompts contain. Given box prompts alone are not sufficient to distinguish infrastructure elements, point prompts is the source we need for recognizing road infrastructure elements in SLIs via adaptive and targeted sampling. SAM digests every prompt—points, boxes, or masks—through the prompt encoder and then fuses the resulting tokens with the pre-computed image embedding in the mask decoder transformer. The crucial detail is that box prompts are treated as a very strong spatial prior. First, SAM supplements the usual four box-corner tokens with a dense positional kernel that is added to every pixel inside the box before decoding; this shifts the transformer's attention toward whatever lies in the boxed region. Second, after decoding, SAM hard-gates the logits so that any activation falling outside the box is forced to zero. Together, these mechanisms give the box a spatial prior that is magnitude-level stronger than the sparse point tokens, effectively overriding contradictory point evidence. In the context of Figure 4.12, the imbalance manifests as follows: when only points are supplied, the model respects the fine-grained cues and returns a clean mask. Adding an imprecise bounding box crops the image embedding, injects the dense kernel, and finally suppresses out-of-box log-

its, so the decoder converges on the coarser and in this case incorrect region. The resulting mask reflects the box's geometry rather than the semantic hints encoded by the point set, even when those points carry high confidence.

Table 4.2: Distinctions between point-prompt and box-prompt processing pathways in SAM.

property	point-prompt path	box-prompt path
Token strength	1 token / point, no dense prior	4 tokens + a *dense* "box kernel" that is added at every pixel inside the box
ROI cropping	None (full 1024×1024 embedding)	the image embedding is window-cropped to the box +16-px margin; pixels outside are never seen by the decoder
Final mask gating	No hard gating	logits outside the box are zeroed before the final sigmoid

Table 4.2 presents the architectural biases that make SAM place disproportionate trust in box prompts. During inference the model adds a dense box kernel K to every pixel inside the proposed region and applies a hard spatial gate that nullifies all mask logits outside that region, both *before* the point–derived tokens have a chance to steer cross–attention in the decoder. Consequently, even a mildly mis-aligned box or one that encloses ambiguous textures will dictate the final segmentation, over-ruling a large set of high-confidence point cues. From a high level, the prompt encoder outputs a sparse token set S containing one vector per point and four vectors for the box corners, and a dense positional kernel K injected at every pixel that satisfies $(x, y) \in$ box. The decoder therefore attends to $\tilde{F} = F + K$, where F is the frozen image embedding, while receiving the composite token sequence Z_0. Because $\|K\|_2 \gg \|S_i\|_2$ for any point token S_i, and because logits outside the box are clamped to $-\infty$, the box prior dominates; points can refine the mask only within the boxed window, but cannot override an erroneous box altogether. These two mechanisms, i.e., dense biasing *and* post-decoder gating, explain why an imprecise box overrides otherwise informative point prompts, as observed in Fig. 4.12.

4.3.3 Comparison with "everything mode"

Table 4.3 demonstrates the benefit of proposed paradigm shift from post-hoc voting on SAM's segmentation masks to the novel proactive sampling strategy targeted at specific infrastructure classes. This advancement yields consistent performance improvements across all major metrics, with particularly notable gains in `mIoU` (+1.1%) and `mDice` (+0.4%) compared to S2 + M. Unlike previous evaluations, we deliberately focused on infrastructure-relevant classes rather than the entire Mapillary Vistas taxonomy, because with the new proactive pipeline there is no need to go through their segmentation any more.

The class-wise breakdown in Table 4.4 further reveals improvements across different infrastructure elements. Substantial performance gains are observed on: *fence*, *barrier*, *wall*, *sidewalk*, *lane marking - crosswalk*, *fire hydrant*, *manhole*, *pothole*, *street light*, *traffic light*, and *traffic sign front*. Among these, three classes exhibited exceptional improvement: *sidewalk* (+14.9% in Dice), *manhole* (+32.9% in Dice), and *sign front* (+8.4% in Dice). These solid improvements suggest these classes previously suffered from misguided confidence in domain voting approaches, where conventional models erroneously enforced semantic coherence at the expense of boundary precision. The proposed method shows compelling results for historically challenging infrastructure elements. For instance, manholes, small but critical components of urban infrastructure, were either entirely missed or poorly segmented by previous approaches. The multi-modal targeted model not only detects these elements but also does so with high performance, demonstrating the value of targeted sampling in practical infrastructure mapping applications. Crosswalk-related classes benefited in a different way. Previously they are hard to segment due to two distinct challenges. For *Crosswalk - Plain*, which marks crosswalk region by parallel lines without zebra patterns, SAM

frequently merges these markings with the broader road surface when generating masks. For *Lane Marking - Crosswalk* that has zebra patterns, each individual stripe is typically captured by separate sampling points in previous grid-based approaches, failing to connect these elements into cohesive crosswalk entities. The new approach seems to address this issue with the bounding box that Grounding DINO produces and our strategy to select representative points to prompt SAM, as in Algorithm 1. These point prompts from big bounding boxes must have encouraged cohesive masks that bridged gaps individual point prompts fail to address. Our targeted approach successfully addresses both threads, suggesting that intelligent sampling strategies can overcome limitations in foundation model segmentation.

The overall results show that proactive and targeted sampling provides an good angle for infrastructure segmentation than reactive voting schemes applied to class-agnostic foundation model outputs. This approach not only improves aggregate metrics but demonstrates particular efficacy for some of the most challenging and safety-critical infrastructure elements.

Table 4.3: Overall metrics comparing multi-modal targeted sampling (proposed) vs. S2 + M (Chatper 3) vs. baseline domain models on **all pixels** and selected infrastructure classes.

Metric	Mask2Former (baseline)	Deeplabv3+ (baseline)	S2 + M (Chatper 3)	Multi-modal targeted (Proposed)
aAcc_select	0.852	0.834	0.847	0.892
mIoU_select	0.494	0.466	0.507	0.518
mDice_select	0.645	0.614	0.658	0.662
mPrecision_select	0.649	0.674	0.694	0.698
mRecall_select	0.647	0.596	0.626	0.629
mFscore_select	0.645	0.614	0.658	0.662

Note."S2" – SAM (two pass), "M" – vote by Mask2Former predictions.

Table 4.4: Class-wise metrics comparing multi-modal targeted sampling (proposed) vs. S2 + M vs. baseline domain models on **all pixels**.

Metric	Model	Curb	Fence	Guard Rail	Barrier	Wall	Bike Lane	Crosswalk P.	Parking	Rail Track	Road	Service Lane	Sidewalk	LM Crosswalk	LM General	Catch Basin	Fire Hydrant	Junction Box	Manhole	Pothole	Street Light	Pole	Sign Frame	Utility Pole	Traffic Light	Sign Front
IoU	Mask2Former	0.537	0.542	0.598	0.521	0.492	0.358	0.415	0.179	0.480	0.869	0.339	0.672	0.719	0.591	0.432	0.445	0.457	0.485	0.093	0.384	0.425	0.599	0.453	0.597	0.662
	Deeplabv3+	0.530	0.563	0.596	0.473	0.444	0.298	0.271	0.134	0.454	0.842	0.308	0.630	0.590	0.569	0.386	0.428	0.415	0.448	0.020	0.482	0.385	0.507	0.493	0.683	0.694
	S2 + M	0.551	0.683	0.748	0.833	0.655	0.420	0.325	0.404	0.415	0.912	0.421	0.514	0.466	0.645	0.567	0.676	0.587	0.321	0.403	0.463	0.445	0.742	0.514	0.803	0.711
	Multi-modal targeted	0.695	0.775	0.742	0.865	0.731	0.231	0.293	0.472	0.525	0.918	0.565	0.683	0.625	0.631	0.267	0.712	0.685	0.685	0.423	0.435	0.452	0.825	0.518	0.865	0.748
Acc	Mask2Former	0.739	0.736	0.756	0.701	0.652	0.583	0.517	0.312	0.696	0.931	0.441	0.847	0.810	0.754	0.537	0.629	0.689	0.611	0.135	0.508	0.635	0.772	0.657	0.733	0.791
	Deeplabv3+	0.760	0.701	0.682	0.722	0.578	0.428	0.334	0.187	0.629	0.932	0.430	0.774	0.641	0.737	0.483	0.519	0.533	0.588	0.020	0.609	0.650	0.599	0.718	0.794	0.844
	S2 + M	0.614	0.833	0.813	0.838	0.781	0.690	0.421	0.837	0.685	0.966	0.481	0.873	0.486	0.756	0.641	0.760	0.738	0.332	0.143	0.532	0.632	0.854	0.745	0.880	0.779
	Multi-modal targeted	0.835	0.875	0.895	0.875	0.845	0.525	0.487	0.455	0.715	0.954	0.685	0.842	0.758	0.803	0.715	0.990	0.815	0.765	0.412	0.756	0.595	0.895	0.619	0.962	0.915
Dice	Mask2Former	0.698	0.703	0.748	0.685	0.660	0.527	0.587	0.304	0.648	0.930	0.506	0.804	0.836	0.743	0.603	0.616	0.627	0.653	0.171	0.555	0.597	0.749	0.624	0.748	0.797
	Deeplabv3+	0.692	0.720	0.747	0.642	0.615	0.459	0.426	0.236	0.624	0.914	0.471	0.773	0.742	0.725	0.557	0.599	0.586	0.619	0.039	0.651	0.555	0.673	0.661	0.812	0.820
	S2 + M	0.710	0.812	0.856	0.909	0.792	0.735	0.491	0.576	0.730	0.954	0.605	0.679	0.636	0.784	0.724	0.807	0.741	0.486	0.171	0.631	0.616	0.851	0.679	0.891	0.831
	Multi-modal targeted	0.752	0.845	0.872	0.912	0.835	0.375	0.512	0.593	0.685	0.953	0.723	0.828	0.765	0.776	0.583	0.845	0.812	0.815	0.456	0.645	0.618	0.902	0.675	0.952	0.915
Fscore	Mask2Former	0.698	0.703	0.748	0.685	0.660	0.527	0.587	0.304	0.648	0.930	0.506	0.804	0.836	0.743	0.603	0.616	0.627	0.653	0.171	0.555	0.597	0.749	0.624	0.748	0.797
	Deeplabv3+	0.692	0.720	0.747	0.642	0.615	0.459	0.426	0.236	0.624	0.914	0.471	0.773	0.742	0.725	0.557	0.599	0.586	0.619	0.039	0.651	0.555	0.673	0.661	0.812	0.820
	S2 + M	0.710	0.812	0.856	0.909	0.792	0.735	0.491	0.576	0.730	0.954	0.605	0.679	0.636	0.784	0.724	0.807	0.741	0.486	0.171	0.631	0.616	0.851	0.679	0.891	0.831
	Multi-modal targeted	0.752	0.845	0.872	0.912	0.835	0.375	0.512	0.593	0.685	0.953	0.723	0.828	0.765	0.776	0.583	0.845	0.812	0.815	0.456	0.645	0.618	0.902	0.675	0.952	0.915
Precision	Mask2Former	0.662	0.673	0.741	0.670	0.668	0.481	0.678	0.296	0.607	0.929	0.595	0.765	0.865	0.733	0.688	0.604	0.575	0.701	0.233	0.612	0.563	0.727	0.593	0.763	0.802
	Deeplabv3+	0.636	0.741	0.825	0.579	0.658	0.494	0.589	0.320	0.620	0.897	0.520	0.772	0.881	0.713	0.657	0.709	0.651	0.653	0.750	0.698	0.485	0.768	0.611	0.831	0.797
	S2 + M	0.843	0.791	0.904	0.992	0.802	0.690	0.612	0.439	0.685	0.942	0.723	0.556	0.918	0.814	0.834	0.860	0.734	0.905	0.201	0.782	0.601	0.892	0.624	0.901	0.891
	Multi-modal targeted	0.615	0.873	0.815	0.925	0.865	0.245	0.523	0.845	0.658	0.965	0.745	0.891	0.825	0.752	0.485	0.790	0.798	0.742	0.293	0.527	0.635	0.918	0.742	0.923	0.843
Recall	Mask2Former	0.739	0.736	0.756	0.701	0.652	0.583	0.517	0.312	0.696	0.931	0.441	0.847	0.810	0.754	0.537	0.629	0.689	0.611	0.135	0.508	0.635	0.772	0.657	0.733	0.791
	Deeplabv3+	0.760	0.701	0.682	0.722	0.578	0.428	0.334	0.187	0.629	0.932	0.430	0.774	0.641	0.737	0.483	0.519	0.533	0.588	0.020	0.609	0.650	0.599	0.718	0.794	0.844
	S2 + M	0.614	0.833	0.813	0.838	0.781	0.650	0.421	0.837	0.645	0.966	0.481	0.873	0.486	0.756	0.641	0.760	0.738	0.332	0.143	0.532	0.632	0.854	0.745	0.880	0.779
	Multi-modal targeted	0.705	0.872	0.852	0.875	0.850	0.375	0.442	0.593	0.715	0.953	0.685	0.815	0.758	0.776	0.427	0.878	0.815	0.765	0.342	0.622	0.613	0.895	0.675	0.941	0.872

Note. "S2" – SAM (two pass), "M" – vote by Mask2Former predictions.

4.3.4 Real-world challenge dataset

Recent Vulnerable Road User (VRU) safety assessment for Oregon Department of Transportation (ODOT) highlighted a critical data gap: overhead street lighting, a key predictor of night-time crash risk, is only inventoried on state highways in Region 1 (Portland and suburbs) of its jurisdiction while the remaining network has *no* structured lighting data. Because the existing risk-prediction models treat lighting presence as a binary covariate, the absence of comprehensive coverage forces safety analysis to omit the variable altogether on most corridors or assume lighting homogeneity, both of which hinder desired risk-based analysis using roadway inventory as a factor for state highways and non-state roadways at the agency. Given the availability of (1) ODOT's archive of forward-facing Digital Video Log imagery at 0.01-mile spacing that are collected every year and (2) public SLIs platforms offered by third-party

vendors that can be used to fill in residual gaps like ramps and access roads, automated recognition of related infrastructure element is more cost-effective than manual annotation practiced by iRAP's current procedures. Therefore the business question of this pilot study is to identify street light luminaries from SLIs available in official archive and third-party map vendors. Each identified street light should also be associated with the closest SLI to indicate the locations of these luminaries. Technically, the study asks whether our text-prompted foundation model pipeline that leverages Grounding DINO for phrase-guided object proposals followed by SAM for precise mask generation can identify luminaires across Oregon's diverse roadway contexts without re-training.

When planning large-scale collection of SLIs, two complementary acquisition strategies are considered. The first relies on an agency's own data archive, which typically cover main carriageways of the highway network. Agencies such as ODOT routinely capture "full-forward" frames along every milepost; these front views provide survey-grade quality at no additional cost and, crucially, remain consistent across jurisdictions. By contrast, side or panoramic views differ in field-of-view and angle, so using the common forward view avoids geometric distortion and simplifies downstream computer vision processing (see Fig. 4.14). If an in-house digital data archive is unavailable or if supplemental coverage is required, the second strategy calls commercial SLI sources through their APIs. Latitudes and longitudes of mileposts from high-resolution roadway network shapefiles (ideally sampled every 0.01 mile, approximatelysa mpled 16 m or even finer) can provide the location parameter for SLI query, as well as the driving direction at each point that can be inferred from its immediate upstream and downstream points or from the polyline. For every point the script issues a request to the third-party vendor's API, passing common parameters like latitude, longitude, heading,

field of view, pitch, and image resolution; the nearest panorama is downloaded automatically, after which optional quality checks can be applied, e.g., verifying that the returned image is geographically proximate to the requested coordinate). High-resolution road network shapefiles must be projected in WGS-84 coordinates (EPSG 4326) so that latitude/longitude values feed directly into the request URL and vendor's technical limits should be noted. For example, some vendors cap a single image at 640 × 640 pixels, in addition to the restriction of the field of view parameter to 120°, and limited daily call quotas. Lastly, consistent camera pitch ($\approx 0°$) and heading aligned to the driving direction to minimize horizon tilt. These two sources of SLI data complement each other. Agency data provides authoritative information on infrastructure they manage, including detailed engineering specifications, maintenance records, and traffic management systems. Meanwhile, third-party vendors collect data across the entire transportation network, including local roads, urban streets, and rural connections that extend beyond agency jurisdiction. The temporal strengths of each source are complementary as well. Agencies often have historical archives with consistent measurements at specific locations, while vendors offer real-time, continuous data collection across broader geographic areas, filling in temporal gaps that may exist in periodic agency measurements. Cross-jurisdictional integration is another area of complementarity, as agency data provides depth within specific administrative boundaries while vendor data offers breadth across jurisdictional lines, enabling seamless regional analysis. This pilot study leverages both sources and benefits from the comprehensive coverage of the highway network.

Given the business context of ODOT's needs and the lack of pixel-level segmentation labels, results are measured per SLI as a classification problem: whether presence of street lights in SLIs can be correctly identified. Results show precision ≥ 0.98 and recall ≥ 0.90 for street

```
import requests, numpy as np
from PIL import Image

base = 'https://maps.[third-party-vendor-domain].com/maps/api/streetview'
params = dict(
        key      = YOUR_API_KEY,                    #
        location = '47.652881,-122.322427',         # Seattle
        size     = '640x640',                       # maximum resolution allowed
        fov      = 120,                             # maximum field of view (widest angle)
        heading  = 180,                             # heading in degrees, relative to North
        pitch    = 0,                               # looking straight ahead relative to default (horizontally)
)

resp = requests.get(base, params=params)
with open('raw.jpg', 'wb') as f: f.write(resp.content)

img = Image.open('raw.jpg').convert('RGB')
img.save('example.jpg')
```

light detection performance on 70 miles of selected segments on Interstate highways, US highways, and OR state highways in ODOT's Region 1. More specifically, Table 4.5 and Table 4.6 present results obtained from an I-5 mainline segment and those obtained from an OR-47 segment respectively. In other words, the detection module is accurate at both (1) determining there is street light in the scene when there indeed exist luminaries and (2) deducing there is not any street light when luminaries are not present. Compared to detecting street lights in SLIs, localizing these SLIs faces more challenges. Most notably, it is required to understand which luminaries detected are serving the driving approach/direction. This could be difficult when street lights are on the left-hand side of the drivable roads and even harder when there are multiple instances of street lights in the scene. Figure 4.15 presents four examples to demonstrate the successful and failed cases: they cover interstate freeways and state highways, mainlines and ramps. In I-5 mainline example, the driving direction is northbound and SLIs on this segment are captured at different timestamps. SLIs captured in an afternoon (and thus appear darker) are from 2019 while SLIs captured around noon (and thus appear brighter) are from 2023. A closer comparison of the SLIs would reveal that the

landscapes and even the highway message boards appear different. These are the additional challenges ingesting data from third-party vendors where coverage of their SLIs may be better than state official archives, but quality control might be limited as the data can be somewhat crowd-sourced. Nonetheless, the detection and localization results are superior as shown by the number of street lights detected and their closeness to the ground truth locations. The next two examples show results obtained from two ramps of I-5 Interstate freeway. The on-ramp segment is served by five street lights, two of which are on the left-hand side whose poles are shared with luminaries serving the mainline. In these cases, the detection module works fine by segmenting corresponding pixels and classifying them as street lights, while the localization module did not have enough confidence that the luminaries are serving the ramp directly, and thereby missing them. The next three street lights are well-defined and on the right-hand side, so they are perfectly detected and localized. The off-ramp example has a similar issue where the system misses a street light on the left-hand side, but it also contains a false positive, a rare event in the developed system. A street light is present in the SLI with its pole on the right-hand side and compared to its downstream images, this instance is not repeated later. In this case, the proposed pipeline made a good judgment and the provided inventory source file missed the instance. Example of OR-47 segment is presented the last in Figure 4.15. The segment includes four records in ODOT inventory. The proposed system misses the first one as the street light overlaps with a utility pole in the SLI and is not detectable as it appears too little in the previous SLI and is not present in the subsequent SLI. After successfully detecting and locating the next three street lights, the proposed system continues to make correct decisions by producing true negative results. The factor having the greatest impact on successfully detecting street lights in Oregon's SLIs is the quality of these images.

While the proposed pipeline demonstrates robust handling of inputs of diverse range of resolution, one significant challenge emerge when working with suboptimal imagery of low resolutions. The limited field of view presents a significant challenge to comprehensive light fixture detection and localization. This limitation becomes especially problematic in areas with irregular street layouts or where lighting fixtures are positioned at varying heights and distances from the roadway. Even with advanced algorithms like the one proposed in this chapter, these physical constraints of image capture cannot be computationally overcome, resulting in systematic blind spots throughout the surveyed region. Cases exist where a street light exists in one SLI but is too small to detect given the limited resolution, it suddenly becomes missing in the next SLI because of its limited field of view. Whereas this light fixture is supposed to be associated with the latter SLI and thus mile post, it may be missing in the detected dataset at all because it was successfully recognized in neither the former nor the latter. Equivalently, the consequences of restricted field of view extend beyond simple detection failures to impact the process longitudinally. When sequential images fail to maintain consistent coverage areas, the detection pipeline struggles to establish continuity between frames, potentially counting the same fixture multiple times or missing others entirely. This issue is compounded at intersections and curved road segments where optimal camera positioning becomes particularly challenging. Without sufficient overlap between consecutive images, the system cannot reliably track individual fixtures across multiple frames, leading to fragmented spatial data that undermines inventory accuracy and completeness. Therefore the best approach moving forward is to record SLIs of high-resolution, like those available in Mapillary Vistas and host these data sources on cloud to facilitate accessibility.

This case study has huge implications on the applicability of our proposed pipeline. Firstly,

Table 4.5: Streetlight detection result on I-5 segment (MP 282.66 to MP 302.64)

	Ground Truth True	Ground Truth False
Predicted True	1,077	12
Predicted False	39	2,868

* The segment contains 3,996 hundredth mile posts counting both directions and 390 street lights in total. Among all corresponding SLIs, 1,116 images contain street lights.

Table 4.6: Streetlight detection result on OR-47 segment (MP 55.18 to MP 71.05)

	Ground Truth True	Ground Truth False
Predicted True	401	5
Predicted False	12	2,754

Note: The segment contains 3,172 hundredth mile posts counting both directions and 171 street lights in total. Among all corresponding SLIs, 413 images contain street lights.

the proposed pipeline demonstrated its capacity in recognizing road infrastructure elements in a zero-shot fashion and this capacity holds beyond street lights. Figure 4.16 shows the proposed pipeline outperforms domain models on recognizing steel bridges common in the area. Despite that these domain models are trained on Mapillary Vistas data, where similar examples of overhead steel bridges or at least overhead concrete bridges exist, the trained models still fail to extrapolate to OR's test SLIs. Our proposed pipeline and the vision foundation models therein, on the other hand, are able to extract useful elements from engineered text description, produce meaningful point prompts, and segment correct pixels of the previously unseen steel bridges. Secondly, the proposed pipeline for targeted sampling of street lights extend beyond basic inventory management. By successfully detecting and cataloging infrastructure elements on highways despite SLI quality limitations, a framework is established to revolutionize how transportation agencies approach safety analysis. Traditional infrastructure inventories often rely on manual surveys and annotations that are both

time-consuming and prone to subjective error. Our automated detection system not only expedites this process but also produces inventories with trustworthy geospatial coordinates, due to the high-resolution highway road network shapefiles. Thirdly, the catalog generated through the pipeline improves conventional safety analysis by providing granular location data that can be directly incorporated into existing safety frameworks. For example, the Highway Safety Information System (HSIS) currently hosts roadway features like binary lighting indicators (yes/no), severely limiting analytical precision. With the catalog potentially enhanced by our automated recogintion, safety researchers can now calculate distances between locations of interest and the nearest lighting fixture, introducing continuous variables that better represent real-world conditions. This shift from categorical to continuous measurement allows for more nuanced statistical modeling that can reveal previously undetectable relationships between infrastructure placement and crash patterns. Moreover, the availability of infrastructure location data changes how transportation agencies can approach both reactive and proactive safety modeling. From a reactive perspective, post-crash investigations can now incorporate actual infrastructure characteristics at the crash location, rather than relying on segment-level generalizations. This enables more accurate determination of contributing factors and targeted countermeasure selection. Even more interesting is the impact on proactive modeling: agencies can now develop predictive models that identify high-risk locations based on infrastructure configurations before crashes occur. By analyzing patterns in existing crash data alongside our comprehensive infrastructure inventory, safety analysis can help identify optimal lighting placements and configurations that maximize safety benefits while minimizing costs. Lastly, beyond immediate analytical applications, our infrastructure detection pipeline establishes the foundation for developing comprehensive digital twins

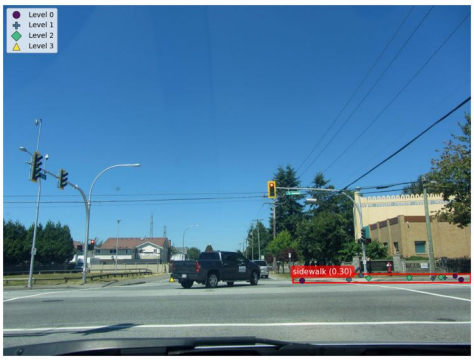
of transportation systems. As additional infrastructure elements are incorporated into the detection system—guardrails, signage, pavement markings—the resulting catalog becomes increasingly representative of the physical environment. This digital representation of the whole highway network regardless of it being rural or urban enables studies to power safety interventions can be virtually tested before physical implementation. Transportation agencies can evaluate multiple infrastructure modification strategies, simulating their potential safety impacts and cost-effectiveness before committing resources. This capability represents a paradigm shift in infrastructure planning, moving from reactive maintenance to proactive optimization based on data-driven insights derived from our comprehensive infrastructure inventory.

4.4 Discussions

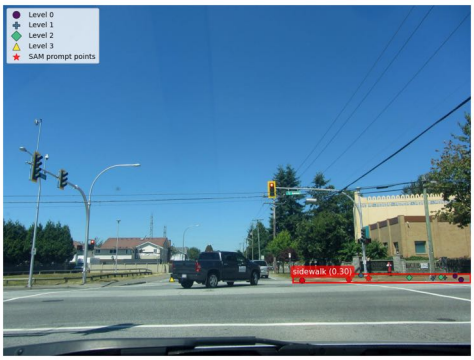
This chapter starts by reflecting on two limitations of the post-hoc voting strategy that aims to inject domain prior's knowledge on pixel-level uncertainty into SAM's segmentation masks produced in a zero-shot fashion. The first limitation concerns the uneven pixel-level coverage provided by SAM's segmentation masks in its "everything mode", which disproportionately favors certain classes within the taxonomy. Moreover, the benefits gained from iterative sampling may not be equally distributed, exacerbating the imbalance. This presents a challenge for road infrastructure recognition from SLIs, as several infrastructure-related classes consistently lag behind—both in segmentation performance and learning dynamics. The second limitation lies in the foundational assumption regarding domain models' segmentation confidence and accuracy. In practice, domain models can be confidently wrong, yet the existing framework depends on them as context-aware, mask-level classifiers. These two limitations of

the post-hoc classification pipeline motivate the core contribution of this chapter: targeted sampling guided by multi-modal input. Rather than classifying every segmentation mask produced by SAM post-generation, the proposed approach focuses on selectively segmenting relevant classes using SAM's "anything mode", informed by pre-processed class descriptions and feature representations transformed into point prompts. Following a comparison of two multi-modal foundation models and a discussion of how to extract visual activations from their attention layers, we present an integrated framework that incorporates these elements. Experimental results demonstrate that formatted class descriptions by a functional template helps Grounding DINO produce consistently good bounding boxes and class activation sample points within. SAM's tarted segmentation build on these point prompts and are successful.

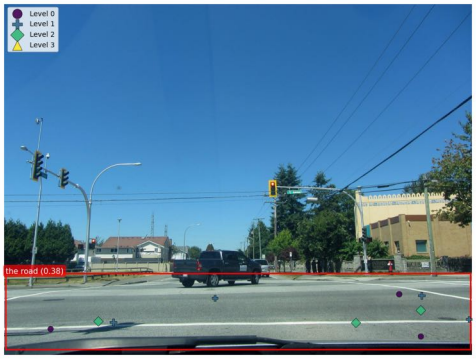
This proposed methodology advances iRAP's safety assessment by opening up a new potential avenue for smart annotation. By embedding class-specific knowledge directly into the prompt design, it allows SAM to focus on segmenting infrastructure elements that are most relevant to road safety assessments. This not only reduces the workload on human annotators but also improves consistency and coverage in identifying critical objects such as curbs, sidewalks, and manholes, guard rails. The integration of multi-modal cues further enhances the model's ability to attend to subtle visual details often overlooked in generic segmentation. Overall, this new angle provides a scalable and adaptive solution for augmenting the annotation process in support of more objective and data-rich Star Ratings assessments.



(a) Grounding DINO output with class name

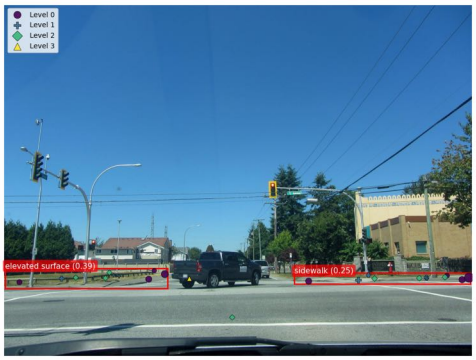


(b) Segmentation result with class name

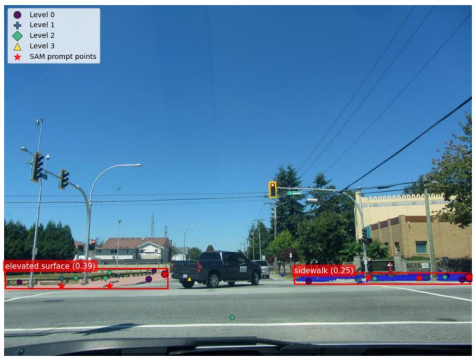


(c) Grounding DINO output with Mapillary Vistas description

(d) Segmentation result with Mapillary Vistas description



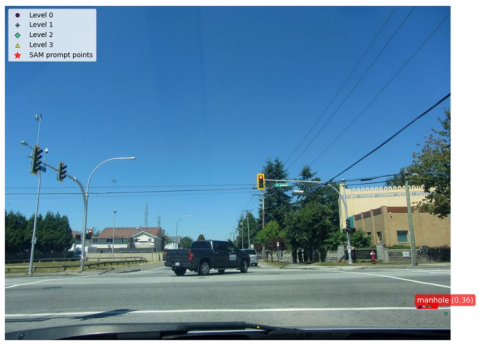
(e) Grounding DINO output with engineered description



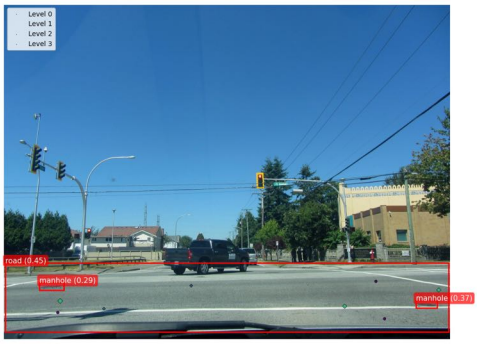
(f) Segmentation result with engineered description

Figure 4.9: Proposed descriptions, rather than descriptions from Mapillary Vistas dataset or plain class name, help Grounding DINO produce the best bounding boxes and thus point prompts for class *sidewalk*. Grounding DINO only found one relevant bounding box with class name "sidewalk" (panel **a**) and thus the segmentation result only covers one instance (panel **b**). Description from Mapillary Vistas dataset misled Grounding DINO to a big and general bounding box that corresponds to "the road" (panel **c**) and thus bad segmentation results (panel **d**). In other words, the model believes "the road" has the highest relevance to to these specific visual elements in the bounding box among all parts of the description sentence and all other text-box pairs did not meet the stability and confidence thresholds. Engineered descriptions present the best support for Grounding DINO (panel **e**), which identifies both pieces of *sidewalk* and generates good point prompts for SAM's successful segmentation (panel **f**). Note: only sample points from selected attention heads are visualized for the bounding boxes.

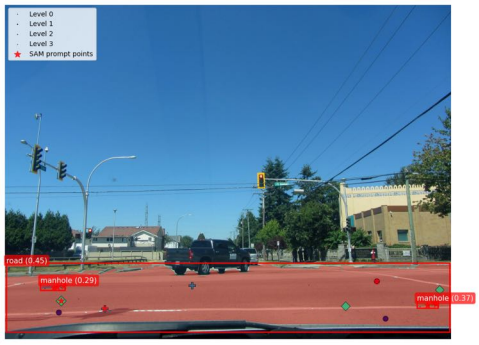
(a) Class name: Grounding DINO



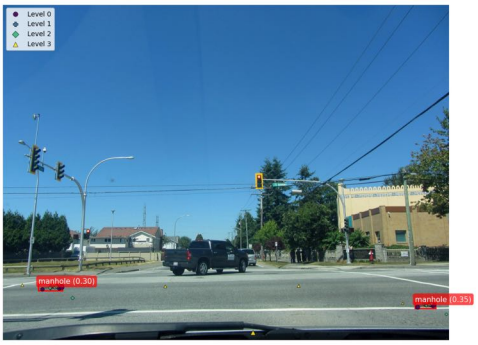
(b) Class name: proposed pipeline



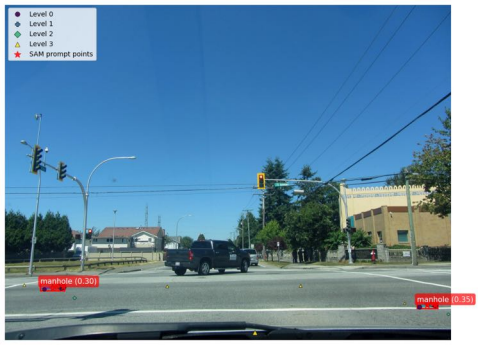
(c) Mapillary description: Grounding DINO



(d) Mapillary description: proposed pipeline



(e) LLM description: Grounding DINO



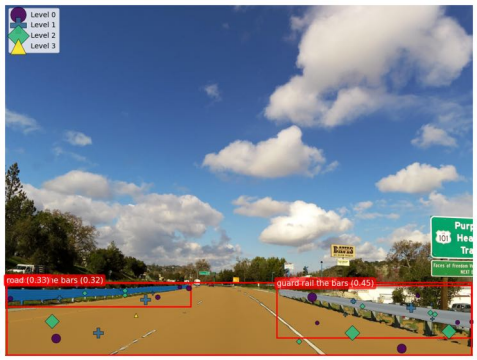
(f) LLM description: proposed pipeline

Figure 4.10: Proposed descriptions, rather than descriptions from Mapillary Vistas dataset or plain class name, help Grounding DINO produce the best bounding boxes and thus point prompts for class *manhole*. Grounding DINO only found one relevant bounding box with plain class name "manhole" (panel **a**) and thus the segmentation result is only partial (panel **b**). Description from Mapillary Vistas dataset guided Grounding DINO to find both manholes in the SLI but word "road" also triggered a big irrelevant bounding box (panel **c**), which led to bad segmentation (panel **d**). Engineered descriptions (panel **e**) present the best support for Grounding DINO, which identifies both pieces of *manhole* with well sized bounding boxes and produces good point prompts for SAM's successful segmentation (panel **f**). Note: only sample points from selected attention heads are visualized for the bounding boxes.

(a) Class name: Grounding DINO

(b) Class name: proposed pipeline

(c) Mapillary description: Grounding DINO



(d) Mapillary description: proposed pipeline

(e) LLM description: Grounding DINO

(f) LLM description: proposed pipeline

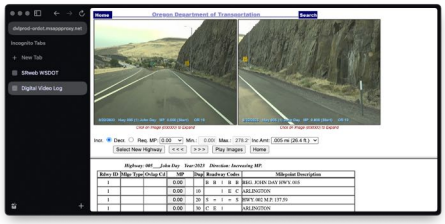
Figure 4.11: Proposed descriptions, rather than descriptions from Mapillary Vistas dataset or plain class name, help Grounding DINO produce the best bounding boxes and thus point prompts for class *guard rail*. Grounding DINO found both relevant bounding box with plain class name "guard rail" , but somehow the generated point prompts were off (panel **a**), which led to inaccurate masks from SAM (panel **b**). Description from Mapillary Vistas dataset guided Grounding DINO to find both manholes in the SLI but word "road" also triggered a big irrelevant bounding box (panel **c**), which led to bad segmentation (panel **d**). Engineered descriptions present the best support for Grounding DINO, which identifies both pieces of *guard rail* with well sized bounding boxes and produces good point prompts (panel **e**) for SAM's successful segmentation (panel **f**). Note: only sample points from selected attention heads are visualized for the bounding boxes.

(a) Points only

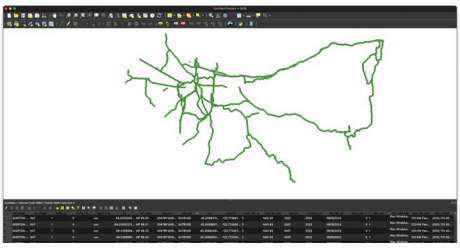
(b) Points + box

Figure 4.12: Feeding box prompt to SAM in addition to point prompts worsens segmentation quality for infrastructure element classes like *sidewalk*: (a) when only point prompts are supplied, the model respects the fine-grained cues and returns a clean mask for *sidewalk*; (b) supplying box prompt in addition to the point prompts produces a coarse mask that reflects the box's geometry but ignores the semantic hints encoded by the point set. In the example provided, the mask from box + point prompts highlights of the inverse of *sidewalk*.

Figure 4.13: Oregon Department of Transportation (ODOT) region map. Street light inventory only exists for state highways in Region 1, around Portland Metropolitan Area, while other regions have more rural and mountainous highways where systemic safety assessment can benefit the most.

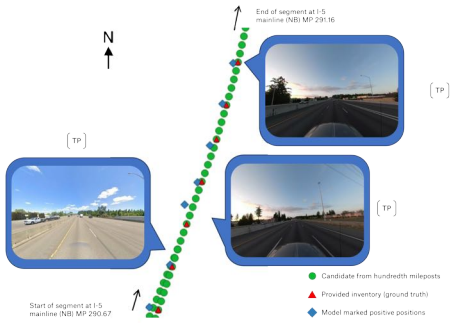


(a) ODOT Digital Video Log website

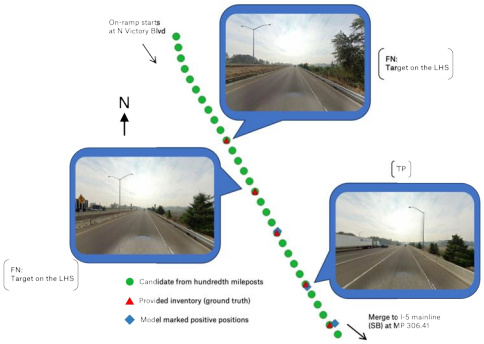


(b) ODOT road network mileposts

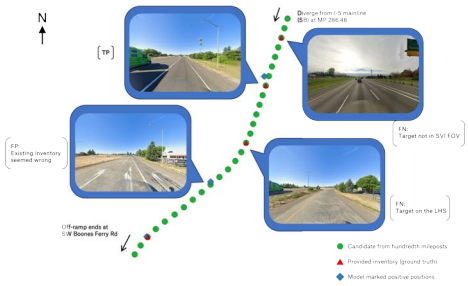
Figure 4.14: Oregon Department of Transportation (ODOT) maintains two essential data sources for this road safety assessment project: (a) the Digital Video Log provides SLIs on state highways for front- and side-views on the right-hand-side collected at 0.005 mile (26.4ft) intervals and the SLIs are of 640×480 resolution from the website archive and are refreshed by survey campaigns on a regular basis; (b) the road network shapefile provides latitude/longitude in WGS84 coordinates for linearized mileposts at 0.01 mile granularity in its Region 1. SLIs in the Digital Video Log are queried for every milepost on the exhibited network for analysis.



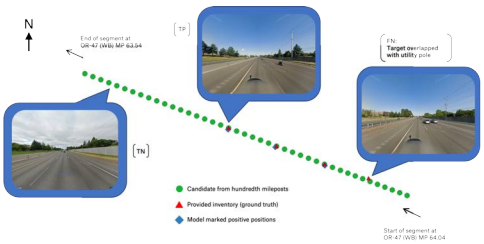
(a) I-5 mainline segment near 45.451N, -122.729E



(b) I-5 on-ramp segment near 45.591N, -122.682E

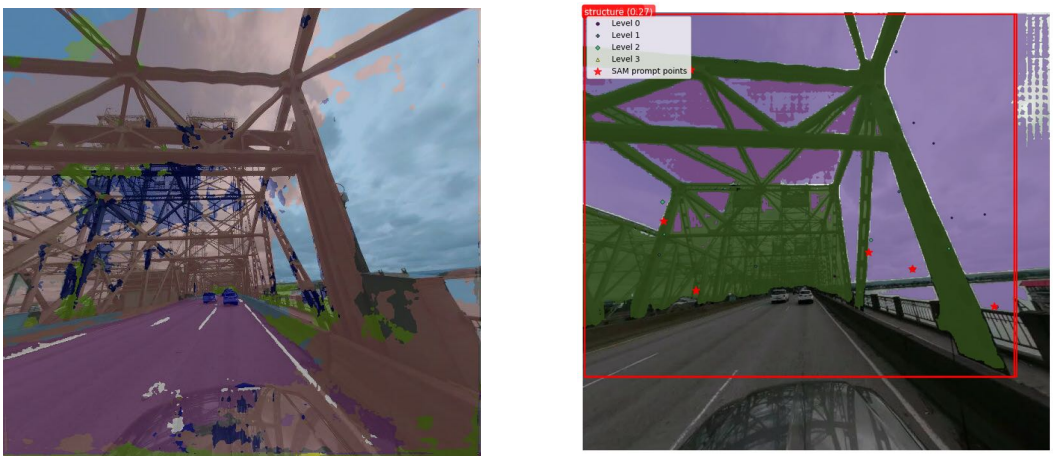


(c) I-5 off-ramp segment near 45.337N, -122.771E



(d) OR-47 segment near 45.544N, -122.872E

Figure 4.15: Proposed pipeline identifies street light fixtures in SLIs and associate them with most reasonable mile posts. The four examples include interstate freeway mainline (panel **a**), freeway ramps (panels **b-c**), as well as state highways (panel **d**). Modules in the proposed pipeline identify both true positives and true negatives consistently, while the localization function is occasionally disrupted by the limited quality of input SLIs. Specifically, limited resolution and field of view in certain situations fail to include light fixtures in the closest SLI and thus the instance is matched to SLIs that contain it immediate upstream.



(a) Domain model **(b)** Proposed pipeline

Figure 4.16: Proposed multi-modal segmentation pipeline outperforms domain model on challenging infrastructure object classes in unconventional scenarios: (a) domain model is confused by overhead steel bridges and thus classifying many pixels as vegetation and trucks, even though overhead bridges have been included in its training set in Mapillary Vistas; (b) the proposed pipeline demonstrate tremendous zero-shot capabilities to find the most relevant text patch from the engineered class description, locate the most likely bounding box corresponding to that query, generate point prompts of good correspondence and quality, and produces segmentation masks that distinguish bridge from sky pixels as much as possible.

5

Multi-source data fusion for safety analysis

5.1 Introduction

Street-level images (SLIs) offer a snapshot of the road environment from the road users' perspective. When algorithmically parsed, these images supply granular information on lane configurations, roadside hazards, midblock designs, traffic control devices, and vulnerable road user (VRU) facilities, which are the precise elements that underpin the *Safe Roads* pil-

lar of the Safe System Approach. By turning a once manual, resource-intensive visual inspection into a scalable computer vision task, automated infrastructure element recognition converts SLIs into geo-referenced asset inventories and condition ratings. Urban and rural agencies alike can therefore diagnose design deficiencies, prioritize low-cost countermeasures, and monitor asset life cycles without repeated field visits, accelerating the feedback loop between network monitoring and proactive roadway design. Yet road safety is ultimately realized through the dynamic interaction between these static environmental features and human driving behavior—the domain of *Safe Speeds*. Free-flow operating speed, speed variance, and unexpected braking all encode drivers' implicit assessment of roadway quality and perceived risk. Coupling reliable, large-scale recognition of infrastructure elements with crowd-sourced speed profiles opens a new analytic frontier: quantifying how specific design attributes modulate real-world speed choice. Such evidence can inform self-explaining road principles, calibrate automated speed management strategies, and refine performance indicators in traffic safety programs. In this light, the contribution of automated SLI segmentation transcends asset mapping; it becomes a bridge linking the static and dynamic facets of the Safe System, enabling data-driven interventions that harmonize road environment, desired operating speed, and ultimately, crash-momentum exposure.

5.1.1 CVD landscape and limits

Connected vehicle data (CVD) represents time-stamped information streams transmitted from vehicles equipped with either original equipment manufacturer (OEM) telematics systems or aftermarket connectivity dongles. These systems continuously record and transmit GPS coordinates alongside onboard diagnostic (OBD) metrics. Typical payload fields in-

clude latitude/longitude coordinates, instantaneous speed, heading, yaw-rate measurements, ignition status, and critical safety events such as hard-braking incidents. CVD's continuous, high-resolution speed traces allow direct observation of real-world operating speeds, e.g., median and 85th percentiles, and speed variance patterns that reveal driver behavior and risk perception. This data collection scales network-wide compared to traditional methods using loop detectors or specialized survey vehicles. For iRAP safety assessments, CVD provides an empirical counterpart to the theoretical "desired speed" factor in Star Rating calculations, enabling data-driven calibration of safety models.

CVD offers advantages over traditional cellular mobility data sources like call detailed records (CDR) and location-based services (LBS). Compared to these cell phone data, CVD achieves spatial accuracy of less than 3 meters, sampling intervals that are uniformly 1-3 seconds, and rich telematics features including yaw rates, hazard variables, trip indicators, and safety events. These attributes significantly surpasses those of cell-phone-based datasets: 50-300 meters spatial accuracy, 10-60s sampling intervals, and plain GPS recordings. As sensors are activated when vehicles are turned on, CVD also record clean trip starts and ends based on vehicle ignition status, which is another shortcut where cell phone data pipelines need a dedicate pre-processing step to infer origin and destinations based on extensive research.

Two critical parameters influence CVD's reliability: sampling rate and penetration rate. Sampling rate is the frequency of data collection along individual trajectories; it impacts map matching quality and speed estimation accuracy. Studies show that while high-frequency data such as those collected at 3-second intervals provides excellent trip trajectory details, it can occasionally confuse map-matching algorithms by producing spatially similar points that create illusionary detours. Meanwhile, very low sampling rates such as greater than 30 sec-

onds blur stopping patterns and distribute stopped duration across multiple road segments, deflating speed estimates. Penetration rate represents the percentage of vehicles in the traffic contributes to the dataset. It determines representativeness of the sample among the total traffic flow. This factor is compounded by CVD's inherent bias toward newer, connected vehicles and their owners, potentially missing patterns from other vehicle classes and road users. These biases can inflate or deflate observed speeds and therefore must be modeled probabilistically.

5.1.2 SLIs and segmentation context

Street-level images (SLIs) provide comprehensive geo-referenced RGB panoramas at high resolution (typically 0.15–0.3 m/pixel), refreshed regularly for the road network and with varying coverage in rural areas. These visual datasets are collected via specialized camera-equipped vehicles by platforms like Digital Video Log (DVL) system operated by ODOT, crowd-sourced platforms like Mapillary, and commercial providers such as Google Street View. The resulting datasets offer clear coverage of the roadway environment with precise geolocation metadata, time stamps, and in some advanced systems, LiDAR point clouds that enhance depth perception. The imagery resolution is sufficient to identify major infrastructure elements in most cases but also deterioration indicators, signage text, and subtle design features like curb heights or drainage patterns that affect safety performance if quality is ensured.

SLIs enable comprehensive assessment of critical infrastructure attributes and serves as the foundational data source for the *Safe Roads* pillar of the Safe System Approach. These elements include guard rail type and condition, crash cushion placement, lane configurations

and markings, roadside hazard proximity, speed-limit and advisory signage, mid-block crossing designs, traffic control device visibility, shoulder width variability, and facilities for vulnerable road users (VRUs) such as pedestrians and cyclists. While connected vehicle data excels at capturing vehicle dynamics, driving speeds, and hard braking events, it provides no information about the physical roadway environment that influences these behaviors. SLIs allow analysis for design deficiencies, prioritize countermeasures, and monitor infrastructure life cycle conditions without repeated field visits, creating a scalable visual inventory that can be processed by automated pipelines rather than site inspections or manual annotations.

The SLI processing workflow proposed in Chapter 3 and Chapter 4 leverage recent advances in vision foundation models, with the latter beginning with Grounding DINO for open-vocabulary object detection, followed by point-prompt extraction to identify points of activation, and culminating in Segment Anything Model (SAM) segmentation to generate precise pixel-level masks of road elements. This zero-shot approach offers significant advantages over traditional supervised segmentation methods, including the ability to identify previously unseen infrastructure elements without domain-specific training, adaptability across diverse geographic contexts, and high-quality segmentation masks that preserve fine details of complex roadway features. The pipeline enables efficient processing of large SLI datasets in inference mode covering hundreds of corridor-miles, with the capability to be parallelized across multiple GPUs for regional or statewide analysis projects.

The segmentation pipeline combines deterministic element detection with sophisticated probabilistic confidence scoring, providing a nuanced understanding of detection reliability across varied roadway environments. Objectness scores from Grounding DINO indicate the certainty that a detected region contains a meaningful object rather than background ele-

ments, while predicted IoU values from SAM quantify the expected quality of the segmentation mask relative to the true object boundaries. These complementary metrics serve as soft quality indicators that allow filtering or weighting detections based on confidence thresholds, enabling uncertainty-aware analysis that distinguishes between high-confidence detections suitable for automated decision-making versus lower-confidence detections that may require human verification. This probabilistic framework is particularly valuable when aggregating detections across multiple images of the same location, allowing for ensemble approaches that improve overall detection reliability through consensus mechanisms that account for varying lighting conditions, occlusions, and viewing angles present in the SLI dataset.

Despite their considerable value for infrastructure assessment, SLIs face several limitations that must be considered when incorporating them into safety analyses. Temporal staleness indicates infrastructure changes, temporary construction, or recently installed safety countermeasures could have taken place before SLIs can be refreshed and thus analysis may be using old data. Visual occlusion from vehicles, vegetation, or weather conditions can obscure critical infrastructure elements, particularly in high-traffic corridors or densely vegetated rural roads where guardrails or signage may be partially hidden. Lighting variation across time of day and seasons affects detection reliability, with harsh shadows, glare, or low-light conditions degrading segmentation quality even with advanced computer vision techniques. Furthermore, SLIs cannot capture dynamic states such as variable speed limit (VSL) panels, temporary construction zone configurations, or time-of-day restrictions that may significantly impact safety performance. Seasonal variation in roadside conditions, particularly in regions with snowfall that may obscure shoulder boundaries or hide roadside hazards, further complicates year-round applicability of insights derived solely from SLIs. These complementary

strengths and gaps motivate coupling SLIs with time-resolved CVD to create a more comprehensive understanding of the road safety environment.

5.1.3 Prior fusion attempts

Prior research has mostly focused on extracting features from SLIs and CVD and applying them to safety regression models. Cai et al.[27] modeled three years of speeding crash counts on 75 mi of Central Florida arterials with SLIs from Google Street View and probe vehicle data from INRIX. Their workflow downloads SLIs every 10 m, semantically segment them with domain deep learning model[194], and paired with monocular-depth estimates, which is projected back so that every pixel is geo-located in 3D road space. From the resulting point clouds the authors extract interpretable, segment-level visual metrics, e.g., tree canopy ratio, building facade ratio, proportion of road length lined by trees, and a Shannon-entropy index of visual complexity. Authors then aggregated five-minute probe speeds from a data vendor to estimate shares of observations exceeding the posted limit. Then these visual and speeding features are added to a pool of conventional roadway and socio-demographic covariates fed to tree-based regression models, which revealed the "speed-over-limit" metric was not a strong predictor of speeding crashes, while green-view factors (tree canopy and roadside tree presence) exert protective effects and visual complexity elevates risk. However, the authors did not normalize the road segments by length of block or vehicle miles traveled (VMT), which became the two strongest predictors of number of speeding crashes on these segments. More importantly, the authors did not consider the uncertainty in the semantic segmentation model output or the representativeness of the speed observations in collected CVD samples. Islam et al.[83] modeled rear-end or sideswipe conflicts with instantaneous vehicle

dynamics, geometric variables, and location context features. They looked for such conflicts in CVD dataset when vehicles were within 25 ft and the estimated time-to-collision (TTC) fell below 3s. Then these conflicts and a 1:4 sample of non-conflict observations were enriched with (1) vehicle driving dynamics such as speed, hard/normal acceleration, yaw rate, stop flag, (2) geometric variables from FDOT GIS layers like speed limit, lane count, median width/type, intersection control, arterial/freeway class, AADT, and (3) land use data like proximity to schools, hospitals, and fire stations. A binary logit including interaction terms then revealed ten significant variables plus two significant interaction terms. Speeding, hard brake, and acceleration as well as yaw rate were positively related to conflicts and speeding in the vicinity of hospitals was also highly correlated with conflicts. Noticeably traveling faster than 140% of the posted limit adds 7 percentage points of conflict probability. These results underline speed discipline and lateral-stability control as prime levers for proactive, surrogate-based safety management using connected-vehicle data. However, the study did not discuss the impact of 3% market penetration rate of the CVD dataset and thus the exclusion of conflicts among other vehicular road users. Authors also applied agency's GIS layers directly, which could bring errors in our context. Abdel-Aty, Ugan, and Islam [1] investigated influence of drivers' visual surroundings on their speeding behavior. They used an interesting dataset that included about 3.5 million SLIs that covered more than four thousand miles in 15 U.S. cities. They were collected from a connected vehicle fleet, consisting of vehicles that purchased the dashcam products which not only captured SLIs, but also uploaded location and speed data every 30s, thus enriching the SLIs with geo-referenced information as well as speed observations. Additionally, weather data were queried from public sources and associated with each observation. These point observations of speed were then compared with

speed limit GIS layers provided by local agencies to identify speeding observations; and the SLIs were processed similarly to Cai et al.[27] by applying semantic segmentation models to calculate portions of pixels that belong to sky, vegetation, building, and road surface within the SLIs. The authors found that visual environments, weather conditions, driver heterogeneity, as well as geographic locations influence speeding behavior significantly. More open spaces were found to encourage speeding while trees and buildings were found related to reduced speeds. Vlachogiannis, Moura, and Macfarlane[178] reversed the equation by detecting the presence and type of traffic control devices–such as traffic lights, stop signs, and right of way–from CVD movement patterns approaching intersections. Their workflow looks at CVD data from two sources: map-matches the trajectories, filters turns and congestion artefacts, then feeds slowdown-profile features plus simple network attributes into an XGBoost classifier. Despite heterogeneous sampling rates and a penetration ratio below 5%, the model delivers 97% accuracy in San Jose and 93–95% when transferred to San Francisco, with permutation tests showing that the minimum speed within 100 m of the node is the single most influential predictor. Learning curves reveal that just five trajectories going through per intersection secure 90% accuracy in-city, while approximately 20 trajectories suffice for cross-city adaptation, quantifying the data budget for scalable deployment. By demonstrating that coarse, noisy GPS alone can reliably label intersection traffic control devices by characterizing when and why performance degrades, the proposed solution complements image-dependent inventories, especially where camera coverage is patchy or vision suffers from occlusion and lighting variability.

5.1.4 Contribution and outline

Building on these research, this study is the first to develop a comprehensive fusion framework that connects infrastructure elements to observed driving behavior at network scale while addressing the methodological limitations of earlier work. This research makes three primary contributions to the field: first, it introduces a probabilistic treatment of CVD that explicitly models sampling and penetration biases through bootstrap resampling and Kolmogorov-Smirnov distribution testing, allowing realistic confidence bounds on speed estimates; second, it applies the pipelines proposed in previous chapters leveraging vision foundation models for zero-shot, open-vocabulary segmentation of roadway features without requiring domain-specific training, dramatically expanding the range of detectable infrastructure elements; and third, it implements this approach on a real-world dataset in Arizona, empirically validating the relationship between detected speed limit signage and observed operating speeds across diverse roadway contexts. These advancements create a scalable methodology for evidence-based safety countermeasure selection that quantifies how specific design attributes influence real-world driver behavior.

The remainder of this chapter is organized as follows. Section 5.2 begins with a focused review of CVD as an emerging big data source: what features and attributes are available, why its high-granularity speed traces are valuable for iRAP and road safety assessment, how CVD compares with similar location-based data sources, and what sampling- and penetration-rate pitfalls exist before adoption. The section then introduces two datasets used in this study: CVD trajectories provided by Pima Association of Governments (PAG) in Arizona and turn-movement counts from Miovision intersection sensors. Section 5.3 details a scalable pipeline that converts raw large-scale CVD trajectories into speed insights, propagates sampling uncer-

tainty, extracts posted speed limits from SLIs via the cascade proposed in Chapter 4, and unifies all sources of uncertainty, including penetration-rate bias, segmentation confidence, and variable-speed-limit stochasticity, within a Bayesian framework that yields link-level posterior distributions of the compliance gap. Section 5.4 quantify map-matching accuracy across simulated sampling rates, examine how compliance gaps shift under realistic CVD penetration scenarios, and visualise network-wide hot-spots where design attributes measured in SLIs coincide with persistent over-speeding. Section 5.5 analyzes the fusion results, highlight diagnostic insights for maintenance and speed-management, and discuss how the proposed framework can refine iRAP Star Ratings workflow as richer CVD and real-time variable-speed-limit feeds become available.

5.2 Connected vehicle data

5.2.1 Overview

Connected vehicle data (CVD) has emerged as a significant transportation big data source, attracting considerable interest within the community due to its potential to enhance traffic management and safety analysis. Enabled by modern sensor and telecommunication technologies, CVD passively gathers data from probe vehicles, covering broader geographic areas and longer time spans than traditional sensors. It has improved temporal granularity and spatial accuracy and includes novel driving event related to dynamic driving behaviors. Meanwhile, CVD is sold in various processed forms by many third-party vendors at various yet costly prices. And assessing the quality of these data products is made challenging by their black-box nature, which complicates transportation agencies' judgment of their added value and justification of substantial investments into the data, staff, and supporting infrastruc-

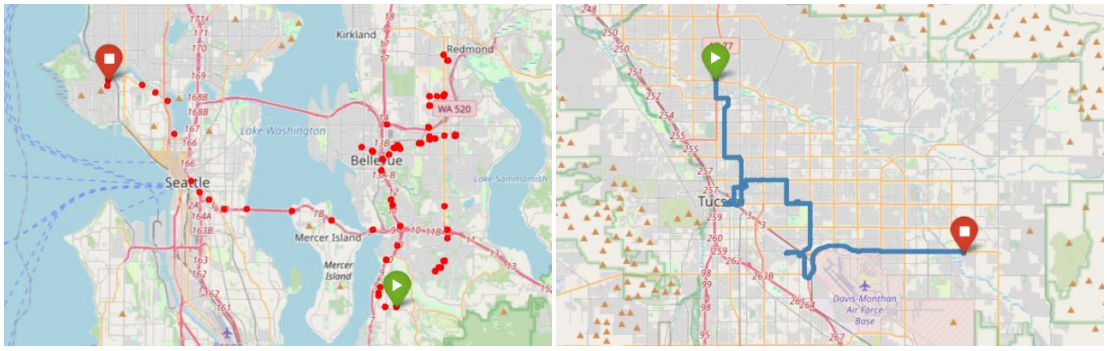
ture.

A comparison between CVD and location-based data sources collected on cell phones reveals several advantages of CVD for safety assessment purposes. Cell phone mobility data include two categories in general: call detailed records (CDRs)[60] and location-based services (LBS)[181]. CDR data are generated and passively collected by mobile carriers when cell phones interact with cell towers, for example when calls start and end, when a text message is sent or received, and when data packages are sent via the carrier's network[190]. CDR data is therefore collected as a byproduct for billing purposes routinely and thereby at a low cost and on a large scale[91,89]. However, as CDR data are collected with the cell phones' interaction with mobile base towers, its spatial and temporal granularities are sparse. Some datasets record locations represented by cell towers[89], some by inferred coordinates via triangulation or other proprietary algorithms[44,10,90,190,180]; as a result, the location accuracy range from 50m in densely populated areas to 300m in general cases[190,201], and this further depends on the carrier's market share[44]. Although CDR data are often sparse in space and time, the large volume and long observation period enable it to be used for interring human travel behaviors. It is thus often used with census[89], travel survey[44,89], land use[91,201] datasets to understand urban congestion, travel demand in origin-destination pairs, trip purposes, and activity patterns, among others. A typical pipeline includes stay and pass-by extraction, home detection, user-day filtering, and expansion to population[90].

Cell phone LBS data are positioning records of the smartphone devices either by GPS or WiFi positioning systems. Therefore its spatiotemporal granularities are much better than CDR, and features like speed and spatial accuracy of the spot measurement are available in addition to timestamped coordinates. The speed feature allows LBS to be widely used

for traffic monitoring and operation management, as these spot observations can contribute to time-mean speed calculation[99,75]. The enhanced spatiotemporal granularities expand the use of LBS across a broader spectrum of mobility topics than CDR, such as travel mode detection[203,228], social distancing compliance[137], route choice behaviors and commuting patterns[200], driving patterns[202], macroscopic fundamental diagram (MFD) calibration[136], and crash frequency modeling[123]. The improved granularity also prompts more robust trip extraction algorithms[181,180] due to data missingness and examination of LBS data quality[175], for example (1) varying observation frequencies among users, geographies, and apps, (2) varying temporal sparsity resulting in temporal clusters of observations, (3) gaps due to lost signals. A typical pipeline to process LBS data shared multiple steps with that of CDR, from basic cleaning and quality checking, to stay point detection/trip extraction, places identification via clustering, trip/tour identification, and to task-specific feature engineering and imputation, before ending at expansion to population[118,138,214,203,200]

Compared to cell phone LBS, CVD is superior in the following aspects. CVD is completely vehicle-based and thus can theoretically upload all vehicular telemetrics data in addition to standard GPS-recorded features if not considering data transmission costs/limits/latency. Features unique to CVD include (1) *yaw rate* which measures the lateral acceleration of the vehicle in degrees per second and has always been collected by specialized sensors[18]; (2) *latency* which indicates the time taken for data to be transmitted from the vehicle to the server, implicitly denoting urban canyons; (3) *ignition status* that marks the start and end of vehicular trips, really useful for departure and parking behaviors; (4) *hazard variables* that monitor critical safety events of the vehicle, including sudden acceleration, horn usage, emergency braking, and electronic stability control activation; (5) *trip variables* that track fuel levels,



(a) A trip as recorded in a typical cell phone LBS dataset **(b)** A trip as recorded in a typical CVD dataset

Figure 5.1: Connected vehicle data (CVD) are collected at higher temporal resolutions consistently, contrasting pings collected in a typical cell-phone location-based services (LBS) dataset. Panel **a** exhibits a trip reconstructed from a typical cell-phone location-based-services (LBS) dataset is represented by widely spaced pings; gaps of minutes or miles would even confuse map matching algorithms with multiple plausible routes. Panel **b** presents a trip captured in a connected-vehicle data (CVD) dataset and appears as a dense, high-frequency polyline whose samples preserve every maneuver and lane change. The consistently finer temporal resolution of CVD supports accurate speed, acceleration, and map-matching analyses for mission-critical safety assessments, whereas LBS is more suited for origin–destination insights.

signal activations, and the use of parking brakes, offering insights into driving behaviors; (6) *occupancy-related indicators* that record data on seat belt usage, the number of passengers, and door status, which are vital for safety and usage analytics; (7) *event attributes* that capture the states of wipers, lights, exterior temperature, and and detailed acceleration patterns, offering comprehensive environmental and operational context. Many of these features have never been made available at scale before and can directly contribute to safety analysis after proper processing. CVD data is almost always of higher quality because vehicular onboard sensors are built for more use cases more critical. CVD marks clear start and stop of trips based on the vehicles' ignition status and thus saves tremendous effort needed for accumulating temporally spanning datasets and inferring home and/or work location [138]. This preserves privacy because device and/or vehicle identifier is no longer needed and CVD datasets typically only record trip identifier to construct trip-level longitudinal behaviors.

Because of its vehicle- and connectivity-based nature, CVD also have certain representativeness issues. Firstly, CVD misses other travel modes and for example does not contain any VRU information, which means other data sources are always needed on top of CVD if pedestrian or cyclist activity is within the study scope. Secondly, CVD data are biased towards newer vehicle models that are equipped with GPS sensors and data onboard transmission units, i.e., "connected". This means CVD data is biased towards certain population groups that can afford "connected" models, but not others. This may not be an issue for general analysis like hotspot identification or speed estimation, but it creates bias in travel behavior understanding and OD estimation, whenever insights from CVD samples are mapped to the general population.

5.2.2 Bias mechanisms

Penetration rate is a critical parameter in mapping the knowledge obtained from CVD datasets to the true population behavior. Several studies investigate this parameter by varying the number of CVD trajectories to simulate different levels of penetration rate. Campos, Wang, and Zong[28] applied K-means clustering to find the minimum number of trajectories for signal coordination performance; Vlachogiannis, Moura, and Macfarlane[178] varied the number of input trajectories at each intersection and investigated how different levels of this factor influenced regulator identification performance of the developed ML model. They claimed models require at most five trajectories per intersection to produce confident performances trained and tested in the same region. To maximize transferrability, twenty trajectories were found required to train the initial models. Cao et al.[30] scales observed CV flow with vehicle identification data from fixed sensors. Instead of scaling by network or hourly average pene-

tration rates, the authors split CVD and their OD pairs into three subsets by comparing with vehicle re-identification results: undetected OD pair set, detected OD Pair set, and matched OD pair set. Penetration rate of each OD pair is then approximated by the penetration rate of paths between the OD. This extended linear project method considers temporal and observability variation among OD pairs. Lastly sparse entries in the tensor are approximated via decomposition techniques. Tavafoghi et al.[172] assumes the distance/position from the intersection of each stopped vehicle observed in CVD follows a uniform distribution, conditioned on the maximum queue length. Non-parametric estimation of the empirical distribution and average queue length then produces estimation of queue length agnostic of penetration level, as the authors claimed. Islam, Abdel-Aty, and Anik[82] noted the number of input trajectories in the pre-crash window varied from 21 to 1 in their study. Abdelraouf, Abdel-Aty, and Mahmoud[2] analyzed penetration rates at 182 detector locations before using the CVD dataset for traffic prediction on the network. The penetration rate ranged between 2.2% to 5.0%, with an average of 3.4%. Researchers in[222,183] used probabilistic models to consider non-CV arrivals. Kandiboina et al.[94] quantified CVD penetration rate on Iowa highways.

Another important parameter is temporal resolution, or sampling rate. Sampling rate has been widely discussed before CVD became available for its impact on speed estimation. Henrickson[72] investigated causal mechanisms that contribute to bias and uncertainty in CVD. Notably, they identified three factors contributing to bias and missingness in GPS data (mostly LBS) before large-scale CVD became available recently: (1) number of contributing vehicles, namely penetration rate; (2) varying sampling frequency of the contributing vehicles; and (3) varying speed distribution of the contributing vehicles. First of all, while all else being equal, LBS datasets record more pings for road segments in high volume and more congested

situations both spatially and temporally, such as a congested highway or peak hours. Next, because of this, the heterogeneity of contributing vehicles that have a mix of driving characteristics and sampling rates results in certain subpopulations being over-represented. In such cases, increased sample size does not necessarily improve the situation because there is no control over the added vehicle. While these are valuable insights, it is critical to point out that many conclusions no longer apply to present-day CVD. To start with, sampling rate of LBS data is typically low contrasted with CVD. This temporal sparsity of LBS leads to time-mean speed estimation, instead of space-mean speed, because the low sampling rate complicates the acquisition of travelled links. Consequently, the number of pings becomes an important factor as it serves as the denominator to produce the average speed. Contributing vehicles that are more frequently sampled by their sensors or those that travel slower on the study segment will unavoidably contribute more to the dataset and cause bias in the speed estimation. This issue is even more exacerbated by the heterogeneous composition of LBS contributing fleet, which consists of various vehicle classes whose data are collected by different types of sensors. What is different in CVD is the improvement in both heterogeneities: sampling rate and vehicle classes. In Wejo datasets used in literature discussed in this section[81,84], all vehicles are passenger vehicles and sampling rate are uniformly every three seconds. While all contributors being passenger vehicles limits CVD's representativeness in large vehicle classes like trucks, it does make CVD represent the most common vehicles on US roads. Uniform sampling rate in CVD also brings tremendous benefits. It not only eliminates the heterogeneity across car models, but also enables space-mean speed to replace time-mean speed in producing speed products, because the number of alternative routes for any three-second interval is greatly limited. Therefore, modern CVD has improved over many bias sources

of LBS datasets but it is still true that vehicles travelling at lower speeds contribute more to a CVD dataset; if aggregated in time-mean fashion, CVD dataset still biases towards the slower moving population. This highlights a key contribution of this dissertation in advocating a universal space-mean production of speed estimates and the procedure supporting it.

Sampling rate is also found critical for the quality of map matching, a pre-processing step that matches GPS traces to the digital map data. For example, Chao et al.[32] conducted a survey of map matching algorithms for trajectory datasets and experimented with three represented algorithms; one of the experiment datasets was configured by Kubička et al.[107] from 100 different areas all over the world, at the resolution of one second per observation. In the experiment, Chao et al.[32] identified confounding effects between sampling rate and other attributes of the source data. (1) Very high sampling rate can create very close pairs of GPS pings, where the succeeding point be matched to the upper stream of its preceding point, resulting unnecessarily detours in the output. This could be a significant problem if many stop points are present in CVD, such as waiting for red lights. Therefore, higher sampling rate does not always lead to better matching quality if the spatial accuracy cannot be guaranteed. (2) Matching breaks are caused by trajectory outliers, where individual observations fall out of a reasonable candidate range. In high sampling rate scenarios, trajectory compression may be a better solution to this issue compared to down-sampling. (3) High map density may affect matching result, but its mixed effect with sampling rate and spatial accuracy is not well explored yet. The paper by Newson and Krumm[132] explored some combinations of these factors when proposing the Hiden Markov Model (HMM) for map matching, but its experimentation scenario is very limited.

5.2.3 Existing safety applications

CVD is widely used in traffic operations management at intersections. Researchers from Purdue University analyzed performances by a cohort of aggregated metrics including time-space diagrams, control delay, level of service (LOS), arrivals on greens, split failure, and downstream blockage[150], for various intersection layouts, including roundabout[152], diamond interchange[153], diverging diamond interchange[155], and continuous flow intersections[151]. Similar measures were used to evaluate effectiveness of signal phasing adjustments[158] and system-wide retiming opportunities[157]. Campos, Wang, and Zong[28] summarized common performance measures and investigated the minimum size of trajectories needed for performance evaluation. Wijnand et al.[191] used hard braking and hard acceleration frequencies from CVD to evaluate Australian roundabouts. These analyses units are per-intersection, meaning the pipeline filters for activities within a spatial range of the selected intersections. Wang et al.[184], on the other hand, start the analysis in a Lagrangian view where they calculate performance index for every single trajectory based on its longitudinal events, such as free-flow speed, free-flow arrival time, number of stops, stop delay, spill overs, and queue distance, before aggregating by intersections. CVD is also used for traffic state and volume estimation at intersections. Saldivar-Carranza, Li, and Bullock[156] proposed a method based on geofencing and K-means clustering to identify individual vehicle turning movement. Zheng and Liu[222] modeled arrivals at signalized intersections as time-dependent Poisson process and and demonstrated CVD at low penetration level can be used for robust estimation of volumes. Tavafoghi et al.[172] estimated empirical distribution and average value of queue lengths, claiming to be agnostic to penetration rates and queue dynamics. Instead of using looking at deterministic time-space diagrams, Wang et al.[183] connected vehicle trajectories with stochastic point-

queue by a probabilistic time-space diagram, which reconstructs the recurrent spatiotemporal traffic state at intersections for optimize signal retiming.

CVD is used for operations management on highways. Khadka, Li, and Wang[97] proposed performance measures based on time-space diagram and showed they were effective for identifying queue propagation at freeway bottlenecks and estimating vehicles' actual queuing time between intersections. Desai et al.[49] evaluated impact of freeway construction work zone diversions with speed profile of the corridor in addition to the performance measures collected at impacted intersections. Ahmad et al.[8] assessed traffic delays and congestion with speed profiles of the corridor during a short-notice wildfire evacuation as compared to before and after the event. Travel time analysis were performed by Jalilifar et al.[86] for border crossing time at the Paso del Norte (PDN) International Bridge in El Paso, TX, where the authors compared results from CVD with existing Bluetooth-based system and found that CVD duration estimates and CVD-generated variables can boost estimation accuracies. Zhong et al.[224] proposed multiple performance indexes across mobility, safety, riding comfort, traffic flow stability, and fuel consumption. Multiple studies used hard braking events as metrics, such as Ahmad et al.[7] for driving patterns during wildfile evacuation, Mathew et al.[124] and Sakhare et al.[149] for freeway queues and queue warning systems, and Desai et al.[48] for work zone safety evaluation.

CVD is used for planning pusposes. Cao et al.[30] estimates dynamic origin-destination flow with automatic vehicle identification (AVI) observations and CVD. The authors split the workflow into two modules, where the first one produces estimates with CVD as prior and the second module calculates the optimal estimate based on this prior knowledge. Nonnegative Tucker decomposition (NTD) is adopted in the first module to address data sparsity

due to CVD penetration rate. Liu et al.[119] synthesized travel itineraries based on CVD trips and CVD data typically do not include vehicle or driver identifiers. They used this synthesized itinerary data to simulate and thus estimate EV charging demands for Virginia.

CVD is also extensively used in safety research. Li et al.[108] used hard braking events to identify roadway hazards at work zones, signalized intersections, interchanges, and entry as well exit ramps. Their assessments showed frequency of hard brakes could be used signals to warrant geometry improvement and advance warning signs. Mussah and Adu-Gyamfi[128] compared CVD-derived variables with historical crash records and found hard acceleration events is strongly correlated crash outcomes in variuos statistical and machine learning models. Features extracted from CVD are also widely used as features to predict real-time crash risks. Mussah and Adu-Gyamfi[128] built a model to predict real-time binary crash instances at five minute resolution in Greater St. Louis, MO, with segment speed from CVD, reference speed, and the differences of speeds between consecutive segments, among others. Zhang and Abdel-Aty[215] modeled crashes in Greater Orlando, FL by collecting an extensive list of CVD safety attributes within five to ten minutes before reported crash occurrence. This list aggregated the variables to one minute level: number of braking, hard brakes, hard acceleration, AEB activation, lane change signal, wiper usage, in addition to lane occupancy, mean speed, volume of the crash segment as well as its immediate upstream and downstream neighbors. A deep learning model is then trained to predict binary crash outcome and then transferred to deployment where data is aggregated hourly. Anik, Islam, and Abdel-Aty[12] investigated crashes at eight intersections in Florida. They extracted nine general features from CVD within pre-defined range of the study intersections: split failure counts and percentage, travel time maximum and average, control delay maximum and average, approach speed maximum

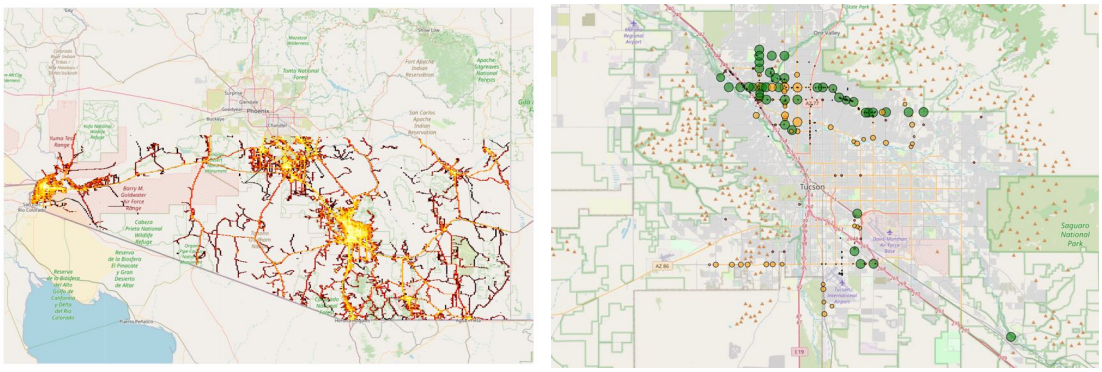
and average, and percent arrival on green every 15 minutes. The total number of features before selection reached 114 for *within-intersection* group and 33 for *approach* group, containing both traffic and weather information. Instead of aggregating counts, Islam, Abdel-Aty, and Anik[82] extracted all CVD trajectories in the five minute window prior to crashes and within 300 feet. Observations that are within 30 seconds prior to the crash moment were labeled with positive label, while others identified in the spatiotemporal window were labeled with zero. Theses trajectories with labels associated with each observation are then converted into a two-dimensional matrix to preserve the raw information of trajectory, after which the matrix was ingested by the deep learning model. Instead of using external crash records, authors in several studies[81,83] identified conflict events within the CVD population, by comparing their trajectories spatiotemporally. These CVD-only conflicts were then modeled with vehicle dynamics, geometric, and non-geometric roadway attributes to reveal the most significant contributing factors. Although CVD-only conflicts may be limited in number and bring biases, these studies did demonstrate the capacity of real-time individual vehicular dynamics features and how maneuvers caused near-misses and even crashes. Such longitudinal insights were also leveraged by Ugan, Abdel-Aty, and Islam[174] in modeling speeding behaviors. Features extracted from longitudinal behavior included time spent as stopped in total, at signalized, and at unsignalized intersections; total length of trip; number of intersections passed through; number of turns in trip; max and average hard brake rates and number of hard brakes in trip; max and average hard acceleration and number of hard acceleration; max, mean, of yaw rates and average during turns; as well as proportions of commercial/industrial/institutional areas and all road classes. The trained ML model is used for understanding how trip factors influence may cause speeding. CVD is used for speed control infrastructure

detection. For example, Vlachogiannis, Moura, and Macfarlane[178] investigated movement patterns of CVD vehicles when approaching intersections. They demonstrated how CVD slowdown and speed features in addition to roadway characteristics can infer the existence and type of regulators. Their sensitivity analysis also explored the size of CVD datasets to reconstruct the traffic regulator presence by varying number of input trajectories.

5.2.4 Research dataset

CVD are crowd-sourced by passenger vehicles, providing extensive spatiotemporal coverage while maintaining enhanced granularities. The research dataset is provided by Pima Association of Governments (PAG), a federally designated metropolitan planning organization (MPO) in Pima County, AZ, whose members including Pima County and the towns, cities and tribes in the county, as well as the Arizona Department of Transportation (ADOT). Specifically, these CVD trajectories cover two months in 2021 (January and August) and have been pre-processed by PAG to remove personally identifiable information and to sort by trip id and collection time. The sorting could be extremely cumbersome as data at this volume are typically stored in file formats optimized for distributed processing, i.e., parquet files, and thus records belonging to the same trip could be scattered in hundreds of files. Sampling intervals of these CVD data are 3 seconds. An accompanying dataset provided by PAG is traffic volume collected by Miovision[125] camera sensors. These sensors are installed at intersections as shown in Figure 5.2 and are primarily used for turning-movement and queueing data collection. While data are reported every 15 minutes, we aggregate them to hourly or daily for estimating penetration rates of CVD.

Map data used in this research are queried from Open Street Map (OSM)[63], a free and



(a) CVD data heatmap in PAG

(b) Miovision sensor locations

Figure 5.2: Connected vehicle data (CVD) provide good coverage of the regional road network. Panel **a** shows heat map of the CVD dataset, which shows these trajectories not only cover urban areas in Tuscon and Yuma, but also major border crossing corridors as well as tribal roads. Panel **b** shows Miovision sensors installed at intersections around Tucson, AZ. Radius of icons represent traffic volume and colors indicate data quality. Green indicate small missing rate, orange means acceptable missingness, and red indicate alarming missingness. Underlying statistics are from Wu, Li, and Xu[177].

crowd-sourced geographic database updated and maintained by volunteers. OSM uses a topological data structure with four core elements: nodes, ways, relations, and tags. *Nodes* are points with a geographic position, stored as coordinates. Besides their usage in *ways*, they are used to represent map features without a size, such as points of interest or mountain peaks. *Ways* are ordered lists of nodes, representing a polyline, or possibly a polygon if they form a closed loop. They are used both for representing linear features such as streets and rivers, and areas, like forests, parks, parking areas, and lakes. *Relations* are ordered lists of *nodes*, *ways* and *relations*, where each member can optionally have a role. *Relations* are used for representing the relationship of existing *nodes* and *ways*, such as turn restrictions on roads, routes that span several existing *ways*. *Tags* are key-value pairs storing metadata about the map objects, e.g., type, name, and physical properties. Hence *tags* are always attached to an object: to a *node*, a *way* or a *relation*. Based on this open and free road network data by OSM, high-performance routing engines were implemented and open-sourced to compute shortest paths and provide

map snapping on large road networks efficiently to serve transportation and supply chain applications.

5.3 Methods

Having established the raw data sources, this section first details the processing pipeline that turn CVD trajectories into link-level speed profiles. We then extract posted speed limits from SLIs, before combining them in a Bayesian framework that propagates detection uncertainty through the network and simulates various CVD penetration rates to estimate speed compliance metric.

5.3.1 Speed processing pipeline

A robust evaluation of speed-management countermeasures begins with a repeatable, auditable workflow that converts raw, CVD pings into statistically explainable and maintainable measures of operating speed. The procedure should (1) preserve the full temporal resolution of connected-vehicle data (CVD), (2) mitigate known artifacts introduced by sampling frequency, penetration bias, and map-matching uncertainty, and (3) emit link-centric outputs that downstream safety tools like iRAP Star Ratings can query without additional preprocessing. To serve these goals, a seven-step pipeline that promotes every intermediate product, e.g., stop-segmented sub-traces, HMM match confidences, and bootstrap replicates, to version-controlled tables and annotates each stage with quantitative goodness metrics such as F-measure for map matching and Kolmogorov-Smirnov distance for penetration re-scaling. Figure 5.3 provides a synoptic view of the pipeline together with the column-oriented data model that underpins it. Black boxes denote deterministic transformations of

the CVD stream, teal boxes mark points where uncertainty is injected or propagated. These CVD-derived speed profiles are joined to SLI-derived posted speed limits in the next section, where the probabilistic framework is discussed.

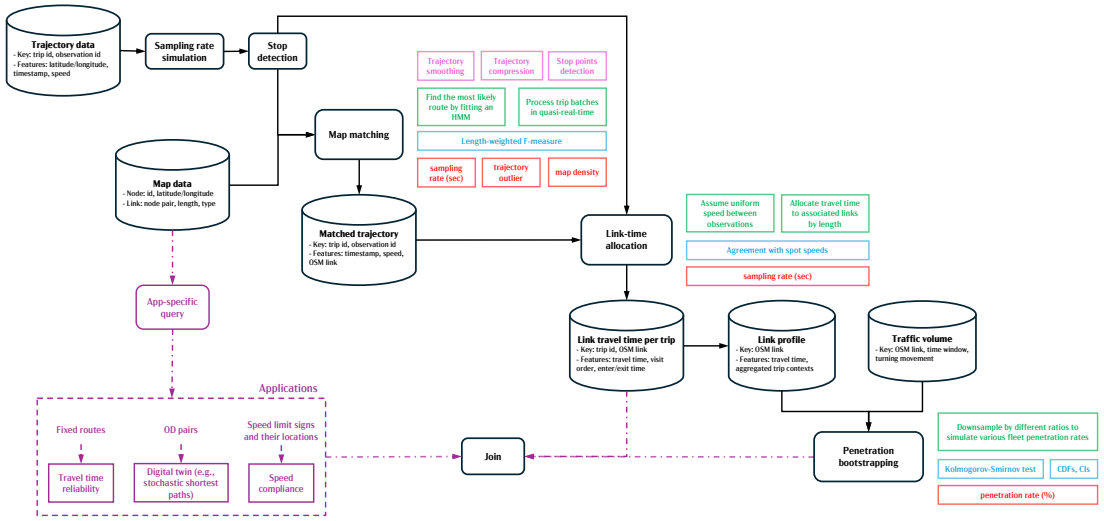


Figure 5.3: Proposed speed processing pipeline uses six steps to process, organize, and link connected vehicle trajectory data with other data sources for safety analysis. Firstly sampling frequencies can be simulated optionally for various levels of data quality, e.g., mimicking LBS out of CVD; then complete stops are detected derived or origiinal trajectories to segment trips into sub-traces. The map matching function finds the most likely sequence of road links for these sub-traces with overlapping map data. Each ping in the sub-trace associates with a matched link so that time spent on each link can be estimated. Duration of complete stops, e.g., for traffic lights, are added to those road links. Link travel time per link per trip can group by trip id to obtain link profiles. These profiles are keyed by *link id* and observations for a link serves as a candidate pool, where the bootstrap function can simulate any desired levels of penetration rate by downsampling or oversampling from that pool. Such bootstrapping is only valid when ground-truth traffic volumes can be joined with link profiles at the desired level, because any simulated penetration rate is supposed to dwarf the ground truth. Lastly, travel times per link per trip and/or bootstrapped results are joined with external data as required by applications. For example, if travel time reliability is the goal of analysis, then routes need to be provided to join with link speed data; speed compliance is the goal of analysis, then location of the speed limit signs need to be associated with road links.

The raw trajectory feed arrives as time-ordered observations $\mathcal{O} = \langle t_i,\ \varphi_i,\ \lambda_i,\ v_i \rangle_{i=1}^{N}$, where (φ, λ) are WGS-84 coordinates and v_i is the measured speed. To assess how sparser data streams would affect map matching and speed estimation frequencies, we first derive a family of down-sampled replica emulating lower sensor sampling, given the original observa-

tion sequence $O = \{\mathbf{s}_k\}_{k=1}^{K}$ with median sampling interval $\Delta t_{\text{raw}} = 3$ s,

$$O^{(\tau)} = \left\{ \mathbf{s}_k \;\middle|\; t_k - t_1 \equiv 0 \pmod{\tau} \right\}, \qquad \tau \in \{9, 15, 30\} \text{ s}, \tag{5.1}$$

where τ is the target sampling period. Each $O^{(\tau)}$ retains the first and last fixes of the original trace to preserve total trip duration. When $\tau = 3$ s, $O^{(\tau)}$ coincides with the raw feed. For every sampling regime τ we process the corresponding observation stream $O^{(\tau)}$. Although the vendor applies smoothing filters, two potential issues persist: duplicated GPS location observations during complete stops and *micro-jitter*—millimetre-scale noise while idling. Both can corrupt map matching algorithms, especially state-transition matching models[32], and produce phantom cruising. Instead of compressing trajectories[162], our procedure detects complete stops of the trajectory indicated by duplicated latitude/longitude pairs with zero speed and partition each trip into non-overlapping *sub-traces*. When data sampling rate is high, this leverages the enhanced granularity to estimate time spent on particular OSM links, as they tend to short and fragmented. When data sampling rate is low, there is little need for extra compression in the first place.

$$\mathcal{T} = \{\mathbf{s}_k\}_{k=1}^{K}, \quad \mathbf{s}_k = \Big\langle t_i, \varphi_i, \lambda_i, v_i \;:\; v_i > 0 \;\vee\; \|(\varphi_{i+1}, \lambda_{i+1}) - (\varphi_i, \lambda_i)\| > 3 \text{ m} \Big\rangle. \tag{5.2}$$

The 3-meter threshold filters out micro-jitter yet retains legitimate creeping in congestion. Each sub-trace inherits an identifier k to denote `trip_id`, enabling downstream joins while preserving the parent trip context. Descriptive statistics recorded at this stage, including median sampling interval, stop count, and stop duration, form the first layer of the pipeline's provenance log.

These down-sampled sequences are then map-matched *independently* to the OSM network using the HMM algorithm. Let $S^{(\tau)} = \{s_j^{(\tau)}\}$ denote the candidate link-state sequence for sampling regime τ; the Viterbi decoder maximises

$$P\big(S^{(\tau)} \mid O^{(\tau)}\big) \;\propto\; \prod_{j} P\big(o_j^{(\tau)} \mid s_j^{(\tau)}\big)\, P\big(s_j^{(\tau)} \mid s_{j-1}^{(\tau)}\big), \tag{5.3}$$

where emissions probabilities follow a 2-D Gaussian ($\sigma_\varphi = \sigma_\lambda = 5$ m) distribution and transitions decay exponentially with great-circle distance between consecutive candidate links, i.e., the same emission and transition kernels used for the raw 3-second data. Executing the match at multiple τ values allows us to quantify how sampling frequency degrades recall, precision, and ultimately speed-profile reliability. Map matching would return for every input point a triplet $\langle \text{link\_id}, \text{offset}, p_{\text{mm}} \rangle$, where p_{mm} is the marginal match probability. Map matching fidelity is tracked with the length-weighted F-measure

$$\text{precision} = \frac{\|T \cap M\|}{\|M\|}, \qquad \text{recall} = \frac{\|T \cap M\|}{\|T\|}, \qquad F = \frac{2\,\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{5.4}$$

with ground-truth paths T derived from a 5 Hz reference set and M is the matched path[32]. Sensitivity analysis reveals that the F-measure remains > 0.92 for sampling intervals up to 15s; quality degrades sharply beyond 30s as alternative routes arise.

The next step estimates time spend on each link. Let two consecutive matched observations be $\langle t_a, \varphi_a, \lambda_a \rangle$ and $\langle t_b, \varphi_b, \lambda_b \rangle$, connected by link sequence $\mathbf{L}_{ab} = [\ell_1, \ldots, \ell_m]$ with total length $L_{ab} = \sum_{r=1}^{m} L(\ell_r)$. Assuming *uniform-speed* between these two matched GPS

observations, traversal time allocated to link ℓ_r is

$$\Delta t(\ell_r) = (t_b - t_a)\,\frac{L(\ell_r)}{L_{ab}}. \tag{5.5}$$

When $m = 1$ the full interval is attributed to a single link; for $m > 1$ the proportional rule above prevents negative speeds on short connectors. Complete-stop segments inherit their measured dwell time, ensuring that queueing delay is not suppressed. Per-link speed is thus

$$v(\ell) = \frac{L(\ell)}{\sum_{q=1}^{Q(\ell)} \Delta t_q(\ell)}, \tag{5.6}$$

with $Q(\ell)$ denoting the number of traversals of ℓ by the same trip. For sampling intervals exceeding 20 s, which are extremely rare cases in this CVD dataset, an additional cubic-spline interpolation layer is invoked to mitigate the piecewise-constant bias documented[72]. All intermediate tables conform to a wide schema keyed by $\langle$trip_id, obs_idx$\rangle$ for point-level products and $\langle$trip_id, link_id$\rangle$ for link-level aggregates; this uniform keying enables vectorized bootstrapping in the probabilistic framework. It is not hard to find out sampling rate has a strong impact on this step as well. When sampling rate is low, the uniform speed assumption during link time allocation may be compromised. Therefore, speeds between ground truth (i.e., high sampling rate) and simulated (i.e., low sampling rate) can be compared to measure this influence. Measured speeds can also be compared with spot speeds observed at each signal ping, as spot speeds in CVD are more reliable than LBS data.

Upon completion of the link–time allocation, each traversal record has the canonical schema

$\langle \text{trip\_id}, \text{link\_id}, v, \Delta t \rangle$. Aggregating over trips yields a *link profile*

$$\text{prof}(\ell) \;=\; \Big\langle \mu_v,\; \sigma_v,\; v_{50},\; v_{85},\; n_{\text{trips}} \Big\rangle, \tag{5.7}$$

where μ_v and v_p denote the mean and p-th percentile of the speed distribution, respectively. This profile is our "best-knowledge" benchmark at the observed fleet penetration. To evaluate sensitivity to lower penetrations, we apply a stratified bootstrap that mimics random under-sampling of the connected fleet:

- choose a target penetration $\pi \in \{1\%, 2\%, 3.5\%, 5\%\}$;
- for every link ℓ sample, *without replacement*, $\lfloor \pi\, n_{\text{trips}}(\ell) \rfloor$ traversals;
- recompute μ_v, v_{50}, v_{85};
- repeat $B = 1{,}000$ times, storing results under key $\langle \text{link\_id}, b \rangle$.

The difference between the bootstrapped and full-sample speed distributions is quantified with the two-sample Kolmogorov–Smirnov statistic $D_{n,m}$ defined in Eq. 5.8; a p-value below 0.05 flags links whose speed estimate is unstable at the chosen π: $D_{n,m}$ denotes the distance statistic calculated between two samples and $F_{1,n}(x)$ and $F_{2,m}(x)$ are the CDFs. The null hypothesis, i.e., two samples are from the same distribution, is rejected at level *alpha* if $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$, where $c(\alpha)$ is a value based on α.

$$D_{n,m} = \sum_{x} |F_{1,n}(x) - F_{2,m}(x)| \tag{5.8}$$

A point estimate of the *true* CVD penetration is required to anchor the bootstrap analysis. Let $C_{\text{CVD}}(h)$ denote the number of distinct CVD vehicles crossing a detector during hour h

and $C_{\mathrm{GT}}(h)$ the corresponding ground-truth count. Aggregating over a period $\mathcal{H}$ as in Equation 5.9 gives the estimate penetration rate. For 7 January 2021 the mean arterial penetration is 5.5 %, with the full range being 4.9–6.1 %.

$$P_{\mathrm{CVD}}(\mathcal{H}) \;=\; \frac{\sum_{h\in\mathcal{H}} C_{\mathrm{CVD}}(h)}{\sum_{h\in\mathcal{H}} C_{\mathrm{GT}}(h)}. \tag{5.9}$$

The vision pipeline produces speed limit detection results based on SLIs by means of a lookup table.

$$\texttt{limit\_tbl} = \big\langle \text{link\_id},\, L,\, p_{\mathrm{det}},\, \text{is\_dynamic},\, \text{last\_seen} \big\rangle, \tag{5.10}$$

where L is the numeric speed limit (mph) decoded from the SLI mask; p_{det} is the detection confidence calculated by the proposed vision pipeline in Chapter 4; is_dynamic $\in \{0, 1\}$ flags variable-message signs; and `last_seen` records the capture date of the source SLI. For *static* R2-series signs the limit value is treated as deterministic when $p_{\mathrm{det}} > 0.70$; lower-score detections are discarded, so p_{det} drops out of the posterior. When multiple static signs map to the same link, the record with the most recent `last_seen` is retained. Links that carry variable speed displays (is_dynamic $= 1$) receive a placeholder $L =$ NaN and are excluded from the compliance analysis unless a controller feed is available. The join is executed once per bootstrap replicate via link_id and materialized as $\langle \text{trip\_id}, \text{link\_id}, v, L \rangle$, ensuring that every speed observation is co-located with a posted limit where one exists.

5.3.2 Probabilistic framework for compliance

Conventional compliance studies typically stop after a single bootstrap estimate of V_{85} and treat the posted limit L as a fixed input. That approach masks three independent error sources already present in our pipeline: fleet penetration sampling and sign detection uncertainty (or controller variability for VSL links) in L. Equally important is the veracity of both inputs. We therefore develop an uncertainty-aware pipeline that quantifies sampling bias in connected-vehicle speeds, and learns a probabilistic speed-limit layer from automatically detected signs in SLIs. The resulting framework outputs a distribution over speed-exceedance intensity rather than a point estimate. By embedding these uncertainties in the proposed Bayesian Monte-Carlo layer, for every link ℓ, a *posterior distribution* of $\Delta(\ell) = V(\ell) - L(\ell)$ can be calculated. This is novel on two fronts: (1) it is the first workflow to propagate error *jointly* from both connected-vehicle speeds and vision-derived speed limits, and (2) it yields a direct probability measure $P\big(\Delta(\ell) > 0\big)$ that can be aligned with speed limit design principles, which no deterministic method can quantify defensibly.

The probabilistic layer also deepens the Safe System interpretation of the results. For the *Safe Speeds* pillar it shows not just *where* speeding occurs, but *how sure* we are of that diagnosis, guiding agencies to schedule imagery refreshes or spot-radar checks only where posterior uncertainty is large. For *Safe Roads*, the same posterior can be cross-filtered with the roadside-severity scores extracted in Chapter 4, producing a combined "speed × severity" hotspot list that prioritizes segments where a high-energy crash would be least forgiving. In short, transforming Δ into a distribution rather than a single number closes an acknowledged research gap and supplies practitioners with a risk metric that is both statistically rigorous and operationally actionable.

The processing pipeline yields, for every bootstrap replicate b and link ℓ, a paired observation $\langle v^{(b)}(\ell),\, L(\ell)\rangle$. A purely point-estimate treatment would ignore three uncertainties: fleet penetration bias, sampling-rate artefacts, and imperfect sign detection or, for VSL corridors, intrinsic variability in the displayed limit. We therefore wrap the deterministic output in a Bayesian layer whose goals are to (1) convert raw bootstrap draws of $v^{(b)}(\ell)$ and detection confidences on $L(\ell)$ into a full posterior distribution of the compliance gap $\Delta = V - L$; and to (2) produce link-level probabilities of speeding that align with iRAP's criterion, enabling objective hotspot selection.

For every link ℓ we treat the *operating* speed and the *posted* limit as random variables due to the stochasticity in vision model.

$$V(\ell) \;\sim\; f_V\big(\,\pi, \mathcal{S}(\ell)\big), \quad L(\ell) \;\sim\; \begin{cases} \delta\big(\hat{L}\big), & \textbf{static sign,}\ p_{\text{det}} > 0.70, \\ \text{Bernoulli}\big(p_{\text{det}}\big)\cdot \hat{L}, & \textbf{static sign,}\ p_{\text{det}} \le 0.70, \\ g_t(\,\cdot\,), & \textbf{variable-speed sign (VSL)}, \end{cases} \tag{5.11}$$

where f_V is the empirical bootstrap distribution of the link's 85-th-percentile speed; π is the fleet penetration level, $\mathcal{S}(\ell)$ denotes the full set of speed samples on ℓ; $\hat{L}$ is the numeric limit read from the SLI mask; $\delta(\cdot)$ denotes a degenerate one-point distribution; p_{det} is the sign-detection confidence; $g_t(\cdot)$ is a time-indexed discrete PMF supplied by the VSL controller. While due to limited access to dynamic data stream for SLIs like those captured by connected dash cameras, signs available in Mapillary dataset and third-party vendors are updated with low cadence and thus we only allow the first two options and if $p_{\text{det}} > 0.70$, so $L(\ell)$ is deterministic; the general form is kept to show how the framework accommodates lower-

confidence detections or future VSL feeds.

Three complementary compliance measures can be computed, given every bootstrap replicate b and each link ℓ that possesses a valid posted limit:

$$\Delta_{50}^{(b)}(\ell) = v_{50}^{(b)}(\ell) - L(\ell), \tag{5.12}$$

$$\Delta_{85}^{(b)}(\ell) = v_{85}^{(b)}(\ell) - L(\ell), \tag{5.13}$$

$$E^{(b)}(\ell) = \mathbb{1}\{v^{(b)}(\ell) > L(\ell)\}, \tag{5.14}$$

where $v_p^{(b)}(\ell)$ is the p-th percentile speed drawn from the bootstrap distribution of $V(\ell)$ at penetration level π. The gap statistics Δ_{50} and Δ_{85} track median and 85^{th}-percentile compliance, respectively, while the binary indicator $E^{(b)}$ records any exceedance event. After B replicates, the empirical speed exceedance probability for each link is summarized by

$$\mu_{\Delta_p}(\ell) = \frac{1}{B}\sum_{b=1}^{B} \Delta_p^{(b)}(\ell), \quad \sigma^2_{\Delta_p}(\ell) = \frac{1}{B-1}\sum_{b=1}^{B}\big(\Delta_p^{(b)}(\ell) - \mu_{\Delta_p}(\ell)\big)^2, \tag{5.15}$$

$$P_{\text{exceed}}(\ell) = \frac{1}{B}\sum_{b=1}^{B} E^{(b)}(\ell). \tag{5.16}$$

Links with $P_{\text{exceed}}(\ell) \geq 0.15$ deserve closer investigation by road authorities, if posted speed limits are expected to match 85th operating speed. This means on average 85% of speed observations should be below this threshold and equivalently, speed compliance failure should be limited to 15% or below. We can leverage this interpretation to identify road segments of *high-risk* or set up separate thresholds for that purpose. All posterior statistics are written to a columnar store keyed by `link_id`, providing a plug-and-play input to the Star Ratings

calculation chain. More specifically, the Monte-Carlo sampling to approximate compliance gap $\Delta(\ell) \;=\; V(\ell) - L(\ell), \ell \in \mathcal{E}_{\text{static}}$ and its posterior CDF $F_\Delta(x \mid \ell) = P\big(\Delta(\ell) \le x\big)$ consists of these steps:

- outer loop (penetration grid): for each penetration level $\pi \in \{1\%, 2\%, 3.5\%, 5\%, \pi_{\text{emp}}\}$, draw $B = 1{,}000$ bootstrap replicates of v_{85}.
- inner loop (detection uncertainty): for *static* signs with $p_{\text{det}} \le 0.70$ flip a Bernoulli coin with that probability; if the draw fails then set $L = \text{NaN}$ and skip the replicate; for $p_{\text{det}} > 0.70$ we accept the limit deterministically.
- compliance metrics: record $\Delta_{50}^{(b)}$, $\Delta_{85}^{(b)}$, and the indicator $E^{(b)} = 1\{v_{85}^{(b)} > L\}$.
- posterior summaries: after B draws we store $\mu_{\Delta_p}, \sigma_{\Delta_p}$, the one-sided 95% CI, and $P_{\text{exceed}} = B^{-1} \sum_b E^{(b)}$.

In summary, this probabilistic layer delivers three novel ingredients absent from prior CVD–SLI fusion work: explicit penetration-rate propagation, detector-confidence propagation without re-training, and link-level posterior risk scores that plug directly into iRAP's decision thresholds.

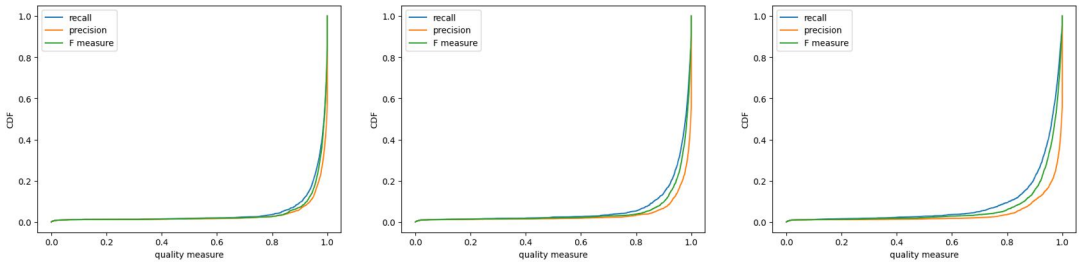
5.4 Results

We restrict the analysis to CVD trajectories that were collected on January 7, 2021 and that traversed a major corridor in the region, consisting of W Ina Rd, E Ina Rd, E Skyling Dr, and E Sunrise dr. This selection is based on the fact that several intersections on this corridor had the largest traffic volume by Miovision sensors. For example, W Ina Rd and N Thornydale Rd was the busiest corridor.

5.4.1 Impact of sampling rate

Given the wide impact of *sampling rate* on data cost and processing pipeline, we simulate various sampling frequencies by down-sampling the original data to evaluate the accuracy of speed estimation, correctness of stop identification, and quality of map matching. To emulate coarser datasets we down-sampled each trajectory to three levels of sampling intervals: 9 seconds, 15 seconds, and 30 seconds. As a comparison, typical LBS feeds arrive every 20 to 30 seconds, albeit inconsistently. Figure 5.4 displays recall, precision, and *F*-measure (Equation 5.4). At 9 s and 15 s intervals the performance curves of recall, precision, and F score in Figure 5.4 are virtually indistinguishable from the 3s baseline, and link-speed errors rise by less than 1 mph in MAE, as presented in Table 5.1. Those levels still resolve stop bars, short weaving sections, and midblock decelerations that matter for safety diagnostics. Once the spacing widens to 30 s, however, two problems emerge. Geometries are missed because an average passenger car traveling 35–40 mph covers 470–590 m in 30 s; this distance is long enough to skip an entire signalized intersection or a freeway off-ramp, so the map-matcher has to guess. Time spent while stopped smears to corresponding links. By spread dwell time at red lights across multiple links, travel times on these links are inflated and inferred speed are depressed. MAE nearly triples (4.7 mph) and the violin plot in Figure 5.4 grows fat around the diagonal, signaling larger directional bias. Temporal resolution seems okay to be relaxed to 15 s with almost no performance penalty; 30 s is a defensible ceiling.

Table 5.1 compares estimated link speeds with spot speeds observed in CVD, using RMSE and MAE as in Equation 5.17, as well as the 15th and 85th percentiles of the errors. The 3 second feed yields the lowest MAE 1.67 mph but the largest RMSE 44.3 mph, a perfect indicator of occasional phantom cruising, i.e., very short detours between nearly coincident pings



(a) Simulated sampling frequency: 9s **(b)** Simulated sampling frequency: 15s **(c)** Simulated sampling frequency: 30s

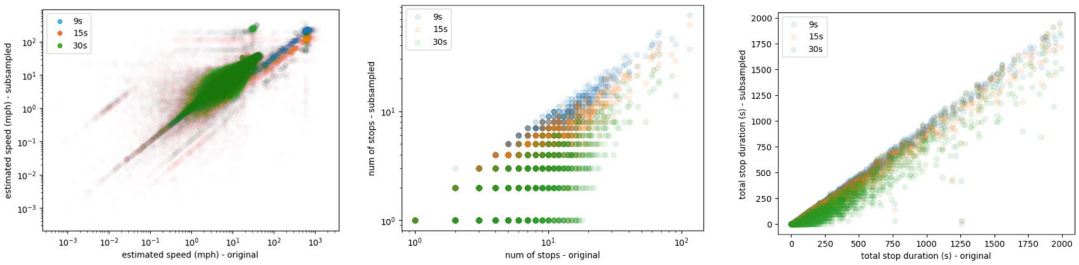
Figure 5.4: Map matching quality does not deteriorates significantly in 9s and 15s sampling rate (panels **a-b**), but drops when the rate is 30s (panel **c**). For each trip, recall, precision, and F measure can be calculated based on Equations 5.4 by comparing map matching results of downsampled trajectory against original CVD trajectory (three-second frequency). These metrics are weighted by lengths of the road links to accommodate their intrinsic differences. After collecting quality metrics per trip, cumulative distribution functions (CDFs) of each metric reveal the full spread of each metric: central tendency, variability, and long-tail behavior.

that imply unrealistic speeds. Figure 5.5 corroborates these findings: speeds on shared links, the number of detected stops, and stop-duration estimates are nearly identical for 3 s, 9 s, and 15 s, while 30 s begins to blur stops and deflate speeds. More specifically, at 9 s and 15 s the detour artifact diminishes, cutting RMSE to 18.8 mph and 13.6 mph, with only modest increases in MAE. 30 s sampling widens all error metrics (MAE 4.73 mph; RMSE 19.8 mph). Knowing that very high sampling can backfire on the map matching function, strategies exist to mitigate these phantom cruising issues accordingly: (1) use a heading filters to drop candidate edges whose bearing deviates > 30 degrees from the observed course;(2) apply context windows to score a candidate path not just on its own likelihood but on continuity with the previous N matched points; (3) impose speed caps to discard edges if implied speed exceeds three times the statutory limit.

$$\text{MAE} = \frac{1}{N}\sum_{i}^{N} |v_i - v_i^{\text{GT}}|, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N} |v_i - v_i^{\text{GT}}|^2} \tag{5.17}$$

Table 5.1: Estimated link speed vs. spot speed observations

Sampling rate (s)	RMSE (mph)	MAE (mph)	15th percentile (mph)	85th percentile (mph)
3	44.3	1.67	-0.63	1.57
9	18.8	2.04	-0.93	2.30
15	13.6	2.68	-1.77	3.35
30	19.8	4.73	-4.49	5.98



(a) Estimated speeds on shared links **(b)** Number of detected stops **(c)** Duration of detected stops

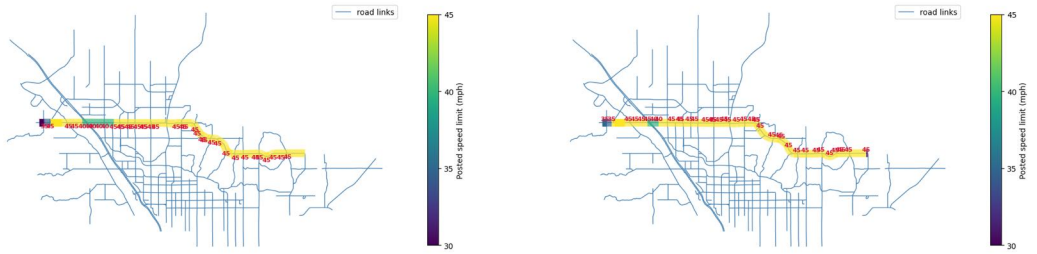
Figure 5.5: Link speed estimation quality also indicates the pipeline does not suffer from loss of sampling frequency until 30s. Panel **a** shows that estimated speeds of frequencies 9s and 15s exhibit controlled dispersion around the plane diagonal, while estimations of 30s sampling frequency presents a huge violin shape. It is also noticed that a subset of estimated speeds from the original sampling frequency are outlying by reaching more than 100 mph and these outliers are also seen for 9s and 15s frequencies, but not 30s frequency. While this indicates potential quality pitfalls in the map matching module, it also tells that behaviors of 9s and 15s stay closer to the original. The other two metrics center around stops that are detected at different sampling frequencies. Impact of low sampling frequency is clear as all data points lie in the lower right half of the plane, indicating different levels of losses. Both the number of detected stops (panel **b**) and duration of these complete stops (panel **c**) tell that behavior of sampling frequencies at 9s and 15s does not deviate from the original as much as that at 30s.

5.4.2 Speed limit detection

Speed limit detection should be modeled as a probabilistic problem, not a frozen catalog lookup, because every stage of the workflow—from field conditions to policy interpretation—contains parts that deterministic maps cannot capture. Firstly, the assumption that a limit is "fixed once recorded" fails as agencies roll out LED regulatory signs and short-lived work-zone plates; a crew can drop a 35 mph orange plate on a normally 55 mph link for the night shift, and the moment a traffic cone occludes the sign the detector's confidence drops. Sec-

ondly, leaning on an "already good enough" GIS layer ignores the routine audit finding that a meaningful share of limits are missing or mismatched because as-built drawings never track post-construction swaps or school-zone plaques. A vision detector can close that gap only if its confidence is calculated per detection; when confidence is low the system must fall back on the GIS prior. Thirdly, the idea that a static "90% accuracy is good enough" misses the core idea of compliance: if mean speed hovers around 47 mph, a slight misread will flip thousands of trips from "compliant" to "violation", pushing exceedance probability high. Propagating that speed limit sign detection and recognition uncertainty creates actionable signals. For example, in the night-shift work-zone example, the posterior becomes bimodal and the wide probability band itself is the alert that human input is needed; when a variable speed sign freezes blank while upstream LEDs fall to 45 mph, a low-confidence prior widens the compliance interval, simultaneously flagging a safety risk and diagnosing a hardware failure; by encoding time-of-day modifiers in the prior, a school zone link instantaneously shifts from 25 mph to 35 mph at 4pm, so the map never falsely alerts "speeding" after school hours. In summary, treating the speed limit detection as a probabilistic problem in our context has wide real-world implications.

The vision pipeline proposed in Chapter 4 delivers a solid performance on standard SLI dataset. It allows for versatile multi-modal input tailored for "speed limit sign" and successfully captures most instances (see Figure 5.6). The only failed cases were due to a combination of external factors. Firstly, geo-referenced speed limit map by Pima county oversimplifies the network: it only uses one polyline to represent a road segment, which is physically wide. This inevitably causes the matched SLI to end up in lanes of the other direction of travel in some cases, which results in the speed limit signs being small and distant in the field of view. And



(a) Posted speed limits on West to East driving direction

(b) Posted speed limits on East to West driving direction

Figure 5.6: Speed limits signs are recognized with previously proposed vision pipeline using SLIs acquired from third-party data vendor. Readings are extracted from these identified speed limit signs and propagated through the road network, assuming road segments between two speed limit signs are governed by the upstream speed limit sign. The chosen corridor features various driving contexts: urban vs. rural and plain vs. mountainous. Notice how certain parts of the corridor are filled with speed limit signs, indicating these segments might be the most speeding violations occur. The two panels show propogated speed limit in two driving directions.

due to sampling intervals and the driving speed of the survey vehicle, there is no guarantee of the appearance of speed limit signs in SLIs, often positioned and rotated. Lastly, occlusion could be serious for these wide segments as vegetation like trees and cactus can be planted in the median and if the SLI collection point is positioned on the wrong side of street, the algorithm must be able to look over or through them.

5.4.3 Speed compliance

Figure 5.8 compares empirical cumulative distribution functions of the compliance gap $\Delta L = v^{(b)} - L$ for the east-bound (EB) approach of the W Ina Rd & N Thornydale Rd corridor under four simulated fleet-penetration scenarios $\pi \in \{1, 2, 3.5, 5\}$ %. This segment connects the I-10 Frontage & W Ina Rd interchange to the signalized intersection at N Thornydale Rd. To reveal intra-segment heterogeneity this link is bisected at its approximately its mid-point (see Figure 5.7), yielding a *fast* upstream half (higher design speed, wider lane, no frontage access) and a *slower* downstream half (adjoining retail frontage, shorter block length). The full

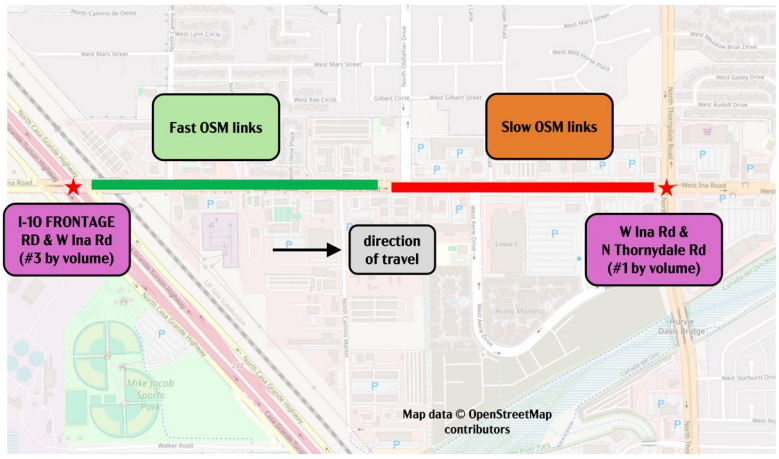
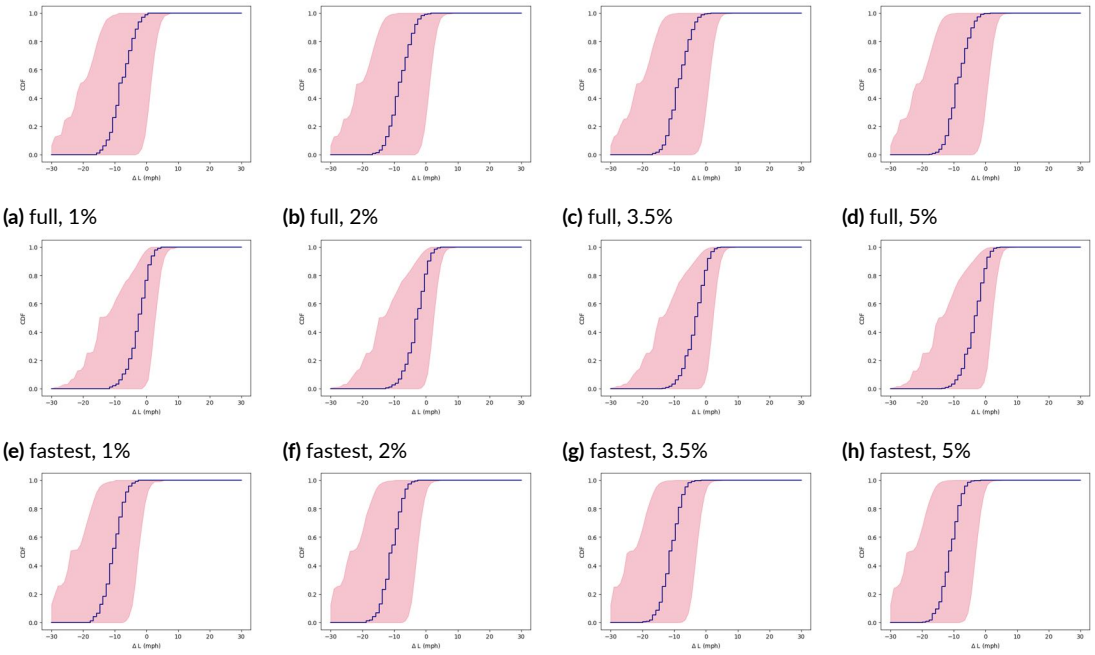


Figure 5.7: The segment connects the busiest intersection to the third busiest intersection by traffic volume according to installed Miovision sensors. First half of the segment is immediately downstream of I-10 Frontage Rd & W Ina Rd and treated as *fast* segment; second half is immediately upstream of the W Ina Rd & N Thronydale Rd. and denoted as *slow* segment.

segment and the slow half share similar speed exceedance profiles: regardless of bootstrapped penetration rate, their empirical CDFs clear at approximately 0, implying compliance. On the contrary, CDFs of the fast section shift right by 6-10 mph. At $\pi = 1\%$ the exceedance probability $P(\Delta > 0) \approx 0.25$, shrinking only slightly to ≈ 0.22 at $\pi = 5\%$. The interval between 10th percentile and 90th percentile is also tighter, indicating that under all simulated cases, the speed gap ΔL moves to positive region consistently fast.

Figure 5.9 and Figure 5.10 visualize distribution across bootstraps of the median ($\Delta_{50}^{(b)} L$) and 85th-percentile ($\Delta_{85}^{(b)} L$) compliance gaps for the same corridor, separating the fastest and slowest halves and contrasting the 1% versus 5% penetration scenarios. For the fastest half the posterior mean is $\mu_{\Delta_{50}} = +2.7$ mph at $\pi = 1\%$ and $+2.5$ mph at $\pi = 5\%$; the 95% credible interval shrinks only from ± 1.8 mph to ± 1.4 mph. By contrast the slow half centers at $\mu_{\Delta_{50}} = -7.4$ mph with a similarly narrow CI, confirming systematic under-speeding near the signalized intersection. The fast half records $\mu_{\Delta_{85}} = +7.6$ mph with CI $= [6.1, 9.1]$

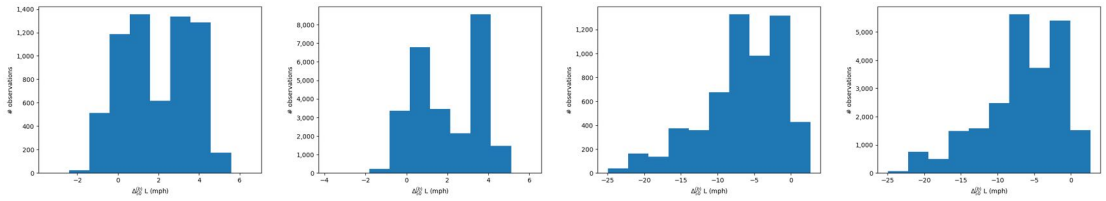


(a) full, 1% **(b)** full, 2% **(c)** full, 3.5% **(d)** full, 5%

(e) fastest, 1% **(f)** fastest, 2% **(g)** fastest, 3.5% **(h)** fastest, 5%

(i) slowest, 1% **(j)** slowest, 2% **(k)** slowest, 3.5% **(l)** slowest, 5%

Figure 5.8: Empirical cumulative distribution function of ΔL estimated for EB approach of W Ina Rd & N Thornydale Rd under different simulated penetration rates. Panels **a-d** present results on the full segment; panels **e-h** present results on the fastest half; panels **i-l** present results on the slowest half. X-axis ΔL calculates observed speed minus speed limit in miles per hour. Pink ribbon is area between 10th and 90th percentiles. Profiles of the *full* segment and the *slow* portion of the segment show almost 100% speed compliance, while profiles of the *fast* portion only demonstrate 75% compliance, i.e., $F_{\Delta}(x \mid \ell) \approx 0.75$.

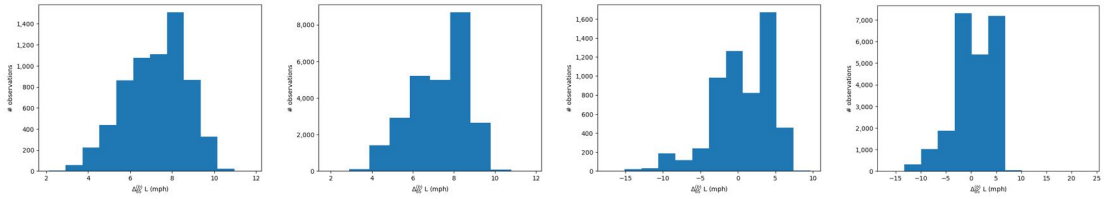


(a) fastest, 1% **(b)** fastest, 5% **(c)** slowest, 1% **(d)** slowest, 5%

Figure 5.9: Bootstrapped median speed exceedance $\Delta_{50}^{(b)}L$ of two different portions of the EB approach of W Ina Rd & N Thornydale Rd segment under two simulated penetration rates. Panels **a-b** present results on the fastest half of the segment, while panels **c-d** present results on the slowest half. Under both scenarios, i.e., 1% and 5%, median driving speeds on the fastest half of the EB approach could be approximately 8 mph higher than the slowest half and the fastest half witnesses around two to three mph median speeding above speed limits.

mph under 1% penetration and 6.8 mph (CI$= [5.9, 7.6]$ mph) under 5%. For the slowest half the distribution straddles zero, with $\mu_{\Delta_{85}} = -1.3$ mph and 46% of replicates yielding $\Delta_{85} > 0$. Even at the lowest penetration level, the Bayesian posterior separates the two halves by approximately 8 mph in both median and 85th percentile speed gap. The modest tightening of CIs from 1% to 5% penetration indicates diminishing marginal value in purchasing higher CVD penetration for corridor-level compliance diagnostics. These histogram-based summaries complement the CDF ribbons by revealing how often a corridor section would be classified as speeding when only a fraction of the connected fleet is observed—a key metric for agencies deciding whether low-cost (1% penetration) data are sufficient.

Figure 5.11 translates the link-level exceedance probabilities $P_{\text{exceed}}(\ell)$ into a geographic heat map for the corridor that includes W Ina Rd, E Ina Rd, E Skyline Dr, and E Sunrise Dr and that has the most highly ranked intersections by traffic volume collected by Miovision sensors. Results from ($\pi = 1\%$) and p($\pi = 5\%$) highlight almost identical hotspots—most prominently E Skyline Dr, east of E Ina Rd–despite the five-fold change in fleet coverage. Cross checking with Figure 5.6 reveals that hotspot segments overlap with wide-lane, low-access arterials, which are typical environments where static 45 mph limits conflict with



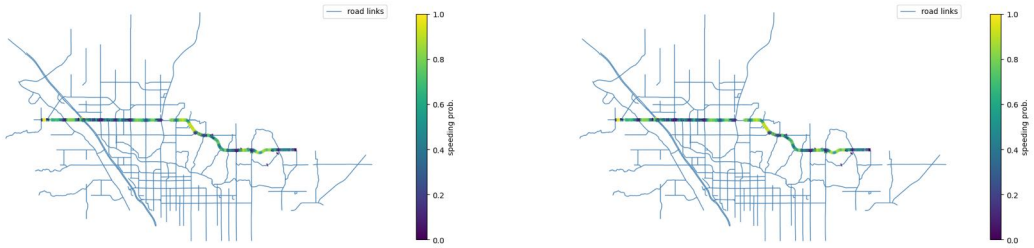
(a) fastest, 1% **(b)** fastest, 5% **(c)** slowest, 1% **(d)** slowest, 5%

Figure 5.10: Bootstrapped 85th percentile of speed exceedance $\Delta_{85}^{(b)}L$ of two different portions of the EB approach of W Ina Rd & N Thornydale Rd segment under two simulated penetration rates. Panels **a-b** present results on the fastest half of the segment, while panels **c-d** present results on the slowest half. Under both scenarios, 85th percentile of driving speeds on the fastest half of the EB approach could be approximately 4 mph higher than the slowest half and the fastest half witnesses around six to eight mph speeding above posted speed limits on average. Median of these 85th percentile speed gaps of the slowest segment is around and above 0, meaning approximately half of these 85th percentile speed observations are equal to corresponding speed limits.

perceived safe speed in traffic. SLIs further reveal sparse roadside development on E Skyline Dr, explaining driver bias towards higher speeds. Near-identical maps at 1% vs. 5% penetration mean agencies can run corridor-level screening with the cheapest CVD data license tier before investing resources to imagery audits for marked segments. It will be worthwhile to investigate how probabilistic speed compliance P_{exceed} compare with Star Ratings collected by iRAP protocols to improve their methodologies.

5.5 Discussions

This chapter demonstrated how high-frequency connected vehicle data (CVD) and street-level images (SLI) can be fused to produce statistically robust, infrastructure-aware speed compliance metrics that feed directly into network-scale safety assessments. We first reviewed CVD's capabilities and biases, then detailed a data processing pipeline that converts three-second telematics pings into link-level speed profiles, while preserving its identifiers so that space mean speeds be calculated as speed measure, single trajectory can be easily pieced back, and



(a) 1% penetration rate **(b)** 5% penetration rate

Figure 5.11: Speed compliance heatmap presents estimated speed exceeding probability along the selected corridor by simulating two CVD penetration rates: (a) 1% and (b) 5%. Several segments are identified by their persistent speeding above the posted speed limits, e.g., E Skyline Dr. The similarity between results obtained from 1% penetration rate and 5% penetration rate indicates the former is sufficient for this analysis and thus a cost-effective stream of input for network-level safety assessment.

simulation is made easy for bootstrapping and downsampling analysis as records are keyed by OSM links. Then these CVD observations and bootstrapped replicas join posted speed limits extracted by a zero-shot segmentation cascade, considering the probabilistic output of vision models. A Bayesian layer propagated two key uncertainties, i.e., fleet-penetration bias and sign detection confidence, to produce posterior distributions for median, 85th percentile, and "any-exceedance" speed gaps. Experiments on PAG's arterial network showed that (1) corridor-level hotspot rankings remain stable down to 1% fleet penetration, (2) the fastest half of a Ina Rd segment exceeds its 45 mph limit by 6–8 mph with ≥80% credibility, and (3) network-wide exceedance probabilities align with self-explaining road features visible in SLIs. The resulting link-level risk heat map integrates seamlessly with proactive safety assessments protocols, providing agencies a cost-effective, data-driven basis for prioritizing speed management interventions. By working out this degenerate example of the general case–our speed limit signs are static and do not include variable counterparts–we prove the soundness of the proposed probabilistic framework. As crowd-sourced SLI data streams become available and ubiquitous in the future such as those adopted by Abdel-Aty, Ugan, and Islam [1], the same

probabilistic framework can extend to far higher temporal cadences and geographic ranges, transforming what is now a periodic audit into a near-real-time crowd-sourced monitoring system.

6

Concluding remarks

6.1 Summary and contributions

Road safety progress in the United States has stalled: annual fatalities remain above 40,000 despite decades of engineering advances, a plateau that exposes the limits of analyzing crashes only after they happen. Because reactive studies based on police reports must wait for crashes to accumulate, such research offers limited guidance on where to intervene before the next

fatality occurs. The Safe System Approach flips that logic by proposing proactive analysis and countermeasures: the International Road Assessment Programme (iRAP) materializes about 50 variables in its Star Ratings protocol and scores every 100 meter of roadway on how effectively it accounts for inevitable human errors and their impacts. Within this framework, *Safe Roads*, allowing for a forgiving physical environment, is recognized as the bedrock of system safety, and thus accurate and robust perception of road environment elements is indispensable. By identifying segments of high risks in advance, agencies can establish clear safety metrics and implement cost-effective improvements before severe crashes would otherwise reveal the danger. Yet iRAP's rating workflow is still anchored to painstaking human coding, covering only a few dozen kilometers per day per annotator and inevitably subject to inconsistent qualities. Meanwhile, ubiquitous street-level imagery now captures the entire road network at regular intervals, offering an accessible visual record of infrastructure presence and conditions. Recent vision foundation models, able to segment and describe objects they are not trained on, open the door to automating road environment perception and thus safety codings at scale with consistent quality. Nonetheless, several gaps still block adapting these vision foundation models for full automation: foundation models like Segment Anything Model (SAM) yield class-agnostic masks that lack iRAP-specific labels, existing prompt recipes are not tuned for the cluttered, "stuff"-heavy roadside scenes that dominate safety assessments, and no off-the-shelf method yet injects the necessary infrastructure semantics into zero-shot segmentation while preserving the models' remarkable breadth and precision. As a result, this dissertation comprehensively tackles the automation gaps from three strategic vantage points: (1) prompt and prior knowledge design that injects iRAP semantics into zero-shot segmentation, (2) customization of these foundation models by fusing

class-agnostic masks with targeted multi-modal guidance, and (3) workflow integration that turns those advances into a scalable, tuning-free tool that can be applied to safety analysis. The methodological innovations, empirical performance gains, and practical pathways for agency adoption are discussed.

This dissertation pursues two main objectives. First, it diagnoses the automation gaps that limit iRAP, analyzing why current tools falter in both accuracy and scale when tasked with recognizing road infrastructure elements from street-level imagery. Second, it advances a unified methodological solution: reformulating infrastructure recognition as a semantic segmentation task and developing a scalable perception pipeline that infuses vision foundation models with domain-aware prompts and priors, thereby embedding iRAP semantics while preserving the models' zero-shot generalization. Additionally, the dissertation demonstrates the broader utility of the proposed approaches through complementary case studies: validating the vision pipeline as a stand-alone tool for automated coding of key infrastructure objects required in rating protocols, and extending its reach by merging the vision-derived inventories with external datasets to power richer, multi-modal safety analyses. By achieving these goals, this dissertation lays solid progress towards fully automated, data-driven road safety assessment, which will accelerate Safe System adoption, guide agencies toward proactive and cost-effective countermeasures, and open new research pathways in foundation model adaptation, uncertainty-aware perception, and infrastructure-behavior fusion.

Chapter 2 performs a comprehensive survey of three intersecting areas of work: current practices and attempts to automate iRAP Star Ratings, techniques for extracting infrastructure information from street-level imagery, and the evolution from classic deep learning models to recent open-vocabulary vision foundation models. The body of literature exposes a

persistent gap: existing methods either return coarse bounding boxes, rely on narrow closed vocabularies, or bypass interpretability by regressing safety scores end-to-end. None delivers the pixel-accurate, adaptable segmentation required to capture cluttered, scale-varying roadside assets, nor do current approaches produce outputs that map transparently onto iRAP infrastructure variables. The review also pinpoints the Segment Anything Model (SAM) as a most promising segmentation backbone due to its zero-shot, high-fidelity mask generation, and materializes the technical requirements that guide the rest of the dissertation: fine-grained masks, generalization to unseen classes, and explicit semantic linkage to Star Rating variables.

Chapter 3 introduces the an end-to-end framework that converts raw masks produced in SAM's "everything-mode" into segmentation maps containing critical infrastructure elements. The method begins with a grid sampler that places sparse points across the image and lets SAM carve out boundaries and produce class-agnostic object masks; it then injects domain priors by pixel-wise uncertainty maps and class probabilities from a segmentation network trained on domain datasets to vote for a label for each mask. A correlation study between SAM coverage and prior reliability motivates an iterative sampler that targets only the uncertain or uncovered regions in the first pass, lifting mask coverage for hard *stuff* classes while adding minimal sub-linear compute. On Mapillary Vistas, the hybrid system boosts evaluation metrics over domain baselines and closer to an oracle upper bound, demonstrating that crisp SAM boundaries and domain semantics can be fused without retraining. By furnishing pixel-accurate, interpretable masks in a compute-efficient loop, SAM with grid-based sampling and domain prior knowledge lays a practical basis for large-scale automated road environment perception.

Chapter 4 advances from the post-hoc labeling to prompt-driven inference by coupling Grounding DINO with SAM in a multi-modal cascade. The motivation originates from two limitations of the post-hoc voting strategy: persistent inferior coverage of "stuff" infrastructure classes and occasional, confidently wrong voting from the domain models. Engineered text descriptions of each infrastructure class are first ingested by the language-vision detector, whose deformable-attention heads yield box and point prompts appropriately sampled on the target object. These prompts steer SAM to generate class-specific masks during inference, eliminating the need for downstream voting and breaking the pipeline's dependence on potentially "confident-yet-wrong" domain models. Systematic analysis show that description length, visual cues, and prompt type jointly influence performance; the best recipe lifts IoU and coverage for historically under-segmented "stuff" classes, such as *guard rail* and *sidewalk*, beyond both the Chapter 3 approach and domain baselines. A real-world case study confirms the efficacy of the proposal: the pipeline pinpoints street light fixtures with reasonable performances without any task-specific retraining. By demonstrating that iRAP semantics can be injected directly through multi-modal prompts, this chapter delivers an even more flexible, label-free route to high-fidelity infrastructure inventories and paves the way for voice- or text-driven annotation tools.

Chapter 5 extends the vision pipeline into a safety analysis framework that couples its infrastructure inventories with large-scale connected vehicle data (CVD). While prior research has paired imagery or geometry with speed traces, none has merged the uncertainties on both sources into a single, coherent probabilistic framework: segmentation confidence on the perception outcomes and sampling or penetration bias in the CVD observations. This chapter fills that gap: posted speed signs detected via the **Chapter 4** pipeline are linked to map-

matched vehicular speeds from CVD inside a Bayesian Monte Carlo bootstrapping algorithm that jointly propagates speed image-level speed limit detection, fleet sampling, and observation frequency uncertainties. Applied to selected study corridors, the system outputs link-level posterior distributions for the speed compliance gap and the probability of systematic over-speeding, producing risk maps that pinpoint corridors where design and high operating speeds coincide. By unifying static infrastructure perception with dynamic behavior under an explicit uncertainty budget, this chapter showcases the broader analytical power unlocked when infrastructure segmentation outputs are woven into probabilistic, network-wide safety assessments.

In summary, this dissertation demonstrates how vision foundation models can transform road safety assessment from a slow, manual exercise into a scalable, zero-shot analytics pipeline. By diagnosing the limits of current iRAP automation, reformulating infrastructure recognition as a semantic segmentation task, and crafting domain-aware prompts and priors for the Segment Anything Model backbone, this work proposes a high-fidelity, scalable perception pipeline that turns raw street-level imagery into richly labeled road environment layers. Case studies confirm that these inventories can facilitate safety assessments as stand-alone inputs or be fused with additional data sources to produce uncertainty-aware risk maps, giving agencies earlier and clearer signals for intervention. The results underscore the value of multi-modal prompting, adaptive sampling, and probabilistic integration for robust road-environment perception, and they chart a path toward next-generation Safe System tools that operate at network scale.

6.2 Future works

This dissertation lays a solid foundation of adapting vision foundation models to road environment perception serving road safety assessments. Several potential extensions can be pursued. Firstly, more sophisticated task-specific changes of vision foundation models are expected to produce better performances. Subsequent studies can explore structural adaptations, e.g., lightweight task-specific decoders, cross-attention blocks that ingest geometric priors, or self-distillation schemes, to let the backbone internalize features of road infrastructure elements instead of injecting their semantics only at inference time. Selective fine-tuning on curated infrastructure tiles or parameter-efficient techniques like LoRA[78] and adapters have proved useful in other fields and thus may further boost recall on minority classes without eroding the models' valuable zero-shot capabilities. Secondly, because street-level imagery offers only a localized, ground-level view, fusing it with high-resolution satellite or aerial images could supply the plan-view geometry and elevation context that SLIs are short of, yielding more complete road design variables such as curvature, intersection type, presence of pedestrian facilities, number of lanes. These heterogeneous sources of emerging transportation big data promise the most efficient pathways for capturing complementary variables and, when pieced together, will enable an automated, end-to-end iRAP assessment workflow. Thirdly, it would be worthwhile to investigate how the probabilistic framework combining SLIs and CVD (as well as their pipelines) can be linked to iRAP Star Ratings. Given its current decision engines is not sufficiently evidence-based, such connection can potentially enhance iRAP's capacity to data-driven modeling and countermeasures.

References

[1] Abdel-Aty, M., Ugan, J., & Islam, Z. (2024). Exploring the influence of drivers' visual surroundings on speeding behavior. *Accident Analysis & Prevention*, 198, 107479.

[2] Abdelraouf, A., Abdel-Aty, M., & Mahmoud, N. (2022). Sequence-to-sequence recurrent graph convolutional networks for traffic estimation and prediction using connected probe vehicle data. *IEEE Transactions on Intelligent Transportation Systems*, 24(1), 1395–1405.

[3] Abou Chacra, D. & Zelek, J. (2016). Road segmentation in street view images using texture information. In *2016 13th Conference on Computer and Robot Vision (CRV)* (pp. 424–431).: IEEE.

[4] Abou Chacra, D. B. & Zelek, J. S. (2017). Fully automated road defect detection using street view images. In *2017 14th Conference on Computer and Robot Vision (CRV)* (pp. 353–360).: IEEE.

[5] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[6] Afrin, T. & Yodo, N. (2020). A survey of road traffic congestion measures towards a sustainable and resilient transportation system. *Sustainability*, 12(11), 4660.

[7] Ahmad, S., Ahmed, H. U., Ali, A., Yang, X., Huang, Y., Guo, M., Ren, Y., & Lu, P. (2024). Evaluating driving behavior patterns during wildfire evacuations in wildland-urban interface zones using connected vehicles data. *Fire Safety Journal*, 142, 104015.

[8] Ahmad, S., Ali, A., Ahmed, H. U., Huang, Y., & Lu, P. (2023). Evaluating traffic operation conditions during wildfire evacuation using connected vehicles data. *Fire*, 6(5), 184.

[9] Ahmadi, M., Lonbar, A. G., Sharifi, A., Beris, A. T., Nouri, M., & Javidi, A. S. (2023). Application of segment anything model for civil infrastructure defect assessment. *arXiv preprint arXiv:2304.12600*.

[10] Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, 58, 240–250.

[11] Alhasoun, F. & González, M. (2019). Streetify: using street view imagery and deep learning for urban streets development. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2001–2006).: IEEE.

[12] Anik, B., Islam, Z., & Abdel-Aty, M. (2023). intformer: A time-embedded attention-based transformer for crash likelihood prediction at intersections using connected vehicle data. *arXiv preprint arXiv:2307.03854*.

[13] Arun, A., Haque, M. M., Washington, S., & Mannering, F. (2023). A physics-informed road user safety field theory for traffic safety assessments applying artificial intelligence-based video analytics. *Analytic methods in accident research*, 37, 100252.

[14] Arunmozhi, A. & Park, J. (2018). Comparison of hog, lbp and haar-like features for on-road vehicle detection. In *2018 IEEE International Conference on Electro/Information Technology (EIT)* (pp. 0362–0367).: IEEE.

[15] Ash, J. E., Zou, Y., Lord, D., & Wang, Y. (2021). Comparison of confidence and prediction intervals for different mixed-poisson regression models. *Journal of Transportation Safety & Security*, 13(3), 357–379.

[16] Aykac, D., Karnowski, T., Ferrell, R., & Goddard, J. S. (2020). Detection and characterization of rumble strips in roadway video logs. *Electronic Imaging*, 32(6), 50–1–50–1.

[17] Bader, M. D., Mooney, S. J., Lee, Y. J., Sheehan, D., Neckerman, K. M., Rundle, A. G., & Teitler, J. O. (2015). Development and deployment of the computer assisted neighborhood visual assessment system (canvas) to measure health-related neighborhood conditions. *Health & place*, 31, 163–172.

[18] Bevly, D. M., Ryu, J., & Gerdes, J. C. (2006). Integrating ins sensors with gps measurements for continuous estimation of vehicle sideslip, roll, and tire cornering stiffness. *IEEE Transactions on Intelligent Transportation Systems*, 7(4), 483–493.

[19] Blasiis, M. R. D., Benedetto, A. D., Fiani, M., & Garozzo, M. (2019). Assessing the effect of pavement distresses by means of lidar technology. In *ASCE International Conference on Computing in Civil Engineering 2019* (pp. 146–153).: American Society of Civil Engineers Reston, VA.

[20] Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L., & Eckstein, L. (2020). The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1929–1934).: IEEE.

[21] Bradbury, K., Saboo, R., L Johnson, T., Malof, J. M., Devarajan, A., Zhang, W., M Collins, L., & G Newell, R. (2016). Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific data*, 3(1), 1–9.

[22] Brkić, I., Miler, M., Ševrović, M., & Medak, D. (2022). Automatic roadside feature detection based on lidar road cross section images. *Sensors*, 22(15), 5510.

[23] Brkić, I., Ševrović, M., Medak, D., & Miler, M. (2023a). Utilizing high resolution satellite imagery for automated road infrastructure safety assessments. *Sensors*, 23(9), 4405.

[24] Brkić, I., Ševrović, M., Medak, D., & Miler, M. (2023b). Utilizing high resolution satellite imagery for automated road infrastructure safety assessments. *Sensors*, 23(9), 4405.

[25] Brock, E., Huang, C., Wu, D., & Liang, Y. (2021). Lidar-based real-time mapping for digital twin development. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6).: IEEE.

[26] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2019). nuscenes: A multimodal dataset for autonomous driving. *CoRR*, abs/1903.11027.

[27] Cai, Q., Abdel-Aty, M., Zheng, O., & Wu, Y. (2022). Applying machine learning and google street view to explore effects of drivers' visual environment on traffic safety. *Transportation research part C: emerging technologies*, 135, 103541.

[28] Campos, E., Wang, A., & Zong, T. (2023). How many trajectories should be collected to evaluate signal coordination performance? cluster analysis using connected vehicle data. *Institute of Transportation Engineers. ITE Journal*, 93(12), 39–46.

[29] Cao, X., Wu, C., Yan, P., & Li, X. (2011). Linear svm classification using boosting hog features for vehicle detection in low-altitude airborne videos. In *2011 18th IEEE international conference on image processing* (pp. 2421–2424).: IEEE.

[30] Cao, Y., Tang, K., Sun, J., & Ji, Y. (2021). Day-to-day dynamic origin–destination flow estimation using connected vehicle trajectories and automatic vehicle identification data. *Transportation Research Part C: Emerging Technologies*, 129, 103241.

[31] Chai, S., Jain, R. K., Teng, S., Liu, J., Li, Y., Tateyama, T., & Chen, Y.-w. (2023). Ladder fine-tuning approach for sam integrating complementary network. *arXiv preprint arXiv:2306.12737*.

[32] Chao, P., Xu, Y., Hua, W., & Zhou, X. (2020). A survey on map-matching algorithms. In *Databases Theory and Applications: 31st Australasian Database Conference, ADC 2020, Melbourne, VIC, Australia, February 3–7, 2020, Proceedings 31* (pp. 121–133).: Springer.

[33] Chen, C. & Xie, Y. (2016). Modeling the effects of aadt on predicting multiple-vehicle crashes at urban and suburban signalized intersections. *Accident Analysis & Prevention*, 91, 72–83.

[34] Chen, H., Song, J., & Yokoya, N. (2024a). Change detection between optical remote sensing imagery and map data via segment anything model (sam). *arXiv preprint arXiv:2401.09019*.

[35] Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., & Shi, Z. (2024b). Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.

[36] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

[37] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

[38] Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

[39] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 801–818).

[40] Chen, T., Mai, Z., Li, R., & Chao, W.-l. (2023). Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*.

[41] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1290–1299).

[42] Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34, 17864–17875.

[43] Coifman, B. & Kim, S. (2009). Speed estimation and length based vehicle classification from freeway single-loop detectors. *Transportation research part C: emerging technologies*, 17(4), 349–364.

[44] Çolak, S., Lima, A., & González, M. C. (2016). Understanding congested travel in urban areas. *Nature communications*, 7(1), 10793.

[45] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

[46] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).: Ieee.

[47] Dennis, E. P., Hong, Q., Wallace, R., Tansil, W., & Smith, M. (2014). Pavement condition monitoring with crowdsourced connected vehicle data. *Transportation Research Record*, 2460(1), 31–38.

[48] Desai, J., Li, H., Mathew, J. K., Cheng, Y.-T., Habib, A., & Bullock, D. M. (2021a). Correlating hard-braking activity with crash occurrences on interstate construction projects in indiana. *Journal of big data analytics in transportation*, 3(1), 27–41.

[49] Desai, J., Saldivar-Carranza, E., Mathew, J. K., Li, H., Platte, T., & Bullock, D. (2021b). Methodology for applying connected vehicle data to evaluate impact of interstate construction work zone diversions. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (pp. 4035–4042).: IEEE.

[50] Du, Y., Weng, Z., Liu, C., & Wu, D. (2020). Dynamic pavement distress image stitching based on fine-grained feature matching. *Journal of Advanced Transportation*, 2020, 1–15.

[51] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303–338.

[52] Federal Highway Administration (FHWA) (2023). Zero Deaths and Safe System. https://highways.dot.gov/safety/zero-deaths. Accessed: 2024-01-16.

[53] Gal, Y. & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).: PMLR.

[54] Ghazouali, S. E., Mhirit, Y., Oukhrid, A., Michelucci, U., & Nouira, H. (2024). Fusionvision: A comprehensive approach of 3d object reconstruction and segmentation from rgb-d cameras using yolo and fast segment anything. *arXiv preprint arXiv:2403.00175*.

[55] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

[56] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

[57] Gloudemans, D., Wang, Y., Ji, J., Zachar, G., Barbour, W., Hall, E., Cebelak, M., Smith, L., & Work, D. B. (2023a). I-24 motion: An instrument for freeway traffic science. *Transportation Research Part C: Emerging Technologies*, 155, 104311.

[58] Gloudemans, D., Wang, Y., Ji, J., Zachar, G., Barbour, W., & Work, D. B. (2023b). I-24 motion: An instrument for freeway traffic science. *arXiv preprint arXiv:2301.11198*.

[59] Godard, C., Aodha, O. M., & Brostow, G. J. (2016). Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677.

[60] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature*, 453(7196), 779–782.

[61] Groves, R. M. & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, 72(2), 167–189.

[62] Haghani, A., Hamedi, M., & Sadabadi, K. F. (2009). I-95 corridor coalition vehicle probe project: Validation of inrix data. *I-95 Corridor Coalition*, 9.

[63] Haklay, M. & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4), 12–18.

[64] Han, D., Zhang, C., Qiao, Y., Qamar, M., Jung, Y., Lee, S., Bae, S.-H., & Hong, C. S. (2023). Segment anything model (sam) meets glass: Mirror and transparent objects cannot be easily detected. *arXiv preprint arXiv:2305.00278*.

[65] Hasan, M. M. (2020). *Application of Artificial Intelligence and Geographic Information System for Developing Automated Walkability Score*. PhD thesis, Western Michigan University.

[66] He, H., Shao, Z., & Tan, J. (2015). Recognition of car makes and models from a single traffic-camera image. *IEEE Transactions on Intelligent Transportation Systems*, 16(6), 3182–3192.

[67] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017a). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

[68] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017b). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

[69] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

[70] He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P. E., & Ou, Y. (2023). Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets.

[71] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE transactions on geoscience and remote sensing*, 60, 1–15.

[72] Henrickson, K. (2018). *A Framework for Understanding and Addressing Bias and Sparsity in Mobile Location-Based Traffic Data*. PhD thesis, University of Washington.

[73] Hetang, C., Xue, H., Le, C., Yue, T., Wang, W., & He, Y. (2024). Segment anything model for road network graph extraction. *arXiv preprint arXiv:2403.16051*.

[74] Heydari, S., Miranda-Moreno, L. F., & Fu, L. (2020). Is speeding more likely during weekend night hours? evidence from sensor-collected data in montreal. *Canadian Journal of Civil Engineering*, 47(9), 1046–1049.

[75] Hoh, B., Gruteser, M., Herring, R., Ban, J., Work, D., Herrera, J.-C., Bayen, A. M., Annavaram, M., & Jacobson, Q. (2008). Virtual trip lines for distributed privacy-preserving traffic monitoring. In *Proceedings of the 6th international conference on Mobile systems, applications, and services* (pp. 15–28).

[76] Hosseini, M., Sevtsuk, A., Miranda, F., Cesar Jr, R. M., & Silva, C. T. (2023). Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101, 101950.

[77] Hou, Q. & Ai, C. (2020). A network-level sidewalk inventory method using mobile lidar and deep learning. *Transportation research part C: emerging technologies*, 119, 102772.

[78] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[79] International Road Assessment Programme (2022). irap coding manual drive on the right edition. https://irap.org/specifications/. Accessed: 2023-07-03.

[80] Islam, Z. & Abdel-Aty, M. (2023a). Traffic conflict prediction using connected vehicle data. *Analytic Methods in Accident Research*, 39, 100275.

[81] Islam, Z. & Abdel-Aty, M. (2023b). Traffic conflict prediction using connected vehicle data. *Analytic methods in accident research*, 39, 100275.

[82] Islam, Z., Abdel-Aty, M., & Anik, B. T. H. (2023a). Transformer-conformer ensemble for crash prediction using connected vehicle trajectory data. *IEEE Open Journal of Intelligent Transportation Systems*.

[83] Islam, Z., Abdel-Aty, M., Anwari, N., & Islam, M. R. (2023b). Understanding the impact of vehicle dynamics, geometric and non-geometric roadway attributes on surrogate safety measure using connected vehicle data. *Accident Analysis & Prevention*, 189, 107125.

[84] Islam, Z., Abdel-Aty, M., & Ugan, J. (2024). Signal phasing and timing prediction using connected vehicle data. *Transportation research record*, 2678(1), 662–673.

[85] Ito, K. & Biljecki, F. (2021). Assessing bikeability with street view imagery and computer vision. *Transportation research part C: emerging technologies*, 132, 103371.

[86] Jalilifar, E., Li, X., Martin, M., & Huang, X. (2024). Exploring the potential of market-available connected vehicle data in border crossing time estimation. *Transactions in Urban Data, Science, and Technology*, (pp. 27541231241226730).

[87] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of yolo algorithm developments. *Procedia Computer Science*, 199, 1066–1073.

[88] Jiang, S., Ferreira, J., & González, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25, 478–510.

[89] Jiang, S., Ferreira, J., & Gonzalez, M. C. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2), 208–219.

[90] Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E., & González, M. C. (2013). A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing* (pp. 1–9).

[91] Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., & González, M. C. (2016). The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37), E5370–E5378.

[92] Jung, J., Olsen, M. J., Hurwitz, D. S., Kashani, A. G., & Buker, K. (2018). 3d virtual intersection sight distance analysis using lidar data. *Transportation research part C: emerging technologies*, 86, 563–579.

[93] Kačan, M., Oršić, M., Šegvić, S., & Ševrović, M. (2020). Multi-task learning for irap attribute classification and road safety assessment. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1–6).: IEEE.

[94] Kandiboina, R., Knickerbocker, S., Bhagat, S., Hawkins, N., & Sharma, A. (2023). Exploring the efficacy of large-scale connected vehicle data in real-time traffic applications. *Transportation Research Record*, (pp. 03611981231191512).

[95] Kang, Y., Zhang, F., Gao, S., Lin, H., & Liu, Y. (2020). A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS*, 26(3), 261–275.

[96] Ke, R., Lutin, J., Spears, J., & Wang, Y. (2017). A cost-effective framework for automated vehicle-pedestrian near-miss detection through onboard monocular vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 25–32).

[97] Khadka, S., Li, P. ., & Wang, Q. (2022). Developing novel performance measures for traffic congestion management and operational planning based on connected vehicle data. *Journal of Urban Planning and Development*, 148(2), 04022016.

[98] Kilani, O., Gouda, M., Weiß, J., & El-Basyouny, K. (2021). Safety assessment of urban intersection sight distance using mobile lidar data. *Sustainability*, 13(16), 9259.

[99] Kim, S. & Coifman, B. (2014a). Comparing inrix speed data against concurrent loop detector stations over several months. *Transportation Research Part C: Emerging Technologies*, 49, 59–72.

[100] Kim, S. & Coifman, B. (2014b). Comparing inrix speed data against concurrent loop detector stations over several months. *Transportation Research Part C: Emerging Technologies*, 49, 59–72.

[101] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023a). Segment anything. *arXiv preprint arXiv:2304.02643*.

[102] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023b). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015–4026).

[103] Kocur, V. & Ftáčnik, M. (2021). Traffic camera calibration via vehicle vanishing point detection. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30* (pp. 628–639).: Springer.

[104] Krajewski, R., Bock, J., Kloeker, L., & Eckstein, L. (2018). The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st international conference on intelligent transportation systems (ITSC)* (pp. 2118–2125).: IEEE.

[105] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6), 066138.

[106] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[107] Kubička, M., Cela, A., Moulin, P., Mounier, H., & Niculescu, S.-I. (2015). Dataset for testing and training of map-matching algorithms. In *2015 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1088–1093).: IEEE.

[108] Li, H., Mathew, J. K., Kim, W., & Bullock, D. M. (2020). Using crowdsourced vehicle braking data to identify roadway hazards.

[109] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730–19742).: PMLR.

[110] Li, S., Cao, J., Ye, P., Ding, Y., Tu, C., & Chen, T. (2024). Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*.

[111] Li, X., Xu, P., & Wu, Y.-J. (2022). Pedestrian crossing volume estimation at signalized intersections using bayesian additive regression trees. *Journal of Intelligent Transportation Systems*, 26(5), 557–571.

[112] Li, Y., Wang, H., Duan, Y., Zhang, J., & Li, X. (2025). A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, (pp. 111409).

[113] Lin, Y.-C., Bullock, D., & Habib, A. (2020). Mapping roadway drainage ditches using mobile lidar. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 187–192.

[114] Lin, Y.-C. & Habib, A. (2022). Semantic segmentation of bridge components and road infrastructure from mobile lidar data. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 6, 100023.

[115] Liu, C., Yang, H., Ke, R., Sun, W., Wang, J., & Wang, Y. (2023a). Cooperative and comprehensive multi-task surveillance sensing and interaction system empowered by edge artificial intelligence. *Transportation Research Record*, (pp. 03611981231160174).

[116] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. (2024). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision* (pp. 38–55).: Springer.

[117] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21–37).: Springer.

[118] Liu, Y. & Dong, B. (2024). Modeling urban scale human mobility through big data analysis and machine learning. In *Building Simulation*, volume 17 (pp. 3–21).: Springer.

[119] Liu, Z., Borlaug, B., Meintz, A., Neuman, C., Wood, E., & Bennett, J. (2023b). Data-driven method for electric vehicle charging demand analysis: Case study in virginia. *Transportation Research Part D: Transport and Environment*, 125, 103994.

[120] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

[121] Loo, B. P., Fan, Z., Lian, T., & Zhang, F. (2023). Using computer vision and machine learning to identify bus safety risk factors. *Accident Analysis & Prevention*, 185, 107017.

[122] Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1), 654.

[123] Mahmoudi, J., Xiong, C., Yang, M., & Luo, W. (2023). Modeling the frequency of pedestrian and bicyclist crashes at intersections: big data-driven evidence from maryland. *Transportation research record*, 2677(3), 1245–1260.

[124] Mathew, J. K., Desai, J. C., Sakhare, R. S., Kim, W., Li, H., & Bullock, D. M. (2021). Big data applications for managing roadways. *Institute of Transportation Engineers. ITE Journal*, 91(2), 28–35.

[125] Miovision (2024). Video detection solutions. https://miovision.com/solutions/video-detection/. Accessed: 2024-05-18.

[126] Morgan, R. (2022). Cdc tracked 20+ million us phones to monitor covid lockdown compliance says new report. https://americanmilitarynews.com/2022/05/cdc-tracked-20-million-us-phones-to-monitor-covid-lockdown-compliance-says-new-report/. Accessed: 2024-05-11.

[127] Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., & de Albuquerque, V. H. C. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12), 22694–22715.

[128] Mussah, A. R. & Adu-Gyamfi, Y. (2022). Machine learning framework for real-time assessment of traffic safety utilizing connected vehicle data. *Sustainability*, 14(22), 15348.

[129] National Highway Traffic Safety Administration (2023). Nhtsa estimates for 2022 show roadway fatalities remain flat after two years of dramatic increases. https://www.nhtsa.gov/press-releases/traffic-crash-death-estimates-2022. Accessed: 2024-01-10.

[130] National Roadway Safety Strategy (NRSS) (2023). The Roadway Safety Problem. https://www.transportation.gov/NRSS/SafetyProblem. Accessed: 2024-01-16.

[131] Neuhold, G., Ollmann, T., Rota Bulo, S., & Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision* (pp. 4990–4999).

[132] Newson, P. & Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 336–343).

[133] Ning, H., Ye, X., Chen, Z., Liu, T., & Cao, T. (2022). Sidewalk extraction using aerial and street view images. *Environment and Planning B: Urban Analytics and City Science*, 49(1), 7–22.

[134] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[135] O'Shea, K. & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

[136] Paipuri, M., Xu, Y., Gonzalez, M. C., & Leclercq, L. (2021). Calibration of multi-region mfd models using mobile phone data. In *hEART 2020, 9th Symposium of the European Association for Research in Transportation-Virtual conference* (pp. 12p).

[137] Pan, Y., Darzi, A., Kabiri, A., Zhao, G., Luo, W., Xiong, C., & Zhang, L. (2020). Quantifying human mobility behaviour changes during the covid-19 outbreak in the united states. *Scientific Reports*, 10(1), 20742.

[138] Pan, Y., Sun, Q., Yang, M., Darzi, A., Zhao, G., Kabiri, A., Xiong, C., & Zhang, L. (2023). Residency and worker status identification based on mobile device location data. *Transportation Research Part C: Emerging Technologies*, 146, 103956.

[139] Peng, Z., Xu, Z., Zeng, Z., Xie, L., Tian, Q., & Shen, W. (2023). Parameter efficient fine-tuning via cross block orchestration for segment anything model. *arXiv preprint arXiv:2311.17112*.

[140] Peng, Z., Xu, Z., Zeng, Z., Yang, X., & Shen, W. (2024). Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 (pp. 4515–4523).

[141] Porzi, L., Rota Bulò, S., & Kontschieder, P. (2021). Improving panoptic segmentation at all scales. In *Computer Vision and Pattern Recognition (CVPR)*.

[142] Qi, X., Chen, Q., Jia, J., & Koltun, V. (2018). Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8808–8816).

[143] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).: PMLR.

[144] Rahman, M. M., Sainju, A. M., Yan, D., & Jiang, Z. (2021). Mapping road safety barriers across street view image sequences: A hybrid object detection and recurrent model. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (pp. 47–50).

[145] Rajič, F., Ke, L., Tai, Y.-W., Tang, C.-K., Danelljan, M., & Yu, F. (2023). Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*.

[146] Redmon, J. & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[147] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

[148] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

[149] Sakhare, R. S., Desai, J. C., Mahlberg, J., Mathew, J. K., Kim, W., Li, H., McGregor, J. D., & Bullock, D. M. (2021). Evaluation of the impact of queue trucks with navigation alerts using connected vehicle data. *Journal of Transportation Technologies*, 11(4), 561–576.

[150] Saldivar-Carranza, E., Li, H., Mathew, J., Hunter, M., Sturdevant, J., & Bullock, D. M. (2021a). Deriving operational traffic signal performance measures from vehicle trajectory data. *Transportation research record*, 2675(9), 1250–1264.

[151] Saldivar-Carranza, E., Li, H., Taylor, M., & Bullock, D. M. (2022a). Continuous flow intersection performance measures using connected vehicle data. *Journal of Transportation Technologies*, 12(4), 861–875.

[152] Saldivar-Carranza, E., Mathew, J. K., Li, H., & Bullock, D. M. (2021b). Roundabout performance analysis using connected vehicle data. *Journal of Transportation Technologies*, 12(1), 42–58.

[153] Saldivar-Carranza, E., Rogers, S., Li, H., & Bullock, D. M. (2022b). Diamond interchange performance measures using connected vehicle data. *Journal of Transportation Technologies*, 12(3), 475–497.

[154] Saldivar-Carranza, E. D., Hunter, M., Li, H., Mathew, J., & Bullock, D. M. (2021c). Longitudinal performance assessment of traffic signal system impacted by long-term interstate construction diversion using connected vehicle data. *Journal of transportation technologies*, 11(4), 644–659.

[155] Saldivar-Carranza, E. D., Li, H., & Bullock, D. M. (2021d). Diverging diamond interchange performance measures using connected vehicle data. *Journal of Transportation Technologies*, 11(4), 628–643.

[156] Saldivar-Carranza, E. D., Li, H., & Bullock, D. M. (2021e). Identifying vehicle turning movements at intersections from trajectory data. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (pp. 4043–4050).: IEEE.

[157] Saldivar-Carranza, E. D., Li, H., Platte, T., & Bullock, D. M. (2023). Systemwide identification of signal retiming opportunities with connected vehicle data to reduce split failures. *Transportation Research Record*, 2677(12), 587–603.

[158] Saldivar-Carranza, E. D., Mathew, J. K., Li, H., Hunter, M., Platte, T., & Bullock, D. M. (2021f). Using connected vehicle data to evaluate traffic signal performance and driver behavior after changing left-turns phasing. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (pp. 4028–4034).: IEEE.

[159] Savathrakis, G. & Argyros, A. (2024). An automated method for the creation of oriented bounding boxes in remote sensing ship detection datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 830–839).

[160] Schrank, D., Eisele, B., Lomax, T., et al. (2019). Urban mobility report 2019.

[161] Shams, A., Sarasua, W. A., Russell, B. T., Davis, W. J., Post, C., Rastiveis, H., Famili, A., & Cassule, L. (2023). Extracting highway cross slopes from airborne and mobile lidar point clouds. *Transportation research record*, 2677(2), 372–384.

[162] Siddique, C. & Ban, X. J. (2019). State-dependent self-adaptive sampling (sas) method for vehicle trajectory data. *Transportation Research Part C: Emerging Technologies*, 100, 224–237.

[163] Smadi, O., Hawkins, N., Hans, Z., Bektaş, B. A., Knickerbocker, S., Nlenanya, I., Souleyrette, R., & Hallmark, S. (2015). *Naturalistic driving study: Development of the roadway information database*. Technical Report SHRP 2 Report S2-S04A-RW-1.

[164] Solari, L., Del Soldato, M., Raspini, F., Barra, A., Bianchini, S., Confuorto, P., Casagli, N., & Crosetto, M. (2020). Review of satellite interferometry for landslide detection in italy. *Remote Sensing*, 12(8), 1351.

[165] Song, W., Workman, S., Hadzic, A., Zhang, X., Green, E., Chen, M., Souleyrette, R., & Jacobs, N. (2018). Farsa: Fully automated roadway safety assessment. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 521–529).: IEEE.

[166] Song, X., Wang, P., Zhou, D., Zhu, R., Guan, C., Dai, Y., Su, H., Li, H., & Yang, R. (2019). Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[167] Song, X., Wu, J., Zhang, H., & Pi, R. (2020). Analysis of crash severity for hazard material transportation using highway safety information system data. *SAGE Open*, 10(3), 2158244020939924.

[168] Sultan, R. I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., & Zhu, D. (2023). Geosam: Fine-tuning sam with sparse and dense visual prompting for automated segmentation of mobility infrastructure. *arXiv preprint arXiv:2311.11319*.

[169] Tan, X., Chen, G., Wang, T., Wang, J., & Zhang, X. (2023). Segment change model (scm) for unsupervised change detection in vhr remote sensing images: a case study of buildings. *arXiv preprint arXiv:2312.16410*.

[170] Tardy, H., Soilán, M., Martín-Jiménez, J. A., & González-Aguilera, D. (2023a). Automatic road inventory using a low-cost mobile mapping system and based on a semantic segmentation deep learning model. *Remote Sensing*, 15(5), 1351.

[171] Tardy, H., Soilán, M., Martín-Jiménez, J. A., & González-Aguilera, D. (2023b). Automatic road inventory using a low-cost mobile mapping system and based on a semantic segmentation deep learning model. *Remote Sensing*, 15(5), 1351.

[172] Tavafoghi, H., Porter, J., Flores, C., Poolla, K., & Varaiya, P. (2021). Queue length estimation from connected vehicles with low and unknown penetration level. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (pp. 1217–1224).: IEEE.

[173] Transportation Research Board (2010). *HCM 2010 : highway capacity manual*. Washington, DC: National Academy of Science.

[174] Ugan, J., Abdel-Aty, M., & Islam, Z. (2023). Using connected vehicle trajectory data to evaluate the effects of speeding. *IEEE Open Journal of Intelligent Transportation Systems*.

[175] Uğurel, E., Guan, X., Wang, Y., Huang, S., Wang, Q., & Chen, C. (2024). Correcting missingness in passively-generated mobile data with multi-task gaussian processes. *Transportation Research Part C: Emerging Technologies*, 161, 104523.

[176] United Nations (UN) (2022). Voluntary global performance targets for road safety risk factors and service delivery mechanisms and corresponding indicators. https://cdn.who.int/media/docs/default-source/documents/un-road-safety-collaboration/targets-and-indicators-visual-clean.pdf?sfvrsn=29627bde_5. Accessed: 2023-07-18.

[177] University of Arizona (2024). *Comparative Analysis and Integration of Region-Wide Traffic Data*. Technical report, University of Arizona. Accessed: 2024-05-18.

[178] Vlachogiannis, D. M., Moura, S., & Macfarlane, J. (2023). Intersense: An xgboost model for traffic regulator identification at intersections through crowdsourced gps data. *Transportation research part C: emerging technologies*, 151, 104112.

[179] Wang, D., Zhang, J., Du, B., Xu, M., Liu, L., Tao, D., & Zhang, L. (2024a). Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36.

[180] Wang, F. & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87, 58–74.

[181] Wang, F., Wang, J., Cao, J., Chen, C., & Ban, X. J. (2019). Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies*, 105, 183–202.

[182] Wang, P., Yang, R., Cao, B., Xu, W., & Lin, Y. (2018). Dels-3d: Deep localization and segmentation with a 3d semantic map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5860–5869).

[183] Wang, X., Jerome, Z., Wang, Z., Zhang, C., Shen, S., Kumar, V. V., Bai, F., Krajewski, P., Deneau, D., Jawad, A., et al. (2024b). Traffic light optimization with low penetration rate vehicle trajectory data. *Nature communications*, 15(1), 1306.

[184] Wang, X., Jerome, Z., Zhang, C., Shen, S., Kumar, V. V., & Liu, H. X. (2023). Trajectory data processing and mobility performance evaluation for urban traffic networks. *Transportation Research Record*, 2677(3), 355–370.

[185] Wang, X., Tong, J., Zong, W., Lv, Y., & Shen, J. (2024c). Trip misreporting mining and expansion method for household travel survey. *Transportation research part A: policy and practice*, 182, 104015.

[186] Wang, Z., Li, H., & Rajagopal, R. (2020). Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (pp. 1013–1020).

[187] Washington State Department of Transportation (2021). State route imagery collection. https://wsdot.wa.gov/about/transportation-data/roadway-data/state-route-imagery-collection. Accessed: 2025-04-01.

[188] Wei, C., Tan, H., Zhong, Y., Yang, Y., & Ma, L. (2024). Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*.

[189] Weld, G., Jang, E., Li, A., Zeng, A., Heimerl, K., & Froehlich, J. E. (2019). Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 196–209).

[190] Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation*, 42, 597–623.

[191] Wijnands, J. S., Zhao, H., Nice, K. A., Thompson, J., Scully, K., Guo, J., & Stevenson, M. (2021). Identifying safe intersection design through unsupervised feature extraction from satellite imagery. *Computer-Aided Civil and Infrastructure Engineering*, 36(3), 346–361.

[192] Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)* (pp. 3645–3649).: IEEE.

[193] World Health Organization (WHO) (2021). Road Traffic Deaths Data by Country. https://apps.who.int/gho/data/view.main.51310?lang=en. Accessed: 2024-01-16.

[194] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.

[195] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34, 12077–12090.

[196] Xie, J., Yang, C., Xie, W., & Zisserman, A. (2024). Moving object segmentation: All you need is sam (and flow). *arXiv preprint arXiv:2404.12389*.

[197] Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al. (2023). Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*.

[198] Xu, M., Su, J., & Liu, Y. (2023a). Aquasam: Underwater image foreground segmentation. *arXiv preprint arXiv:2308.04218*.

[199] Xu, R., Zhang, S., Xiong, P., Lin, A. Y., & Ma, J. (2022). Towards better driver safety: Empowering personal navigation technologies with road safety awareness. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 3004–3009).: IEEE.

[200] Xu, Y., Clemente, R. D., & González, M. C. (2021). Understanding vehicular routing behavior with location-based service data. *EPJ Data Science*, 10(1), 1–17.

[201] Xu, Y., Olmos, L. E., Mateo, D., Hernando, A., Yang, X., & González, M. C. (2023b). Urban dynamics through the lens of human mobility. *Nature computational science*, 3(7), 611–620.

[202] Yang, D., Xie, K., Ozbay, K., & Yang, H. (2019). Exploring spatial and temporal patterns of large-scale smartphone-based dangerous driving event data. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 116–121).: IEEE.

[203] Yang, M., Pan, Y., Darzi, A., Ghader, S., Xiong, C., & Zhang, L. (2022). A data-driven travel mode share estimation framework based on mobile device location data. *Transportation*, 49(5), 1339–1383.

[204] Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., & Loy, C. C. (2024). Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint arXiv:2401.02955*.

[205] Yuan, R., Xiang, Q., Huang, Y., & Gu, X. (2023). Investigating the difference in factors influencing the injury severity between daytime and nighttime speeding-related crashes. *Canadian Journal of Civil Engineering*, 51(1), 60–72.

[206] Zeng, Z., Zhu, W., Ke, R., Ash, J., Wang, Y., Xu, J., & Xu, X. (2017). A generalized nonlinear model-based mixed multinomial logit approach for crash data analysis. *Accident Analysis & Prevention*, 99, 51–65.

[207] Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., & Hong, C. S. (2023a). Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

[208] Zhang, F., Wegner, J. D., Yang, B., & Liu, Y. (2023b). Street-level imagery analytics and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199, 195–196.

[209] Zhang, F., Wu, L., Zhu, D., & Liu, Y. (2019). Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS journal of photogrammetry and remote sensing*, 153, 48–58.

[210] Zhang, H., Su, Y., Xu, X., & Jia, K. (2023c). Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. *arXiv preprint arXiv:2312.03502*.

[211] Zhang, J., Ma, K., Kapse, S., Saltz, J., Vakalopoulou, M., Prasanna, P., & Samaras, D. (2023d). Sam-path: A segment anything model for semantic segmentation in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 161–170).: Springer.

[212] Zhang, J., Yang, X., Jiang, R., Shao, W., & Zhang, L. (2024a). Rsam-seg: A sam-based approach with prior knowledge integration for remote sensing image semantic segmentation. *arXiv preprint arXiv:2402.19004*.

[213] Zhang, J., Zhou, Z., Mai, G., Mu, L., Hu, M., & Li, S. (2023e). Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*.

[214] Zhang, L., Ghader, S., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A., & Zhao, G. (2020). Data analytics and modeling methods for tracking and predicting origin-destination travel trends based on mobile device data. *Federal Highway Administration Exploratory Advanced Research Program*.

[215] Zhang, S. & Abdel-Aty, M. (2022). Real-time crash potential prediction on freeways using connected vehicle data. *Analytic Methods in Accident Research*, 36, 100239.

[216] Zhang, W., Liu, Y., Zheng, X., & Wang, L. (2024b). Goodsam: Bridging domain and capacity gaps via segment anything model for distortion-aware panoramic semantic segmentation. *arXiv preprint arXiv:2403.16370*.

[217] Zhang, X., Liu, Y., Lin, Y., Liao, Q., & Li, Y. (2024c). Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 (pp. 22520–22528).

[218] Zhang, Y. & Zhao, X. (2024). Mesa: Matching everything by segmenting anything. *arXiv preprint arXiv:2401.16741*.

[219] Zhang, Y., Zhou, T., Wang, S., Liang, P., Zhang, Y., & Chen, D. Z. (2023f). Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 129–139).: Springer.

[220] Zhang, Z., Cai, H., & Han, S. (2024d). Efficientvit-sam: Accelerated segment anything model without performance loss. *arXiv preprint arXiv:2402.05008*.

[221] Zheng, H., Zhang, K., & Nie, Y. M. (2021). Plunge and rebound of a taxi market through covid-19 lockdown: Lessons learned from shenzhen, china. *Transportation Research Part A: Policy and Practice*, 150, 349–366.

[222] Zheng, J. & Liu, H. X. (2017). Estimating traffic volumes for signalized intersections using connected vehicle data. *Transportation Research Part C: Emerging Technologies*, 79, 347–362.

[223] Zhong, Z., Tang, Z., He, T., Fang, H., & Yuan, C. (2024). Convolution meets lora: Parameter efficient finetuning for segment anything model. *arXiv preprint arXiv:2401.17868*.

[224] Zhong, Z., Zhao, L., Dimitrijevic, B., Besenski, D., & Reif, J. L. J. A. (2022). Framework for highway traffic profiling using connected vehicle data. In *2022 IEEE 7th International Conference on Intelligent Transportation Engineering (ICITE)* (pp. 417–423).: IEEE.

[225] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633–641).

[226] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127, 302–321.

[227] Zhou, C., Li, X., Loy, C. C., & Dai, B. (2023). Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*.

[228] Zhou, Y., Zhang, Y., Yuan, Q., Yang, C., Guo, T., & Wang, Y. (2022). The smartphone-based person travel survey system: data collection, trip extraction, and travel mode detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(12), 23399–23407.

[229] Zhuang, Y. (2022). *Efficient and Robust Machine Learning Methods for Challenging Traffic Video Sensing Applications*. PhD thesis, University of Washington.