Advancing Urban Accessibility through AI: Scalable Machine Learning Approaches to Pedestrian Path Network Mapping and Assessment

Yuxiang Zhang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Anat Caspi, Co-Chair
Linda Shapiro, Co-Chair
Joshua Smith

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Advancing Urban Accessibility through AI: Scalable Machine Learning Approaches to Pedestrian Path Network Mapping and Assessment

Yuxiang Zhang

Co-chairs of the Supervisory Committee:

Affiliate Assistant Professor Anat Caspi
Department of Electrical and Computer Engineering

Professor Linda Shapiro
Department of Electrical and Computer Engineering

Pedestrian paths are central to a healthy and accessible transportation network. A connected pedestrian path network map detailing the location, attributes, and connectivity of sidewalks, crossings, and curbs is essential to building an accessible transportation system. While automobile road networks have been extensively mapped, mapping the transportation network for the paths that serve pedestrians is inconsistent, incomplete, or missing. This fundamental lack of network information creates gaps in personal travel, accessible transportation analytics, and city planning. Typical mapping methods mainly rely on human surveyors' collections and annotations. These methods are non-standardized, laborious, costly, unscalable, and difficult to keep current. In this dissertation, we address this problem by developing scalable machine-learning approaches for the generation of pedestrian path network maps and the assessment of paths and infrastructures in the pedestrian environment.

We start by introducing a novel dataset that consists of aerial satellite imagery data, street map imagery data, and rasterized geographic information system (GIS) annotations for important classes in the pedestrian environment in multiple cities. The dataset can be used in many scene-understanding tasks for analyzing pedestrian environments. We also introduce an end-to-end artificial intelligence (AI) pipeline for inferring

connected pedestrian path networks map using existing street network information and aerial satellite images. The pipeline uses a multi-input semantic segmentation network trained on our dataset to generate predictions for important classes in the pedestrian environment and uses these predictions together with existing street information to infer a connected pedestrian path network.

Next, we develop an automated system for the mapping and assessment of sidewalk networks on mobile physical devices, commonly termed edge devices. Our system leverages advances in efficient neural networks, image sensing, Global Positioning System (GPS), and compact hardware to power sidewalk mapping on the edge. The physical system runs on a lightweight and low-power embedded device, facilitating deployment on any battery-driven mobility device, such as a powered wheelchair, stroller, or scooter for real-time on-device pedestrian paths mapping and assessment.

Lastly, to aid model selection in practical applications, we propose segmentation evaluation metrics that decompose to explainable terms and are sensitive to over- and under-segmentation errors. This new approach confers additional desirable properties, including robustness to segmentation boundaries. We contextualized the application of our metrics in current model selection problems that arise in practice when attempting to match the context of use to region-based segmentation performance in supervised datasets.

In summary, this dissertation introduces scalable approaches in response to the problem domain which requires the ability to perform fast, accurate mapping and assessments of the pedestrian paths network, through the production of open, shared data with limited reliance on human laborers.

Acknowledgements

I wish to express my appreciation to each and every individual who has accompanied me on this remarkable journey. Each of you, in your unique way, has provided invaluable support, making my endeavor not just worthwhile, but truly extraordinary.

To begin with, I would like to convey my sincerest gratitude to my advisor Dr. Anat Caspi. Her guidance, expertise, and encouragement played a significant role in shaping my academic, professional, and personal growth. I am honored by the opportunity to work under her mentorship.

I would also like to express my earnest thanks to the rest of my esteemed dissertation committee members: Dr. Linda Shapiro, Dr. Joshua Smith, and Dr. William Howe. Their insightful feedback and invaluable insights have greatly contributed to the development and refinement of this work. I am grateful for their time, effort, and dedication.

Next, I would like to acknowledge my diligent colleagues, Sachin Mehta, Nicholas Bolten, Suresh Devalapalli, and many others for their constant support and collaboration. Their insights and constructive criticism have been indispensable to my research work.

Lastly, I am mostly and perpetually grateful to my loving family for their unyielding support and unconditional love. They have been my motivation to remain steadfast during the most challenging moments of this journey. To my parents, Wenzhu Zhang and Mengju Yu, all of this would not have been possible without your support and sacrifices. To my fiancée, Zhenhui, your love and belief in me have been my pillar of support during this entire endeavor. I am fortunate and blessed to have the support and encouragement from each of you alongside me.

DEDICATION

To my family

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Challenge of Pedestrian Paths Mapping and Assessment

Pedestrian paths are at the heart of walkable and accessible urban cities, offering a mode of travel that connects nearly all other travel options and community activities. Having a comprehensive knowledge of the location, connectivity, and properties of pedestrian paths including sidewalks, crossings, and curbs is essential to any accessible transportation system. In addition, a fully connected pedestrian path network map is important to many city planning tasks and personalized wayfinding applications [1]. While automobile road networks have been extensively mapped [2, 3, 4] and even federally mandated in the U.S. [5], mapping the transportation network for the paths that serve pedestrians is inconsistent, insufficient, or even absent. This information gap creates impediments in both personal travel and accessible transportation analytics. For individuals, traveling on pedestrian paths is challenging to varying degrees for many populations because of the uncertainty in navigation due to discontinuity in sidewalks, missing crossings, and unpredictable barriers. At the city planning level, planners are often impacted by the lack of networked data so they are not able to identify bottlenecks in the pedestrian layer and prioritize resource investments in pedestrian networks. Pedestrian network data, which consists of location, attributes, and connectivity of paths, can alleviate this information gap, promote personal accessibility to public transportation, and provide crucial information to city planners during the decision-making processes. Therefore, comprehensively mapped information in the pedestrian environment is needed [6, 7].

Typical methods for collecting data for pedestrian paths mapping and assessment rely on (1) manually drawing and connecting points and lines, entering classification and tagging information for each path via open data collection tools (e.g. OpenStreetMap (OSM)) [1], and (2) field surveyors' on-site collections [8, 9, 10] in the targeted area. Such mapping tasks are laborious and infeasible if they are to be done by paratransit pathway review teams for many reasons. Firstly, the effort for data collection and analysis is considerable because of the requirement for large amounts of data. Secondly, even once the data is collected, it cannot account for, nor be easily maintained, through future changes to the environment (e.g. rerouting or reconstruction of the sidewalk). Thirdly, data collected by different organizations are often maintained in different formats. The lack of commonly used data standards creates barriers to sharing the data across organizations and reusing the data in downstream applications.

Presently, there are no scalable methods and tools for cities and organizations to collect pedestrian path data in consistent and scalable ways, nor are there data standards to enable sharing across different cities and organizations. This information gap creates barriers in mapping, assessing, and analyzing the pedestrian layer in a transportation system. Hence, affordable and scalable methods that produce consistent pedestrian path data are needed.

1.2 Contributions and Dissertation Outline

To study the under-addressed problem of mapping pedestrian path networks, this research presents scalable approaches for both large-scale pedestrian paths network map data generation and detailed pedestrian environment assessment data production.

Chapter 2 describes our literature review of the prior work. In particular, we discuss pedestrian path network map generation and pedestrian environment mapping that relates to collecting and generating data required for pedestrian paths assessment, pedestrian routing, and city planning.

Chapter 3 describes our first study that leverages and integrates different globally-available data types for proactive pedestrian path network data generation. Towards this end, we introduce a novel dataset, followed by a system to create a connected pedestrian path network map, to address the problem that a pedestrian path network map and the information needed to create such a map is missing. Specifically, we present a method for rasterizing GIS mapping annotations and create a dataset that consists of aerial satellite images, street

network image tiles, and the corresponding rasterized annotations. We also develop an end-to-end system to infer a connected pedestrian path network, detailing the locations and connectivity of sidewalks and crossings. The system uses a multi-input convolutional neural network (CNN) model trained and validated using our dataset to predict the locations of pedestrian paths and uses the predictions together with existing street network information to infer a connected pedestrian path network for previously unseen regions.

Chapter 4 presents an automated pedestrian path mapping and assessment system. To capture the additional information that cannot be easily learned from the aerial imagery data, we develop an automated mapping system that captures street-view images, segments infrastructure elements in the scene, and generates accurate sidewalk infrastructure and connectivity mapping information. The entire system is built on a low-power and light-weight edge device, which can be used as an add-on kit to any mobility device (e.g. power wheelchair, stroller, and scooter), making the system easy to deploy and scale for standardized data collection and mapping applications. With our system, the operation time for mapping pedestrian paths and infrastructures can be reduced by over 80 % compared to human surveyors' collection. Moreover, our system generates the data in a standardized way following data standardization protocols as in the Opensidewalk Schema [11], therefore, it can provide the data foundation for a variety of essential use cases and downstream activities. This data can also provide base data for first-responder curbside access information, or rich analytic tools for data-driven urban planning. Since the method allows for computation on the edge, the system uniquely enables robotic navigation tasks and other applications that have low power consumption requirements and computational resource constraints.

Chapter 5 describes our approach to explaining model performance and aiding model selection in practical applications. Since semantic segmentation models are used in both of our mapping methods with aerial-view and street-view images, to achieve better model explainability and to aid model selection in practical applications, we introduce quantitative evaluation metrics for semantic segmentation that provide granular accounting for over- and under-segmentation of semantically-similar pixel regions. We express the empirical probability that a predicted segmentation mask is over- or under-segmenting. We also penalize model segmentations that repeatedly over-segment or under-segment the same region, which causes large variations between model prediction and ground truth. We demonstrate how these issues affect currently used segmentation algorithms across a variety of segmentation datasets and models, and that currently avail-

able metrics do not differentiate between models since they do not measure these issues directly.

Chapter 6 provides a comprehensive summary of the contributions and key findings of this research, along with directions for future research in the field, underscoring the novelty and value of the methodologies and results presented in this dissertation.

This work addresses the under-studied problem of mapping pedestrian path networks by developing scalable practical methods that leverage multiple data sources to generate pedestrian path network maps, collect data in the pedestrian environment, and assess pedestrian paths and infrastructures.

Chapter 2

Background Literature

In this chapter, we discuss the related works that are in response to collecting and generating the data required for pedestrian path assessment, pedestrian routing, and city planning. Our discussions on the background literature are centered around two primary problem domains: pedestrian paths network map data generation (Section 2.1) and pedestrian environment mapping and assessment (Section 2.2).

2.1 Pedestrian Paths Network Map Generation

Pedestrian paths network map generation is the process of generating connected and comprehensive maps that details the pedestrian paths networks. The map data should include sidewalks, crossings, footpaths, and other features that are important to the pedestrian environment. This process is central to various applications including urban planning, pedestrian routing, transportation analysis, and paths accessibility and walkability assessment. We focus on four subdomains of prior work: data collection and integration, street and road network mapping, imagery datasets for built environments mapping, and pedestrian routing.

2.1.1 Data Collection and Integration

Data collection and integration is the process of gathering and integrating data from multiple sources that are required to generate pedestrian path network maps. The data sources include but are not limited to satellite imagery, street-level images, and survey data. The most commonly used tool is the OpenStreetMap (OSM) [12]. OSM is an open and collaborative platform that combines satellite imagery, street-level imagery, GPS

tracks, and user-contributed data, providing crowdsourced information that can be used for pedestrian paths annotation and pedestrian paths network analysis. The integrated data on OSM enables applications involving parcel indication and building footprint assessment [13, 14, 15, 16]. Other open-source software was also proposed. For example, Brovell et al. proposed a software architecture for collecting and integrating volunteered geographic information (VGI) for the creation of vector-based geographic datasets [17]. These prior studies highlight the importance and challenges of data collection and integration for proactive pedestrian paths network map generation.

2.1.2 Street and Road Network Mapping

The outcome of street and road network mapping is a map that can be used in automobile vehicles, but often does not contains rich information for the environments that serve pedestrians' travel. Many studies have been made on street mapping with aerial imagery data and other auxiliary data. Wu et al. [18] used the OSM center line as labeled data and extracted roads from very-high-resolution (VHR) satellite images. Sun et al. [19] added crowd-sourced GPS data to the satellite images to assist road extraction with CNN-based semantic segmentation networks. Similarly, Zhou et al. [20] fused remote sensing images and GPS data for road detection and extraction. There are other learning-based studies on road extraction recently. Lu et al. [21] proposed a novel globally aware network with multi-scale residual learning for road detection. Pan et al. [22] proposed an approach for extracting road networks from VHR remote-sensing images with a fully convolutional neural network. Mattyus et al. [2] proposed an approach that directly estimates road topology from aerial images. Mi et al. [3] proposed a hierarchical graph generation model to produce road lane graphs with LiDAR data. However, works in this domain solely focus on automobile road detection and extraction. In addition, land use and zoning mapping are often related to street mapping as it provides an analysis of the residential, commercial, and industrial zones that relate to the placement of streets and roads. For example, Ewing et al. [23] conducted a meta-analysis of the built environment and its impacts on travel behaviors. Similarly, the relations between the built environment and travel were discussed in [24, 25].

These prior studies collectively provide a foundation for mapping the outdoor streets, but mapping the environments that serve pedestrians' travel has not been widely studied. To map pedestrian paths, Li et al. [26] presented a semi-automated method that uses parcel-level data and roadway centerline data. The

method generates sidewalks that should exist in theory and human editing is required. Recently, a computer vision-based approach for generating sidewalk network datasets from satellite imagery data was proposed in [27]. The method uses satellite images to segment sidewalks, crosswalks, and footpaths in cities, then simplify the segmented polygons with the Douglas-Peucker algorithm [28] and extract the centerlines of the sidewalks with a dense Voronoi diagram [29]. This work highlights the challenges of feature detection from only satellite imagery because of the vegetation obstructions over footpaths, and the difficulties in correctly representing the path connectivity when fitting centerlines to discrete polygons.

2.1.3 Imagery Datasets for Built Environments Mapping

Currently, there are several datasets containing aerial imagery for mapping the outdoor built environments. The TorontoCity dataset [30] contains aerial satellite images for road curb extraction and road centerline estimation. PRRS [31] presents a dataset for building extraction and Digital Surface Model (DSM) estimation using satellite imagery data. In addition, the DeepGlobe dataset [32] and the ISPRS dataset [33] contain imagery and annotations for tasks including road extraction, building detection, and land cover classification. These datasets enable researchers to study different tasks involving the use of aerial imagery data (e.g. land use classification and road network extraction), but they do not directly and specifically target the pedestrian path networks, nor provide annotations needed for semantically understanding the pedestrian environment. Other imagery datasets provide labeled (e.g. KITTI [34], Cityscapes [35]) or unlabeled (e.g. StreetLearn [36]) street-view images but are not available in large quantity and do not target the pedestrian environment. To summarize, currently, there are no large-scale datasets available that specifically target the pedestrian environment, nor pedestrian path network generation methods that are applied to the large-scale area.

2.1.4 Pedestrian Routing

Pedestrian routing includes two parts, the first part is to create an informative map that accurately and effectively represents the pedestrian networks, and the second part is to develop algorithms that model pedestrian movement and generate optimal routes based on the distance of the route, the walking speed of the pedestrian, and the accessibility of the paths. As discussed in Section 2.1.1, OSM provides crowdsourced information required for pedestrian routing. Haklay et al. [37] compared OSM to professionally produced

maps and concluded that the quality of data that VGI-based maps like OSM can provide was comparable to professionally produced maps. Similarly, other studies comparing OSM to authoritative data suggested that OSM data varies across regions but in general provides a comparable representation of pedestrian networks for pedestrian routing and accessibility assessment [38, 39, 40]. A recent work, Accessmap [41] provides an open-source, interactive web map that provides individualized pedestrian infrastructure information. It also offers automatic individual routing planning services based on several customized factors including distance, slope, and access to public amenities.

2.2 Pedestrian Environment Mapping and Assessment

Besides generating pedestrian paths network maps, assessment of the pedestrian environment is also an essential component in understanding and analyzing the connectivity and accessibility of pedestrian paths. This process involves collecting data in the pedestrian environment and processing the data to analyze the elements in the built environment that can affect pedestrian travel. We focus on three areas of prior work: Outdoor built environment mapping and semantic segmentation, pedestrian paths mapping and assessment, and accessibility assessment.

2.2.1 Outdoor Built Environment Mapping and Semantic Segmentation

Automobile environment mapping has been widely studied [4]. Teichmann et al. [42] proposed a model for joint classification, detection, and semantic segmentation in automobile roads. Treml et al. [43] developed a network architecture for automobile road environment image segmentation that maintains high accuracy while being efficient. Recently, Wu et al. [44] proposed a panoptic perception model for simultaneous traffic object detection, drivable area segmentation, and lane detection.

Semantic segmentation plays an important role in these prior works for mapping the outdoor environment and many semantic segmentation architectures have been studied. Semantic segmentation architectures include convolutional networks with encoder-decoder architectures [45, 46, 47, 48, 49, 50]. In addition, region-based methods that utilize region proposals to improve the segmentation performance were also studied [51, 52, 53, 54, 55, 56, 57]. In recent years, attention-based models have been studied in many semantic segmentation works [58, 59, 60, 61, 62]. Importantly, most of the semantic segmentation networks

used in automobile mapping tasks are computationally expensive and are not suitable for mapping on edge devices. In response, efficient architectures (e.g. ENet [63], MobileNetv2 [64], ShuffleNetv2 [65], and ESPNetv2 [66]) were designed to enable computing with constraint resources.

2.2.2 Pedestrian Paths Mapping and Assessment

Although a large number of studies have been done on automobile road mapping, only limited mapping attempts have been made to the environments that serve pedestrians. Karimi et al. [67] discussed several pedestrian map generation approaches and experiments in a small-scale area to show preliminary mapping results. The result of each approach heavily depends on the availability and quality of the input data. Recently, with the advance in remote sensing technologies, more imagery-based approaches have been studied. Ahmetovic et al. [68] used satellite images to detect zebra crossing candidates and then used street-level images to validate and detect sidewalks. Similarly, Ghilardi et al. [69] used a support vector machine (SVM) classifier to detect and locate crosswalks with the aid of information from road maps. Ning et al. [70] extracted sidewalks from aerial images with a neural network and restored the occluded segments from street-level images. These previous works on extracting and locating sidewalks or crossings demonstrate the advance in pedestrian environment mapping, but they do not generate a comprehensive connected pedestrian path network required by city planning and pedestrian route-finding applications. Other than mapping with aerial images as the core inputs, there are other works that focus on using on-the-ground data. For example, Hou et al. [71] developed a system that uses LiDAR data and point cloud segmentation to extract sidewalks. But physical systems are often used in the methods that use on-the-ground data, and the physical systems need to be deployed at a larger scale to effectively generate a pedestrian path network. Other studies that map pedestrian path networks at a large scale include [26, 27] which we discussed in Section 2.1.2.

2.2.3 Accessibility Assessment

Accessibility and walkability are also important aspects when assessing pedestrian paths. Manaugh et al. [72] examined indices of walkability and developed a walkability index that incorporates many factors including the width, surface types, and locations of pedestrian pathways. In addition, Bolten et al. [1] conducted a case study on pedestrian data collection and management, and discuss its impacts on the acces-

sibility of pedestrian paths. Amenities and street furniture are also commonly related to sidewalks and the pedestrian environment in general. The impacts of pedestrian amenities and street furniture (e.g. bicycle racks, trashcans, lighting fixtures, and signs) on the walkability of urban cities were studied in [40, 73]. These studies collectively contribute to the understanding of pedestrian path accessibility and walkability in urban built environments.

Chapter 3

Generation of Pedestrian Path Networks Data at Scale from Multiple Open Data Sources

3.1 Introduction

In this chapter, we introduce a unique, annotated image dataset that captures critical elements within a pedestrian environment. Subsequently, we outline a methodology for the creation of an interconnected pedestrian path network map. This innovative dataset serves as the fundamental basis for the analysis and comprehension of pedestrian pathways using computer vision methodologies. Additionally, our proposed technique offers a scalable solution for the generation of comprehensive data on pedestrian paths.

A connected pedestrian path network detailing the location and connectivity of sidewalks, crossings, and curbs is essential to a healthy transportation network. Sidewalks and crossings connect all other transportation modes and are the key to building an accessible city. Creating a fully connected pedestrian path network is essential for many city planning tasks and wayfinding applications [1, 41]. Mapping information for the paths that serve pedestrians is incomplete or missing, as many cities do not have an accurate map of pedestrian paths. This fundamental lack of information creates bottlenecks in the transportation network and artificially disconnects neighborhoods. More importantly, this information gap impacts pedestrians

with disabilities' travel in the cities. A connected pedestrian path network map consisting of the location and connectivity of sidewalks and crossings would close this information gap and improve accessibility for all.

A pedestrian map requires different information than a map for car navigation. It must show common paths (sidewalks, crossings), their connectivity, transitions, and additional attributes of those paths. Altogether this is difficult to gather manually due to the large amount of data and potential for error [1, 74]. In addition, mappings that are done by human mappers often contain errors due to many factors, including low image quality and ambiguous mapping protocols. We validated and evaluated several areas mapped through a human annotation campaign carried out by the OpenSidewalks Project [75]. We found that although mappers were trained before doing the mapping and were provided with instructions and mapping protocol, different types of errors were still made in the mapping process. Example errors are visualized in Figure 3.1, and Table 3.1 summarizes our analysis on human-generated maps, where we analyzed the three elemental geometry features (sidewalks, crossings, and sidewalk links that connect sidewalks to crossings) [1]. These geometry-based errors can lead to inaccurate depictions of pedestrian paths and internal path connectivity, hindering downstream applications and particularly impacting routing algorithms and network analysis. Machine learning methods that globally generate connected pedestrian path networks efficiently would improve mapping accuracy, reduce manual efforts, and improve transportation planning.

In this chapter, we present a novel imagery dataset for analyzing the pedestrian environment, and a method to generate connected pedestrian path network maps at scale. The contributions of the work described in this chapter are two folds: (1) we present a novel dataset, the Annotations for Pedestrian Environment (APE) dataset (Figure 3.2), created through a rasterization process of GIS mapping annotations. The APE dataset includes aerial satellite images, street map tiles, and rasterized annotations. (2) We also develop the Pedestrian Path Network Graph Inference Method (Prophet) (Figure 3.3), an end-to-end process to infer a connected pedestrian path network, using a multi-input segmentation network trained on the APE dataset to predict pedestrian path locations and integrate them with existing street network data to complete a pedestrian path network.

[1]The nodes that are analyzed in Table 1b include the curb nodes (the intermediate nodes shared by a crossing line and a link line that connects a crossing to a sidewalk), and the crossing nodes (the nodes at the intersection of a crossing line and a road line). Sidewalks are mapped as lines in common GIS data thus there is no designated tag for sidewalk nodes in the mapped annotations.

(a) **(b)**

Figure 3.1: Example of annotations from human mappers. Red text shows different types of features in a pedestrian path network map. Yellow boxes show examples of errors made by human mapper: (a) crossing lines are not connected to the sidewalk lines with link lines (b) the curb nodes are falsely mapped to the middle of the road.

Table 3.1: Analysis on annotations from a human annotation campaign

(a) Statistics of validated annotations

	Annotation count	**Annotation length**
links	1526	9423.9 m
sidewalks	3537	37527.1 m
crossings	1594	8543.9 m

(b) Number of errors on Nodes

Node type	**Location**	**Classification**	**Tag**
curb	301	108	1
crossing	51	35	226

(c) Number of errors on LineStrings

Line type	**Location**	**Classification**	**Tag**
links	337	N/A	59
sidewalks	396	N/A	59
crossing	375	15	0

The rest of this chapter is organized as follows. Section 3.2 introduces the APE dataset. Section 3.3 details the implementation of Prophet. Experiment results are shown in Section 3.4. Discussions are made in Section 3.5.

3.2 The Annotations for Pedestrian Environment Dataset

To address the lack of large-scale datasets that specifically target the pedestrian environment, we develop a method to rasterize more commonly found GIS annotations and present the Annotations for Pedestrian Environment (APE) dataset, which contains labels for essential classes that compose a connected pedestrian path network graph. The GIS annotations for pedestrian path environments, either created by city agencies or crowdsourcing by local mappers (e.g. OpenStreetMap (OSM)), are primarily created for use in vector GIS data systems. The commonly used GIS data formats (e.g. Shapefile map or GeoJSON) cannot be directly used in computer vision tasks that required input data to be imagery-like. In creating the APE dataset, we rasterize GIS data from multiple sources in multiple geographic areas to create the labeled data needed for understanding pedestrian environments in computer vision tasks.

3.2.1 Coverage

The APE dataset covers select regions in the following areas: (1) Los Angeles, CA, United States (2) Bellevue, WA, United States (3) Quito, Ecuador (4) Sao Paulo, Brazil (5) Santiago, Chile (6) Gran Valparaiso, Chile. Consequently, the APE dataset encompasses diverse built environments from both North and South America. Cities in South America are more densely populated with higher street and intersection density, but simpler in shape compared to North American cities [76]. The OpenSidewalks Project [75] provides crowdsourced GIS annotations for select regions in each of these 6 cities in a consistent schema [77], with annotations created by trained and veteran mappers [2]. These annotations are available in standard GIS formats in OSM. In addition, the City of Los Angeles provides a GIS dataset with additional information for the Los Angeles area, including annotations for sidewalks, crossings, corner bulbs, and driveways [78]. The APE dataset consists of a total of $14,800$ samples and spans approximately $2,700 km^2$ land area.

[2] A mapper is considered trained if the mapper is given clear written instructions and had completed at least one mapping training session. A trained mapper is considered to be a veteran mapper if the mapper has completed several mapping tasks and the mapping outcomes are validated in a peer-review process.

Figure 3.2: Samples from the APE dataset. The first three rows show samples from North American cities. The last two rows show samples from South American cities, where the available aerial satellite images have lower resolutions.

3.2.2 Collecting Images and Annotations

Each dataset sample includes (1) the aerial satellite image, (2) the street map imagery tile, and (3) the rasterized annotations for pedestrian paths. The aerial satellite images and the street map imagery tiles are acquired from Bing Maps along every major road for each area described in Section 3.2.1. The GIS annotations mainly come in three geometry types: *Point*, *LineString*, and *Polygon*. The nodes representing crossings, curbs, and link endpoints are usually annotated as the *Point* geometry. For each of these annotated objects with the *Point* geometry, we convert the point to a circle by adding a buffer zone with a fixed radius and then rasterize the circle in the image. Sidewalks and crossings usually have two types of representations: the centerline representation and the polygon boundary representation. (1) If they are represented by their centerline as the *LineString* geometry, for each *LineString*, we convert it into a *Polygon* by adding a buffer zone to each side of the *LineString*, then rasterize the *Polygon* as a filled polygon shape in the image. (2) If they are already annotated by their boundary as the *Polygon* geometry, we directly rasterize them as filled polygon shapes in the image. [3] Each set of an aerial satellite image, a street map image, and a rasterized annotation are aligned with their precise geographic bounding boxes.

3.2.3 Classes

Our dataset provides annotations for three distinct classes needed to semantically segment the pedestrian environment. As shown in Figure 3.2, these classes include (1) *Corner bulb* (2) *Sidewalk* (3) *Crossing*. Corner bulbs are commonly used when describing a transportation network since they serve as a transition zone connecting a sidewalk to curb ramps, crossings, or another sidewalk. The nodes representing sidewalk endpoints, link endpoints, and curbs are usually located within the corner bulbs. Sidewalks and crossings are essential elements in an urban pedestrian path network graph, as the lines representing sidewalks and crossings are essentially the edges a pedestrian will traverse. The focus of our work is on mapping pedestrian paths, thus, all other annotated classes (including roads, buildings, and trees) collectively comprise *background class* in the following experiments. However, these classes can also be rasterized with our method (Section 3.2.2) and added to the APE dataset as additional class labels.

[3] For the data currently in the APE dataset, the polygon boundary representation is only applied to the sidewalk annotations from the City of Los Angeles. The sidewalk annotations from the OpenSidewalks project are all in the centerline representation per the OpenSidewalks standard.

3.2.4 Challenges

As previously noted in autonomous-driving datasets, imagery dataset bias can be introduced by unrepresentative geographic locations of images [79, 80, 81]. This bias generally refers to a systematic error that results from the training dataset only representing a limited geographic region or a limited type of environment. To expand APE's ability to generalize to other regions and environments, we created a more balanced and representative dataset that includes images from a diverse range of regions and environments, including imagery from both North and South American urban contexts. However, this introduces a different challenge having to do with geopolitical biases in satellite image collection: only lower-resolution aerial satellite images were openly available for South American cities. This presented challenges in model learning. Figure 3.2 displays samples from the APE dataset, with the first three rows representing North American cities and the last two rows representing lower-resolution images in South American cities. Differences in model performance when predicting pedestrian networks in different cities are discussed in Section 3.4.

Another challenge for learning from aerial satellite images is that the important classes such as sidewalks and crossings are occluded in some satellite images [70]. For example, in the second and third rows shown in Figure 3.2, part of the pixels that are labeled as *sidewalk* are occluded by buildings, the shade from buildings, and vegetation in the aerial satellite images. In these cases, learning from aerial satellite images alone is very challenging and street map imagery tiles can provide crucial auxiliary information. Section 3.4 shows a quantitative analysis of the models trained with (1) only aerial satellite images (2) only street map imagery tiles, and (3) both aerial satellite images and street map imagery tiles.

3.3 The Pedestrian Path Network Graph Inference Method

The APE dataset enables us to automate the process of generating pedestrian path network data at scale. In this section, we introduce Pedestrian Path Network Graph Inference Method (Prophet), an end-to-end process to infer a connected pedestrian path network. As shown in Figure 3.3, Prophet consists of three main steps. First, Section 3.3.1 details the creation of a hypothesized graph with existing street network data using a tool we developed called *Pedestrianfer*. Second, Section 3.3.2 describes the multi-input segmentation network trained on the APE dataset for generating pixel-wise prediction masks for the important classes

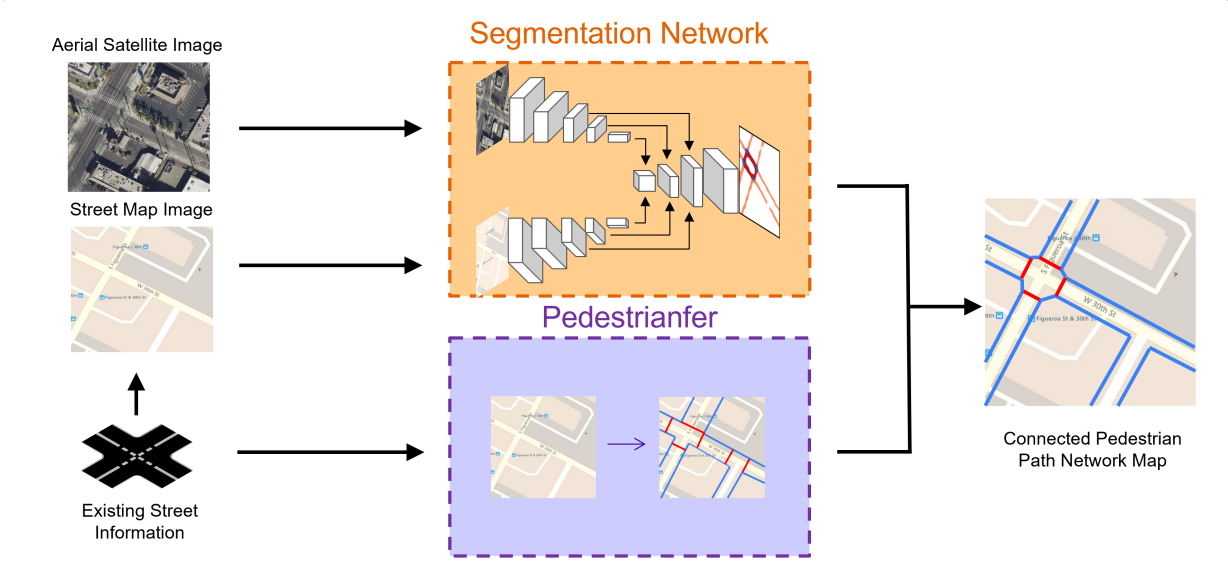


Figure 3.3: Overview of the Pedestrian path network graph inference method (Prophet), an end-to-end process for inferring pedestrian path network graph: Section 3.3.1 outlines *Pedestrianfer*, generating a pedestrian path network from incomplete information. Section 3.3.2 details the segmentation network using aerial and road map images to predict class labels of pixel locations in the pedestrian map. Section 3.3.3 explains how *Pedestrianfer* and segmentation network information are combined for accurate graph inference.

in the pedestrian environment. Lastly, Section 3.3.3 describes the process of using the predictions from the segmentation network to optimize the hypothesized graph into an accurate connected pedestrian path network graph.

3.3.1 Pedestrianfer

To infer a hypothesized pedestrian path network from incomplete information, we developed *Pedestrianfer* (Figure 3.4). *Pedestrianfer* performs three discrete inference tasks: (1) inferring an (optimistic) sidewalk network from the existing street networks, (2) inferring street crossing locations from a street network and a sidewalk network, and (3) creating a preliminary conjecture of curb interface locations around the street crossings. *Pedestrainfer* is similar in outcome to the method [26] discussed in Section 2.1.2 but with the following key differences: (1) the only required input to our method is a street network, and sidewalk locations are estimated from that street network, (2) full intersection-to-intersection sidewalk paths are estimated rather than being broken into 50-meter segments, and (3) street crossing pathways are generated via a cost function that weighs several properties and does not require manual intervention. Additionally, Pedestrianfer

Figure 3.4: Hypothesized graph generated with *Pedestrianfer*. Yellow dots represent the hypothesized nodes for the curb modes and the nodes on the sidewalk. Blue lines represent the hypothesized sidewalks. Red lines represent the hypothesized crossings. Links are part of the crossings between the curb nodes and the nodes on the sidewalk.

generates pathways according to the OpenSidewalks Schema specification [77].

Pedestrianfer first infers sidewalk networks from a street network under two alternative regimes, each representing a different hypothesis on the built environment. The alternatives depend on whether the data source is a vector layer of street network data with metadata on the presence (and optionally, offset distance) of sidewalks. If metadata is present, sidewalks are placed only where indicated. If not, *Pedestrianfer* retrieves a street network and hypothesizes full sidewalks, i.e. the scenario in which all streets have a connected sidewalk on either side. In the latter case, *Pedestrianfer* retrieves an existing street network, such as a street network from OSM. In either case, the *Pedestrianfer* process involves 4 steps: (1) creating a directed graph representation of the street network, (2) generating all right-hand-turn paths that start and stop at the same node (closed paths), (3) drawing sidewalks via line offset algorithm, and (4) trimming or joining them based on the path context, e.g. trimming overlapping sidewalk lines when they are neighbors along the path.

Next, *Pedestrianfer* uses a street network and a sidewalk network (generated by *Pedestrianfer* in the

previous step if it does not already exist) to infer crossings. In this step, *Pedestrianfer* iterates over each street intersection node in the street network and generates (1) all street lines associated with the intersection, directed outwards from the intersection and up to half of the distance to the next intersection, and (2) a set of candidate sidewalks to connect with a crossing on each side of each street. *Pedestrianfer* then generates candidate crossings that are lines drawn from a sidewalk on the left side of a street to a sidewalk on the right side of a street. Multiple candidates are generated for each street of an intersection by selecting a series of points along that street (every 1 meter) and generating a crossing that connects the closest corresponding left and right sidewalks. Metrics of the candidates are generated, including the distance of the street point to the intersection, the crossing line length, and the angle between the crossing line and the street it crosses. A cost function is then used to heuristically select the *best* crossing: one that minimizes a linear combination of the distance to the street intersection, crossing line length, and non-orthogonal crossing angles. Therefore, *Pedestrianfer* estimates street crossing locations that are near intersections, which align with common (albeit not universal) policies and pedestrian safety measures. In the case that crossing locations with ground markings are already known and present in an associated dataset (as *Point* data), *Pedestrianfer* will generate a crossing near that point by projecting to the nearest known sidewalk candidates on each side.

Lastly, *Pedestrianfer* splits crossing lines into three segments: (1) the originating sidewalk surface, (2) the street surface, and (3) the destination sidewalk surface. These correspond to typical surfaces a pedestrian would travel in an urban area. Sidewalk-street transitions, which often have vertical displacements or are where curb-cut ramps meet the street, can be interpreted as potential curb interfaces. *Pedestrianfer* does not use any metadata to determine where to split each crossing into these 3 segments, but rather provides a low-information hypothesis on sidewalk-curb-street locations, which can be improved with manual mapping (GIS software, crowdsourcing in OSM) or learning-based methods such as those described in Section 3.3.2 and Section 3.3.3.

3.3.2 Segmentation Network

In order to verify, correct, and refine the *Pedestrianfer* hypothesized path network graph, we use inference from a CNN-based segmentation network. The segmentation network has a siamese-like structure, allowing it to fuse and utilize the information from both the aerial satellite image and the street map tiles. As shown

in Figure 3.3, the segmentation network has two identical branches, the aerial satellite images are used as the input in one branch and the street map tiles are used in the other. Each branch has its encoder-decoder structure. In addition, to fuse the information from both branches, we concatenate feature maps from both branches at different layers hierarchically during the up-sample process. Experiments described in Section 3.4 used FCN-8s [46] as the backbone model, but it can be replaced by any other segmentation model with a similar structure.

3.3.3 Optimization for Full Pedestrian Path Network Graph

Optimizing node geolocation

Pedestrianfer generates a hypothesized path network graph that outlines the potential location and connectivity of sidewalks and crossings. To obtain a more accurate pedestrian path network, we use information from the segmentation network to refine the hypothesized path network graph. Our strategy is to first find the optimized geolocation of the nodes in the graph, then connect the nodes with edges to complete the graph.

At each intersection, there are nodes representing sidewalks endpoints and curbs in a hypothesized graph generated by *Pedestrianfer*. Ideally in a correct pedestrian path network, these nodes should all locate in a connected region being segmented as the *corner bulb* class (an example is given in Figure 3.5 (c) and Figure 3.5 (d)). To infer the correct geolocation of these nodes, we aim to find a parameterized affine transformation to warp each set of hypothesized nodes in a corner, represented by their pixel coordinates, to a new set of coordinates, so they better align with the *corner bulb* class in the predicted segmentation mask. We start by connecting the nodes at each corner (shown in Figure 3.5a) to form a closed polygon (shown in Figure 3.5b), then we find a parameterized affine transformation so that the sum of the probability of each pixel under each closed polygon being the *corner bulb* class is the greatest. Mathematically, this optimization process can be defined as follows. For the total of n points:

$[(x_1, y_1), (x_2, y_2), ...(x_n, y_n)]$, in a corner in a given Image I, the affine transformation that warps them to a set of new points: $[(x'_1, y'_1), (x'_2, y'_2), ...(x'_n, y'_n)]$ can be described as:

$$\begin{bmatrix} x'_1 & ... & x'_n \\ y'_1 & ... & y'_n \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} x_1 & ... & x_n \\ y_1 & ... & y_n \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = AX + t \tag{3.1}$$

In words, X is a set of points representing the coordinates of the nodes in a hypothesized graph at a given street corner. The new set of coordinates $[(x'_1, y'_1), (x'_2, y'_2), ...(x'_n, y'_n)]$, and the transformation parameters A and t are to be found using the information obtained from the segmentation network.

Assuming there are a total of m pixels that fall into the polygon enclosed by the n corner points X, and the probability (as predicted by the segmentation network) that each of these m image pixels to be the *corner bulb* class is denoted as p_i, then we define the function that finds these pixels and their probability as f:

$$f : X \longmapsto [p_1, p_2, ...p_m] \tag{3.2}$$

and define a function g that sums $[p_1, p_2, ...p_m]$ as:

$$g(f(AX + t)) = g([p_1, p_2, ...p_m]) = \sum_{i=1}^{m} p_i \tag{3.3}$$

The goal is to maximize g so that the pixels under the new polygons have the greatest sum of probability (shown in Figure 3.5c). Thus, the optimization problem can be expressed as:

$$\begin{aligned} &\underset{A,t}{\text{minimize}} \quad -g\left(f\left(AX + t\right)\right) \\ &\text{subject to} \quad \forall i \in [1, n] \quad 0 < x'_i < I_{width}, 0 < y'_i < I_{height} \end{aligned} \tag{3.4}$$

There is no closed-form solution to Equation 3.4, and the gradient of f cannot be explicitly found. Hence we use the simultaneous perturbation stochastic approximation (SPSA) [82] method to find the optimal parameters A and t of the objective function g.

After optimizing sidewalk and curb node geolocation, we connect them with the information from the hypothesized graph to generate new sidewalks and crossing edges. The optimization process is shown in Figure 3.5. Figures 3.5a and 3.5b show the original node locations in the *Pedestrianfer* hypothesized graph.

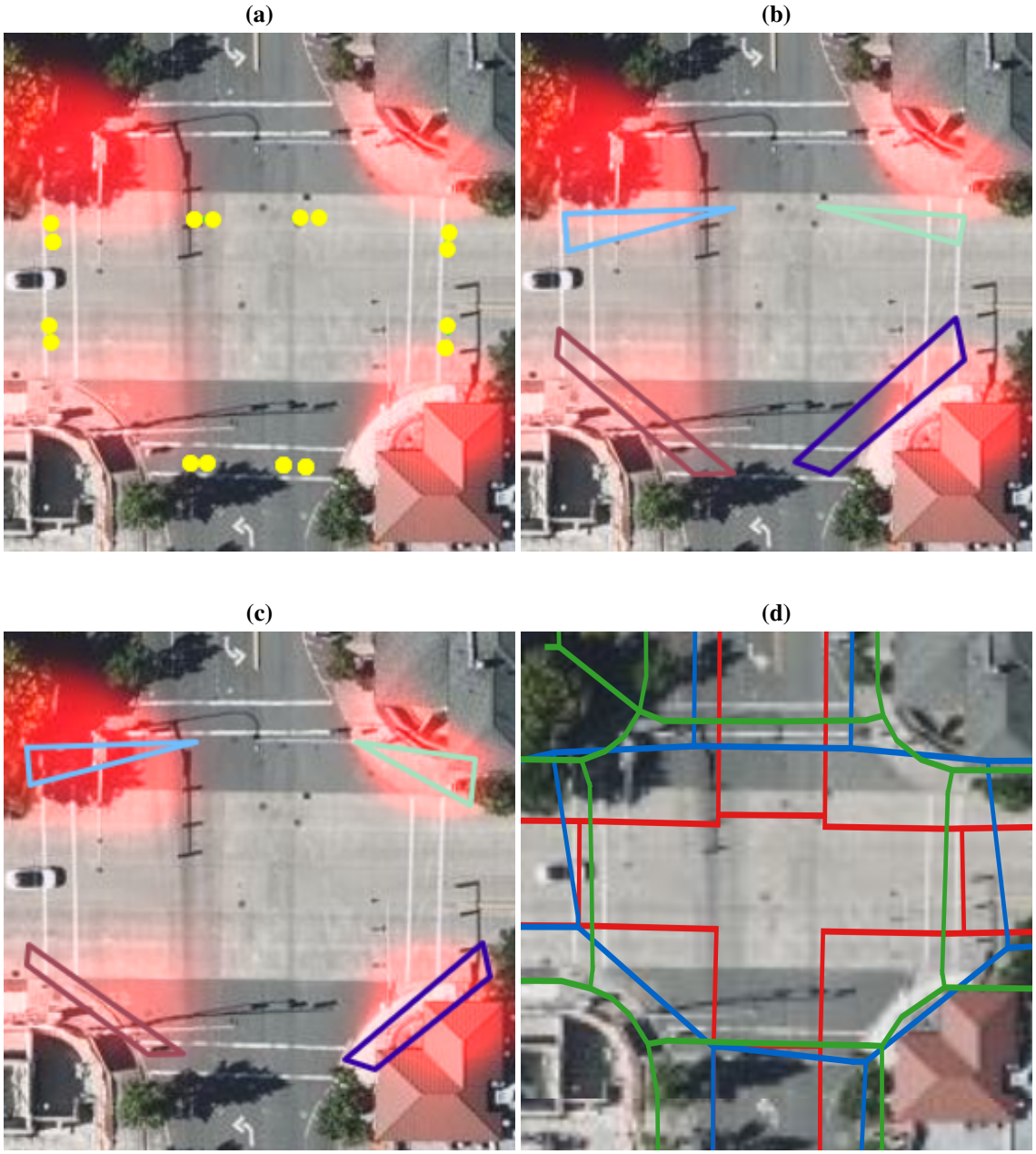


Figure 3.5: Illustration of node location optimization, the probability of each pixel being the *corner bulb* is shown as a heat map in red in (a) - (c). (a) *Pedestrianfer* hypothesized nodes (b) Polygons formed by the hypothesized nodes in each corner (c) New set of nodes optimized with information from the segmentation network (d) Compare to Human annotation: Green is human annotation graph Red is Pedestrianfer Hypothesized graph. Blue is the optimized graph.

In this example, *Pedestrianfer* approximates the nodes erroneously in the middle of the road. Figure 3.5c shows the locations of the nodes post optimization, where the nodes and the enclosed polygon formed by the nodes are moved to the middle of a region predicted as the *corner bulb* class. The improved graph (post-optimization) in Figure 3.5d, shows graph elements considerably closer to the human-generated (ground truth) graph. A more detailed quantitative evaluation is made in Section 3.4.4.

Graph refinement with class probability

The optimized graph is used with probability masks of each class from the segmentation network to further improve the predicted path network map accuracy. For a set of hypothesized nodes at a given corner, if the closed polygon they form does not overlap with enough high-probability pixels representing the *corner bulb class*, we consider these nodes to be falsely hypothesized. Mathematically, we define μ_p as:

$$\mu_p = g(p)/m \tag{3.5}$$

If μ_p is less than a set threshold, these nodes are considered to be false-hypothesized and therefore directly removed from the graph without optimizing with equation 3.4. The threshold is chosen to balance the edges' precision and recall in this post-processing step. Higher thresholds lead to higher precision.

For downstream applications (e.g. wayfinding) that use the predicted graph, confidence values are added to edges to inform the optimization for high-confidence routes, i.e., the application may choose to avoid low-confidence edges and instead optimize for higher-confidence paths. For each edge created by connecting two nodes in the graph, we assign a confidence value as an attribute of the edge as follows: each *LineString* representing the *crossing* class, is converted into a polygon by adding a buffer to each side of the *LineString*, then we compute the mean probability of the pixels in the polygon in the *crossing* class, similar to Equation 3.5. The mean probability is used as the confidence value and stored as an attribute of the edge in the graph. Similarly, the confidence values of the *sidewalk* edges are computed and stored as an attribute of the *sidewalk* edges. The confidence value improves the graph for various transportation network analyses, sidewalk network scoring, personalized routing [41], and other downstream applications.

3.4 Experiments

Figure 3.6: Qualitative results on the validation set. The segmentation results of 3 different models are shown in columns 4-6. (1) Trained with the aerial satellite image branch only (2) Trained with the street map tile branch only (3) Trained with both the aerial satellite image branch and the street map image tile branch. The model that uses both aerial satellite images and street map images generates better predictions than the models that use only one input. A detailed discussion of these samples is made in Section 3.4.2. Quantitative analysis of models' performance is made in Section 3.4.3 and Table 3.2.

3.4.1 Training the Segmentation Network

Three segmentation networks were trained with different setups: aerial satellite image branch only, street map image tile branch only, and both aerial and street map image tile branches. All three models used the

Table 3.2: Quantitative segmentation results: the model that uses both aerial satellite images and street map tiles outperforms models that use only one branch of data

Method	Pixel-wise						Instance-wise (Corner)	
	BG IoU	SW IoU	Corner IoU	Crossing IoU	mIoU	Accuracy	Precision	Recall
Satellite Only	0.89	0.42	0.47	0.42	0.55	0.89	0.59	0.68
Street Only	0.85	0.33	0.48	0.37	0.51	0.86	0.60	0.63
Satellite + Street	0.91	0.47	0.57	0.49	**0.61**	0.92	0.73	0.69

Table 3.3: Quantitative evaluation: graph level analysis

(a) Sidewalks

	Precision			Recall			F1		
	LA	**Bellevue**	**Quito**	**LA**	**Bellevue**	**Quito**	**LA**	**Bellevue**	**Quito**
Pedestrianfer (Baseline)	68.1	69.2	67.1	73.4	73.2	71.2	70.7	71.1	69.1
Pedestrianfer + Segmentation	83.5	82.4	75.6	77.8	77.0	73.6	80.5	79.6	74.6

(b) Crossings

	Precision			Recall			F1		
	LA	**Bellevue**	**Quito**	**LA**	**Bellevue**	**Quito**	**LA**	**Bellevue**	**Quito**
Pedestrianfer (Baseline)	73.5	69.2	71.9	82.1	71.8	75.8	77.6	70.5	73.8
Pedestrianfer + Segmentation	76.7	72.2	74.4	87.7	75.1	78.7	81.8	73.6	76.5

same dataset split and data augmentation techniques. Performance comparisons are shown in Figure 3.6 and Table 3.2.

Dataset

We train and validate with an $80\%/20\%$ split of the APE imagery and annotation samples in the Los Angeles area. Specifically, we chose to use the rasterized data from the City of Los Angeles annotations because of the amount of data available. Bellevue and Quito annotations from OpenSidewalks are used as unseen test sets.

Data Augmentation

To improve the robustness of the segmentation network, data augmentations were applied to the training data during the training stage. First, random rotating, cropping, and resizing were applied. Again, we note the resolution differential in open aerial satellite images in North America (Los Angeles and Bellevue) as compared to those from South America (Examples are shown in Figure 3.2). We enhanced the segmentation

network's performance on low-resolution images by using Gaussian kernels with random sizes to artificially generate low-resolution images in training.

3.4.2 Qualitative Analysis

Figure 3.6 visualizes the segmentation results on the validation set. The models' shape segmentation and identified locations for *sidewalk*, *crossing*, and *corner bulb* align well with the ground-truth segmentation. These qualitative examples show the difficulty in predicting pedestrian path network classes with single-source input and the improvement gained from adding other input sources. The first and second rows of Figure 3.6 show that different kinds of errors are produced when the model is trained with only one branch of input data. In the example shown in the first row, the prediction made with the model trained with only street images generated false-positive *sidewalk* predictions. Adding aerial satellite images in model training helped remove these spurious predictions. Predictions made with only aerial satellite images incurred other limitations, as shown in the second-row example, when sidewalks were occluded in the satellite image by vegetation, the model trained with only those images generated false-negative *sidewalk* predictions. Adding street image tiles in training helped predict many of the occluded sidewalks. To further demonstrate the effectiveness of using both the aerial satellite images and the street image tiles as inputs to the model, the examples in row 3 and row 4 of Figure 3.6 show different cases in which using both branches of data generated better results than using one branch of input data alone. In the third row example, the model trained with only one branch of data generated different false-positive *sidewalk* predictions, and using both branches of data to train the model helped remove these false-positive predictions. Similarly, in the fourth-row example, models trained with only one branch of data generated false-positive predictions on *sidewalk* and *crossing*, while using both branches of data helped remove these false predictions. A quantitative analysis of the performance of our method when testing in different cities is provided in Section 3.4.4.

3.4.3 Quantitative Analysis

Table 3.2 shows the quantitative experiment results for the models trained with the 3 different setups: (1) trained with the aerial satellite image branch only, (2) trained with the street map image tile branch only, and (3) trained with both the aerial satellite images and street map image tiles. The metrics we included are

(1) overall mIoU, (2) the mIoU for each of the 4 classes (*background* (BG), *sidewalk* (SW), *corner bulb*, and, *crossing*) in the dataset, (3) pixel accuracy, and (4) instance-wise precision and recall for the important *corner bulb* class.

From Table 3.2 we observe that the model trained with both the aerial satellite images and street map image tiles has a higher mIoU on each of the classes than the models trained with only one branch of data, demonstrating its better pixel-wise performance. In addition, higher precision and recall on the important *corner bulb* class demonstrates its better instance-wise performance. These metrics combined shows the effectiveness of the model for segmenting the pedestrian environment with both the aerial satellite images and the street map image tiles, confirming the discussions we made in Section 3.4.2 based on the observations from Figure 3.6. A graph-level analysis is made in Section 3.4.4.

3.4.4 Quantitative Evaluation of the Inferred Path Network Graph

Although the models perform well with the pixel-wise and instance-wise measures, these metrics do not fully reflect the accuracy of the connected graph prediction. mIoU (or other pixel-wise measures) cannot measure how close a predicted graph is to the ground truth graph. In order to measure the similarity between the predicted graphs to the ground truth graph, we compare (1) the pedestrian path network graph generated by *Pedestrianfer*, and (2) the graph optimized using the segmentation network, to the ground truth graph generated and validated by human mappers from the OpenSidewalks project. *Pedestrianfer* is used as the baseline method as it is similar to the method proposed by [26] and it represents the methods that derive a pedestrian path network purely from existing street network information. The metrics we used for evaluating the graphs are similar to the ones proposed by [2] for road topology measurements, namely the precision, recall, and F1 score based on the assignments of predicted sidewalk edges (or crossing edges) to the corresponding edges in the ground truth graph. The results are summarized in Table 3.3. The first row of each sub-table shows the performance of the baseline model, and the second row of each sub-table shows the performance of our method where the information from the segmentation networks that use both aerial satellite images and street images is used to optimize the predicted graphs. Table 3.3a shows evaluation results on sidewalks and Table 3.3b shows evaluation results on crossings. The graph generated with the segmentation network outperforms the baseline model, especially in the precision category because the

baseline model that uses only street network information tends to be overly optimistic about the existence of pedestrian paths. High precision, recall, and F1 score demonstrate the effectiveness of our method in inferring a pedestrian path network with both the existing street network and the information learned with the segmentation model. In addition, our method maintains high precision and recall when tested on unseen datasets (Bellevue and Quito) and outperforms the baseline methods, despite the test areas having significantly different built environments and aerial satellite image resolution (with notably low-resolution imagery in Quito). Most importantly, since we start from a connected graph hypothesized by *Pedestrianfer* and utilize the segmentation network alone to correct node locations and remove false-hypothesized geometries, the outcome graph stays as a connected graph that can be directly used as a routable network graph.

3.5 Discussion

In this chapter, we introduce the APE dataset to address the dearth of data and methods to automate mapping for pedestrian path networks. APE includes aerial images, street map tiles, and annotated pedestrian environments. It provides the research community with a new and challenging dataset to address pedestrian network path predictions through machine learning methods. The dataset covers diverse urban areas and we demonstrated it can be used for computer vision tasks. Additionally, we develop Prophet to infer pedestrian path networks using segmentation and street data, and validated the method's accuracy through comparison to human-annotated data. Through our experiments, we were able to both metricize human performance in pedestrian networks annotation tasks like drawing geometries and labeling (metrics achieved through double annotations) and also evaluated the baseline performance of a commonly used segmentation model. The APE dataset and Prophet provide valuable initial contributions to pedestrian environment research, specifically in contributing to many wayfinding and planning tasks that rely on detailed pedestrian transportation networks.

Chapter 4

Detailed Assessment of Urban Pedestrian Paths on Portable Edge Devices

4.1 Introduction

To capture additional detailed information that cannot be easily learned from the aerial imagery data, in this chapter, we introduce an automated mapping and assessment system that captures street-level images, segments the infrastructures in the scene, and generates accurate sidewalk infrastructure and connectivity mapping information.

Sidewalks offer a mode of travel that connects nearly all other travel options and community activities in urban cities. However, sidewalks are not safe and accessible to all. Many travelers, particularly those with disabilities, encounter sidewalk obstacles, like steep inclines, damaged concrete, or curb barriers, that limit their access to the benefits of urban life. Detailed networked information about the paths, connectivity of paths, avoidable obstacles, and other amenities along paths, provide pedestrians with different mobility concerns the ability to discover routes appropriate for them, plan trips, and foreshadow what they may find on the ground [83, 84, 85]. Additionally, in the United States, under Section 504 of the Rehabilitation Act (1973) and Title II of the Americans with Disabilities Act (ADA) (1990) [86], any public entity subject to ADA federal regulation must perform self-evaluation of their public right of way (PROW) to identify barriers in the paths and specify a remediation plan. While driver-routing, asset management, and

transportation planning have motivated extensive mapping, surveying, and monitoring of automobile roads (e.g., [3, 35, 87, 88]), pedestrian street-side environments have not been regularly surveyed or monitored [6]. Studies in the U.S. focusing on the PROW, which consists of civic environments and paths for public use (including sidewalks and footpaths), have demonstrated that government agencies often lack knowledge about their own PROW and the barriers within those pathways [89, 90]. Importantly, both public and private stakeholders need and would benefit from routable graphs describing sidewalk networks. Public stakeholders require data collection and monitoring of sidewalk environments including first responders in routing to emergent settings, road construction coordinators, Safe Route to Schools programs, Vision Zero programs, as well as paratransit teams to manage door-to-door transportation and determine rider eligibility. Private uses involve shipping and freight applications identifying curbside-to-door routes, and street-side autonomous navigating devices, e.g. delivery robots and autonomous wheelchairs. There is evidence to support the idea that the information gap between the cartographic information needed for pedestrians with disabilities to navigate in the sidewalks and the existing mapped information severely deteriorates the performance of transportation networks and impacts the security and safety of travelers [91, 92, 93]. Our goal is to introduce a system for automated and ongoing monitoring of pedestrian path networks and an analytic software solution that can leverage streetside footage to obtain useful, open, shared pedestrian network data that is geographically accurate, consistent, and usable by both public and private stakeholders. The desired outcome is open, shared, interpretable data that enhances available tools, visualizations, and metrics to manage transportation networks, accessibility, and traversability at scale [74].

In this chapter, we present an open-source system tested through a partnership with King County Metro Access, which serves as an applied case study environment. Our pilot was tested in three different municipal environments to demonstrate the feasibility of system adoption in real-world environments, and in the hands of real-world agencies.

The system contribution, Open Automated Sidewalks Inspection System (OASIS), is an analytic system that provides inferred sidewalk network output via a human-guided device composed of direct-to-consumer off-the-shelf electronics. It consists of a number of steps: OASIS captures street-view images, segments infrastructure and street furniture in the scene, and generates mapping information for sidewalk assessment in real-time – potentially avoiding image data storage, per user preference. OASIS is built on a low-power

and lightweight edge device to enable easy integration into other powered mobility devices. The pedestrian paths assessment process, which is typically done by human surveyors, is automated and confers an efficient and scalable way to assist with sidewalk mapping and review. With OASIS, we demonstrate the ability to generate pathway assessment data for multiple municipalities, in a standardized format that can be reused by other stakeholders or consumed by downstream applications.

The rest of this chapter is as follows. Section 4.2 details the real-world setting and organizational partner for our pilot. Section 4.3 details the implementation and methods utilized in building our system. Section 4.4 describes the hardware used in the preliminary proof of concept implementation. Quantitative and qualitative results from the pilot study are provided in Section 4.5, followed by discussion and conclusions in Section 4.6.

4.2 Case Study Stakeholders

Transportation agencies are federally mandated to provide ADA-complementary paratransit service. These are transportation services provided to individuals with disabilities to supplement fixed-route services that are inaccessible to them. King County Metro (Washington State) provides such paratransit services through its Metro Access program. Metro Access has a direct interest in assessing pedestrian mobility along the PROW: (1) like other regional transit providers, they have a legal obligation to provide origin-to-destination paratransit rides to eligible individuals that cannot reasonably reach other means of transportation via the PROW, (2) they must determine eligibility for use of their services via local PROW surveys on a per-individual, and sometimes per-transit-request, basis, and (3) if possible, clients are preferentially directed to fixed transit through routes that are feasible and practical to the rider but may not necessarily be the most direct paths between origin and fixed-route stop. Since 2006 Metro Access staff, a team of pathway reviewers, has been slowly compiling a detailed database of pathways and barriers (including uneven terrain, elevation change greater than eight degrees, unmarked intersections, and improper/missing curb ramps). Over time, Metro Access has grown the effort to a dedicated team that collects the data in person for specific paths for frequent trips made by riders. The manual data collection consists of 1019 kilometers of sidewalks and crossings in King County. The team visits the origins and destinations of the rides in person and reviews the paths to bus stops for both legs of the trip. Included in the review are taking measurements and

photographs of the sites. Eligibility reviews benefit both agencies and riders. Redirecting riders to the fixed route system frees up funds that can be put into other services (including back into the paratransit system) and can provide a better experience to the individuals who request rides: the paratransit system must be booked days in advance and has large windows for pickup and drop-off, whereas fixed transit can often be independently navigated and may have more convenient time intervals.

Until 2019, pathway reports collected by the pathway review team were not suitable to create pedestrian networks necessary for long-term needs assessment: reports were missing the vast majority of relevant public pedestrian pathways, potential obstacles and obstructions were stored as points separate from pathways (and therefore could not be automatically assessed via software without significant transformation and guesswork). Additionally, it was difficult to assess data for "staleness", and it was collected in a purposefully incomplete and subjective manner: a pathway was reviewed primarily for the individual requesting paratransit services, so any information deemed irrelevant to that eligibility request was not collected. Much of the pathway review work was redundant in that the team collected and stored point of interest (POI) information readily available on other GIS services (OpenStreetMap, Google, etc). Finally, personally identifiable information (PII) could have been inferred from the data, presenting a privacy liability as well as a challenge to leverage sharing of data.

Metro Access is yet another user story demonstrating that comprehensively mapped information in the pedestrian environment is needed, and resonates concerns raised by other studies [6, 7, 94]. The challenges faced by the Metro Access team were also prevalent with other stakeholders as described by [1]. In general, surveying mapping tasks are laborious and infeasible if they are regularly collected manually. First, data collection and analysis are considerable because of the requirement for large amounts of data. Second, even once the data is collected, it cannot account for, nor be easily maintained through future changes to the environment (e.g. rerouting or reconstruction of the sidewalk). Presently, there are no devices for cities to collect this data in consistent and scalable ways, nor are open data standards being used to enable sharing with or across cities and organizations.

In recent years, Metro Access with partners, has been using an open data specification named OpenSidewalks [75] to perform collections of GIS layers to store the results of these local surveys for future use: if there is already a known path or obstruction, a remote review can be performed to establish eligibility and

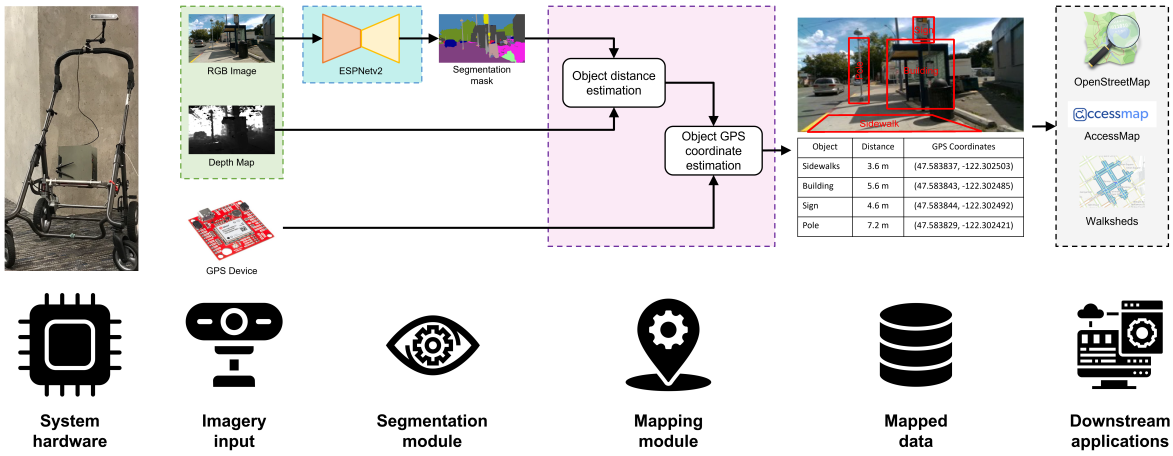


Figure 4.1: Overview of OASIS: Geographic location of the objects in the scene and the connectivity of pedestrian path networks are inferred on-device with data from a stereo camera and a GPS device.

any survey can focus on specific changes to infrastructure since the last review, rather than making redundant observations [95]. However, the general challenges of manual review remain. As pilot partners, we intended not only to address the technical challenges of automating the laborious parts of collecting OpenSidewalks transportation layer, and associated measurements, but also to improve the possibility of technology adoption through addressing organizational concerns around protecting personally identifiable information, as well as avoiding the costs and security concerns around large imagery storage and archival.

4.3 OASIS: An Open Automated Sidewalks Inspection System

In this section, we describe the Open Automated Sidewalks Inspection System (OASIS) for mapping sidewalks on portable devices commonly termed edge devices. OASIS, as shown in Figure 4.1, consists of two main modules: (1) segmentation module (Section 4.3.1) and (2) mapping module (Section 4.3.2). The segmentation module classifies each pixel in an input image into semantic categories of interest. The mapping module takes the semantic segmentation mask, the depth map, and the GPS coordinates of the camera as an input and produces the GPS coordinates of each object present in the environment as an output. This not only helps us to understand the environment, but also allows us to register objects seen by the system to mapping libraries, such as OpenStreetMap (OSM), so that such information can be used in downstream applications. Figure 4.1 shows the overall system and each component of OASIS is described in the following

sub-sections.

4.3.1 Segmentation Module

Semantic segmentation partitions a scene into several meaningful parts, decomposing it into its pixels and having each pixel labeled as belonging to a certain class of object. OASIS is built around the object classes that are meaningful in the context of outdoor pedestrian environment mapping. These classes are paths (road, sidewalk), and impassable obstacles which may be static (building, wall, fence, pole, traffic light, traffic sign, vegetation) or moving (person, rider, car, truck, bus, train, motorcycle, bicycle). Additionally, we segment sky and terrain to avoid false-positive object detection of street-side items in these areas. An integer ID is assigned to each pixel, representing its inclusion in the corresponding object class. The segmentation process uses an Image $I \in Z^{3\times w\times h}$ as input, after the segmentation module assigns one of these integer IDs to each pixel in Image I, a new image $S \in Z^{w\times h}$ called the image segmentation mask is produced. Note that although OASIS is capable of segmenting objects among all the classes mentioned, the classes we choose for further analysis in this prototype system are specific to those static objects in the pedestrian environments: pole, traffic light, traffic sign, and building, which are commonly presented in the street environment and typically hard or too frequent to map each instance by hand. As discussed in Section 2.2.1, the state-of-the-art segmentation models are the Convolutional Neural Networks (CNN). We choose ESPNetv2 [66] as the base segmentation model in our prototype system. ESPNetv2 greatly reduced the computational power needed while maintaining comparable results to other state-of-the-art models, which is important to deploy the model on an edge device. Each RGB image I captured by the stereo cameras is used as input to ESPNetv2, generating the resulting segmentation mask S. Each pixel in the segmentation mask has an integer value representing the class that has the highest probability among all classes as predicted by the model.

Though the CNN models have high reported segmentation accuracy in outdoor environments, the models for the purpose of outdoor mapping have generally been trained on imagery acquired from a vehicle dash-board, e.g. Cityscapes [35] and KITTI [34]. The errors introduced by this dataset bias are non-negligible in our domain setting of pedestrian environment mapping. We use two approaches to resolve this problem. Firstly, our application aims to perform semantic segmentation while the camera is traveling on the

sidewalks. This implies that the images captured and fed to OASIS are consecutively in time. Instead of separately using each frame in the image stream for inferring, we propose that utilizing the temporal information in consecutive frames helps improve inference results. We compute the union of masks from CNN across a number of temporally adjacent frames. This straightforward approach is shown to reduce over-segmentation errors and increase overall prediction accuracy. Secondly, observing that sidewalk is the class that suffers the most from the change of viewpoint between training samples and testing samples, we retrain the model on the more diverse dataset coco-stuff [96], and fine-tune on a dataset containing our hand-labeled images captured in diverse sidewalk environments. Because a model trained with images captured from the car dashboard has an inherent bias that the sidewalk has a higher probability to be on the sides of the input image, a model trained on an automobile dataset often incorrectly predicts the majority of the sidewalk as an automobile road (see Figure 4.4 for an Example). This error is not acceptable in applications that require knowledge of the location and connectivity of sidewalks. After retraining on our pedestrian-centric dataset, the model prediction accuracy is greatly improved in real sidewalk environments. The evaluation of our segmentation module is discussed in Section 4.5.2.

4.3.2 Mapping Module

In this section, we detail the design of our mapping module. With the output of the segmentation module, the mapping module tracks and registers each static object in the scene, and infers sidewalk walkability and connectivity by inferring the location and width of each sidewalk fragment.

Tracking and registering static objects

After the segmentation mask S of each image I is generated, it is combined with its corresponding depth map to compute the distance from the objects to the system. A depth map is an image (array) in which each pixel contains information about the depth from each point in the 3D scene at that pixel to the camera, from the same viewpoint that the corresponding RGB image is captured, see inputs in Figure 4.1 for an example of an RGB image and its corresponding depth map.

Using the segmentation mask, we found the pixels representing each segmented object in the scene. The arithmetic mean of the values of these pixels in the depth map is computed to represent the depth of

the segmented object. After the distance information of each segmented object is obtained, this distance is combined with the system's GPS information to estimate the GPS location of each object in the current scene. The system extracts the latitude and longitude at the time that an image is taken, and then infers the relative direction of each object using its pixel coordinates and the heading of the camera. Using the GPS location of the starting point (system location), camera heading, and distance inferred, the GPS of each object seen by the camera can be estimated with the Haversine formula.

Since our goal is to infer the geographic information for the sidewalk environment, if the same static object appears in different frames, the system needs to be able to recognize it. In order to keep track of each unique object in consecutive frames, we implemented an object tracker using the centroid tracking method. Using the segmentation mask, the centroid (the geometric center) of each segmented object is computed. When a new frame comes from the camera, the object tracker takes in a set of new centroids coordinates and polygons of the segmented objects, which are produced by the CNN model for the current frame. The object tracker tries to match new objects with the existing objects from previous frames. This is done by computing the Euclidean distance between each pair of new object centroids and the existing object centroids. Each new object will be associated with an existing object where the Euclidean distance is the smallest. If the number of existing objects exceeds the number of new objects, the existing objects that cannot be paired with a new object are designated as temporarily lost. If these objects remain in the temporarily lost state for a duration exceeding a predefined threshold (5 frames in our prototype system) of frames, they are subsequently marked as permanently lost. If the number of existing objects is less than the number of new objects, each new object that cannot be associated with an existing object will be assigned a new object ID. Since we are tracking objects while the camera is moving, the pixel location of object centroids will change because of the movement of the scene even if the object itself is static. To incorporate this information, we compute the homography transformation between adjacent frames, which describes their spatial relationship in terms of rotation and translation. Then we use this transformation to warp the centroid of each object from the current frame to the next frame. This step brings each existing object centroid closer to the new object centroid and improves the object tracking process. Section 4.5.2 evaluates the mapping results for static objects.

Mapping sidewalks

Unlike the registration method we applied to the static objects, we use a different approach to map the sidewalk. Unlike other object classes, for a sidewalk, we not only want to know its geolocation, but the width of a sidewalk is also an important characteristic for measuring pedestrian path walkability. We did not attempt to track the sidewalks in sequential frames, instead, we directly compute the GPS coordinates of each sidewalk fragment with the Haversine formula. The width of each sidewalk fragment is then estimated with camera parameters through perspective projection.

As the geolocation and width of each sidewalk fragment being inferred, more importantly, the connectivity and walkability of a pedestrian path network are inferred. For inferring path connectivity, after the sidewalks for a neighborhood are inferred, a disconnection in the path network is identified when any two nearest nodes in the predicted path network have a substantial gap in between. The inferred width can be used for evaluating sidewalks' accessibility and walkability. For example, sidewalks with insufficient width for certain mobility devices (e.g. wheelchairs) to pass will be marked as inaccessible to those users.

4.4 Hardware Setup

In this section, we describe the hardware setup of our prototype system used in our pilot study (Section 4.5). The schematic illustrating the overall hardware setup is shown in Figure 4.2. The system consists of three hardware modules: (1) imaging module (Section 4.4.1), (2) GPS module (Section 4.4.2), and (3) computation module (Section 4.4.3). The imagery data from the imaging module and the GPS data (synchronized with each image) from the GPS module are processed at the computation module to produce mapping results.

4.4.1 Imaging Module

We use a stereo camera and a three-axis camera jig as the imaging module. The stereo camera is used to capture the RGB images and the corresponding depth image (shown in Figure 4.1), which are necessary for inferring the geolocation of infrastructures in the pedestrian environment and sidewalk connectivity. Based on our study comparing the performance of different stereo cameras in practical settings, we use a ZED

stereo camera (ZED) for better performance in diverse environments. The ZED stereo camera is a passive stereo vision camera that uses two lenses on each side of it to capture the scene. It then finds the matching points between the left view and right view of the same scene, and uses stereo vision algorithms to compute the depth of each sensed object through triangulation methods.

The ZED stereo camera is mounted onto the collection jig parallel to the ground. The spatial resolution was set to 1280×720 pixels for both the RGB images and the depth maps. The frame rate was set to 15 frames per second. The Field of View (FOV) of the ZED camera is $90^\circ \times 60^\circ$ (Horizontal x Vertical). The camera is calibrated per manufacturer specifications with tools provided by the manufacturers before collecting data.

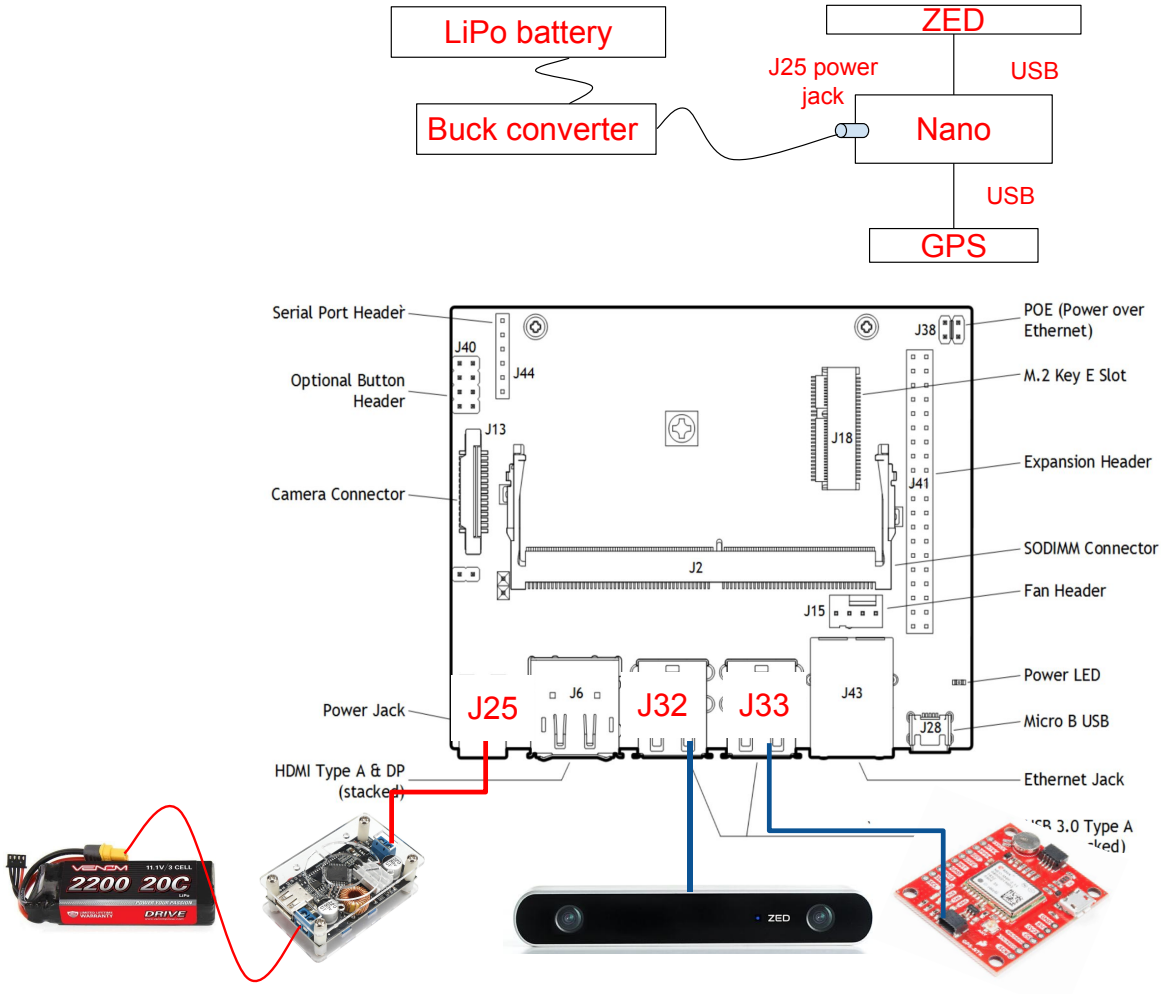


Figure 4.2: System schematic of OASIS

4.4.2 GPS Module

A GPS receiver is used to record the GPS information while the images are taken. This information, in conjunction with the imagery data, is used to infer the geolocation of each object of interest. The receiver in our prototype system is a u-blox ZED-F9P module (F9P). It has an accuracy of ± 0.75 meters when tested in outdoor pedestrian environments. The GPS location of the system is recorded every 0.5 seconds using data from the GPS receiver. At this recording rate, while the system travels in the sidewalks at normal speeds, when each captured image is synchronized with the GPS data, it has a location error of < 0.7 meters, providing sufficient resolution for mapping.

4.4.3 Computation Module

The computing device is used for on-board semantic segmentation and geographic information inference. Because NVIDIA's GPU is required to fully utilize the Software Development Kit (SDK) of the ZED camera, and to optimally run CNN-based segmentation models, we use NVIDIA Jetson Nano which operates at under 10 watts as the computing device. With this device, we are able to map the sidewalks on portable edge devices.

4.5 Pilot Study and Experiments

This section describes the pilot study we conducted with OASIS. The study consists of two parts. Section 4.5.1 provides an evaluation of OASIS at the hardware level. In Section 4.5.2, we use OASIS to map 3 neighborhoods and compare its mapping outcomes to the ones generated by human surveyors from local paratransit pathway review teams.

4.5.1 Analysis of OASIS at Hardware-level

The system performance is tested when OASIS runs continuous mapping in the pedestrian environment in two power modes. OASIS can operate at two power modes: (1) low-power (max. power: 5 W) and (2) high-power (max. power: 10 W). We evaluated OASIS under both power modes and measured different hardware metrics, including power consumption, memory usage, CPU usage, and GPU usage, for

Table 4.1: Average resources usage on OASIS during mapping

Power mode	Avg. power consumption	RAM usage	GPU utilization	CPU utilization
Low (5w)	3.98 W	3006 MB	82%	$11.4\%@921MHz$
High (10w)	5.87 W	3058 MB	84%	$10.5\%@1479MHz$

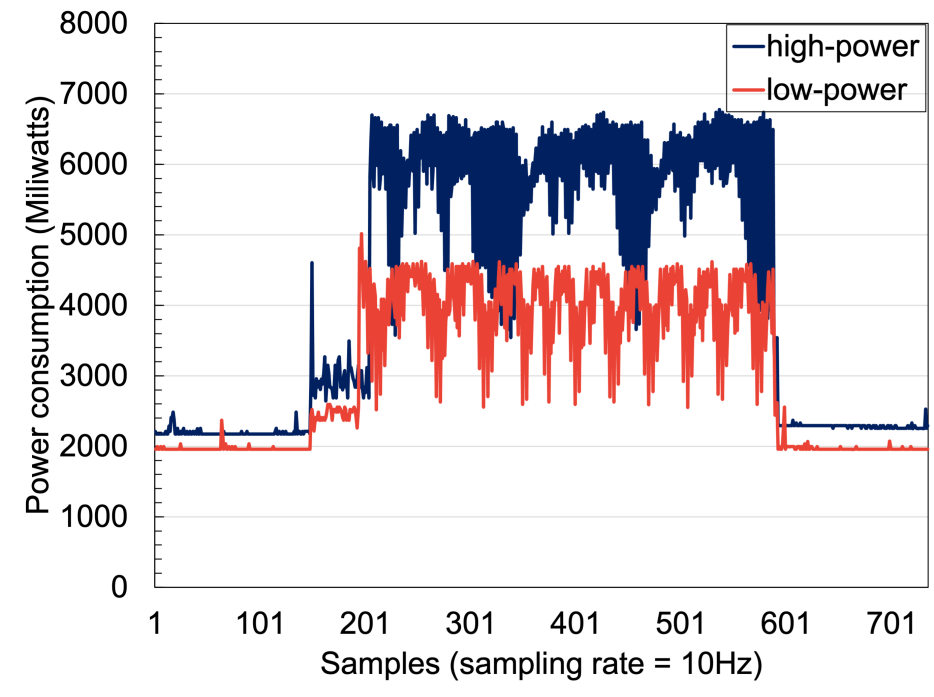


Figure 4.3: Power consumption of OASIS when mapping in two different power modes

hardware-level analysis. Figure 4.3 shows the power consumption changes over time when OASIS runs at the two different power modes, also comparing to the base power consumption when OASIS is idle in each mode. Table 4.1 shows the average resource usage of OASIS for a duration of 5 minutes during which the system is constantly running and generating mapping results. Overall, these metrics, especially low power consumption (about 4.5 watts and 6.5 watts in low-power and high-power mode respectively), suggest that OASIS can be integrated with a powered mobility device (e.g. a wheelchair or scooter) easily to power downstream applications.

4.5.2 Analysis of Pedestrian Environment Mapping with OASIS

In this section, we describe our study for mapping sidewalks and infrastructures in real pedestrian environments and compare our mapping outcome to the ones generated by human surveyors from the local para-

transit pathway review team. The studies were conducted in three geographically different areas: Bellevue, Redmond, and Seattle.

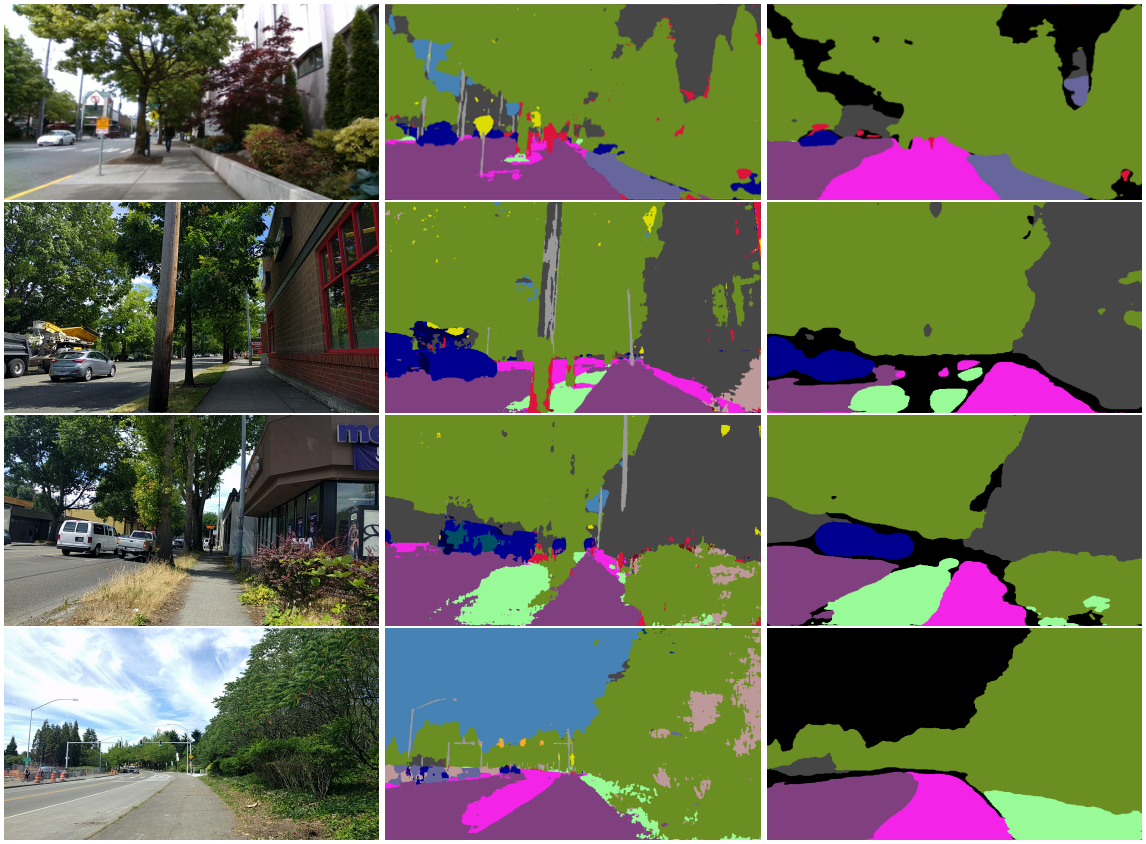


Figure 4.4: Qualitative performance of the segmentation module. Left: Input image. Middle: Segmentation masks produced by ESPNetv2. Right: Segmentation masks produced by ESPNetv2-PED. In the segmentation masks, the sidewalk is represented in magenta, the automobile road is represented in dark purple, and the terrain is represented in light green.

Table 4.2: Performance of the segmentation module in the pedestrian environment

	Building			Pole			Traffic Light			Traffic Sign			Average		
	IoU	Precision	Recall	IoU	Precision	Recall	IoU	Precision	Recall	IoU	Precision	Recall	IoU	Precision	Recall
ESPNetV2	0.815	0.654	0.895	0.441	0.841	0.659	0.423	0.577	0.645	0.539	0.781	0.618	0.547	0.713	0.704
ESPNetV2-PED	0.869	0.691	0.932	0.502	0.882	0.721	0.485	0.613	0.713	0.572	0.815	0.686	0.607	0.750	0.763

Segmentation Module Analysis

We start by evaluating the performance of the segmentation module in the pedestrian environment by comparing different segmentation models. ESPNetv2 trained on the commonly used Cityscapes dataset is used

(a) Mapped infrastructures

(b) Mapped sidewalks

Figure 4.5: Qualitative mapping results: (a) The predicted geolocations of the infrastructures on the sidewalks. (b) The predicted geolocations of sidewalks and the inferred width of each sidewalk fragment.

as the baseline, and for comparison, ESPNetv2-PED denotes the model we trained for the pedestrian environment with the method described in Section 4.3.1. The models' performance is visualized in Figure 4.4. The middle column shows the qualitative segmentation results from ESPNetv2 and the right column shows the segmentation result from ESPNetv2-PED. With ESPNetv2, because of the inherent bias that the training data is captured from a car dashboard, this model incorrectly predicts the majority of sidewalks to be the automobile road, making it impossible to generate correct sidewalk mapping information. On the other hand, with ESPNetv2-PED, we can see that the model correctly segments the sidewalk after incorporating pedestrian-centric data in training, enabling us to correctly infer sidewalk mapping information. Quantitatively, we use three standard metrics to evaluate the segmentation module: (1) pixel-wise intersection over union (IoU), (2) precision, and (3) recall of our objects of interest. As shown in Table 4.2, high precision and recall on the common objects demonstrate the system successfully detects objects in real pedestrian environments. We also observe that ESPNetv2-PED outperforms the baseline model in the pedestrian environment in every category.

Mapping Module Analysis

While the system travels in the pedestrian pathways, street-side images are captured by the camera and processed by the system to generate mapping predictions. Figure 4.5a shows examples, where the infrastructures of pedestrians' interest are segmented by the system, and their distance and geolocations are inferred. On average, 62 obstacles that may impede pedestrians' travel are identified and located per kilometer of sidewalks. As an example, the bus station in Figure 4.5a is predicted as a building and it will be treated as an obstacle in the navigation setting, similar to the situation where part of a building extends into the sidewalk and blocks the pedestrians' way. Besides mapping the infrastructures in the sidewalks, more importantly, we map the sidewalks to infer pedestrian path connectivity and walkability. Figure 4.5b shows examples of the mapped sidewalk fragments, with their geolocation and width inferred. Using the inferred mapping of these fragments, a pedestrian path network is generated.

To quantitatively evaluate the mapping module, we compare the predicted geolocation of each static object to its real geolocation in the frames that were manually annotated. The location error of our mapped objects of interest (measured as the distance between two points in meters) is shown in Table 4.3. The low average error demonstrates the efficacy of OASIS in mapping the common infrastructures on sidewalks. More importantly, to evaluate the predicted sidewalk network graph, we provide quantitative measures for both the predicted location and the predicted width for each of the three cities as shown in Table 4.4. The predictions are compared to the data collected and maintained by the city. Low RMSE demonstrates that the system effectively predicts the location and width of sidewalks. In addition, to evaluate the mapped sidewalks at the network graph level, we use metrics similar to the ones proposed by [2] for road topology measurements, namely the precision, recall, and F1 score based on the assignments of predicted sidewalk edges (or crossing edges) to the corresponding edges in the ground truth graph. As shown in the last three columns in Table 4.4, our method maintains high precision and recall across three areas with significantly different built environments, demonstrating the mapping efficacy and accuracy of our method.

Table 4.3: Error measure of mapped static objects locations

	Building	**Pole**	**T light**	**T sign**	**Average**
Mean Error (m)	0.511	0.620	0.424	0.441	0.496
Standard Deviation Error (m)	0.217	0.122	0.102	0.089	0.136
Root Mean Squared Error (m)	0.554	0.632	0.436	0.449	0.514

Table 4.4: Quantitative evaluation: sidewalks mapped by OASIS compared to human annotations

	Location error (m)			Width error (m)			Precision	Recall	F1
	Mean	STDEV	RMSE	Mean	STDEV	RMSE			
Redmond	0.47	0.12	1.49	0.11	0.03	0.11	0.94	0.98	0.96
Bellevue	0.68	0.22	0.71	0.28	0.07	0.29	0.92	0.98	0.95
Seattle	0.53	0.11	0.54	0.14	0.04	0.15	0.96	0.99	0.97
Average	0.56	0.15	0.91	0.18	0.05	0.18	0.94	0.98	0.96

4.5.3 Towards Efficient and Scalable Pathway Review Process with OASIS

One main consideration in our study was the feasible deployment and adoption of OASIS to assist pathway review teams in larger-scale applications. During our pilot study, we found OASIS has five benefits (1) While the system can be guided at a typical walking speed, the average time for mapping one mile of sidewalks is reduced six-fold (specifically, it took 120 min on average for an unassisted reviewer to collect one mile of pathways as opposed to 20 minutes with OASIS). OASIS can reduce the time and cost of collecting sidewalk mapping data by reducing the effort of human surveyors. (2) As discussed in Section 4.5.2, our method can provide accurate mapping information that is difficult to collect manually with consistency (e.g. width of sidewalks and every infrastructure in the scene which is tedious to collect), in addition to providing accurate location and connectivity data of the sidewalks. (3) The data output is immediately available in an open and standardized format per Opensidewalk Schema specifications [11], resolving a common issue with denoising and postprocessing that the manual surveyor collection requires. This facilitates the data entry stage and improves the probability of data reuse and consumption by other stakeholders or downstream applications. (4) The output of OASIS is the mapped network data in the pedestrian layer, raw image data can be discarded for public privacy concerns. (5) Being built around portable edge devices, OASIS can be easily integrated into powered mobility devices and the data can be easily maintained up-to-date with minimal human intervention. Though further adoption studies are required, the pilot findings suggest that OASIS can be adopted in audits and applications where fast and accurate assessment and re-assessment of sidewalks are needed. In addition, OASIS has minimal reliance on skilled surveyors once deployed at scale, making it suitable for larger-scale applications and organizational efforts.

4.6 Discussion

Novel technologies helping cities manage data pipelines about everything inclusive of its services and public infrastructure will be the driving force for future cities. Pedestrian ways, sidewalks, and footways are of primary concern for sustainable, accessible cities that are attempting to influence city inhabitants to make use of active transportation options rather than private vehicles. In this chapter, we propose a novel urban sidewalk assessment approach using computer vision (machine learning) techniques on portable (edge) devices. We have implemented and experimented with OASIS and deployed it toward mapping sidewalk environments. OASIS enables private or public organizations to map sidewalks and their connectivity (in the form of a routable graph) quickly and accurately.

A primary contribution of OASIS for infrastructure mapping is in producing consistent, standardized data at scale. There is an opportunity for the field to contribute to an effective and resilient data ecosystem that will provide many stakeholders, including transportation agencies, municipalities, and public and private civic actors alike with relevant data to analyze, metricize, prioritize, and possibly manage mobility and accessibility at city scale. Performing all computing on the edge may be attractive to municipalities where the citizenry would reject the use of cloud infrastructure to obtain and potentially keep sidewalk imagery that they might deem as an invasion of privacy or security. The availability of the data on a portable edge device platform opens further opportunities to inform (via a map) automated applications that navigate in sidewalk environments (such as autonomous wheelchairs and self-driving delivery robots). The output mapping results that are open and shared are a crucial step towards this vision since they are not reliant on any proprietary street-side imagery contributions. Specifically, in the case of accessibility data in the PROW, prior work demonstrated that many small and underfunded stakeholders confront similar problems in performing relevant, consistent data collection. Having data contributions through collaborative mapping tools and distributed through collaborative data commons like OSM can improve overall non-motorized transportation assessment and improvement. Travelers stand to benefit through downstream use in routing and wayfinding applications (e.g. Moovel, TransitApp, AccessMap [97, 75]). The use of collaborative tools and shared data allows for citizen participation in data vetting and collection which can increase trust among agencies and accessibility advocates. Concluding, we believe that an automated sidewalks assessment systems approach is of core importance for the development and management of future urban active transportation networks.

Chapter 5

Segmentation Evaluation Metrics for Model Selection and Explainability in Practical Applications

5.1 Introduction

As discussed in Section 2.2.1, semantic segmentation plays an important role in mapping outdoor environments. We employ semantic segmentation models in both of our methods: Prophet (Chapter 3) and OASIS (Chapter 4). However, many currently used metrics do not provide a proper meter of region-based agreement with a given ground-truth or provide an expectation about broken regions that model segmentation may yield, nor offer model performance explainability that can aid model selection in practical applications. In this chapter, we introduce quantitative evaluation metrics that provide granular accounting for over- and under-segmentation for semantic segmentation.

Semantic segmentation is a crucial task in computer vision that groups image parts into the same object class. It is widely used in real-world visual scene understanding applications, such as autonomous driving, robotics, and medical diagnostics. Recent advancements in machine learning, particularly neural networks [59, 98, 99, 100], and hardware technology [46, 49, 101, 102, 103] have significantly improved semantic segmentation performance and made time-sensitive domains, like autonomous driving [4, 44, 104], practi-

(a) Over-segmentation

(b) Under-segmentation

Figure 5.1: Example that illustrates over- and under-segmentation issues in semantic segmentation. (**Left** to **right**: RGB image, ground-truth, prediction, and error metrics). In (a), the back of the bus is over-segmented (one region is segmented into three). In (b), people are under-segmented (three groups are segmented as one). Though the IoU error (1 - IoU score) is low in both cases, the IoU metric does not reflect these distinctions. However, these issues can be explained using the proposed metrics for over- (ROM) and under-segmentation (RUM).

cally feasible.

5.1.1 Motivation

In the field of autonomous driving, there has been extensive research and mapping for automobile roads, but pedestrian street-side environments have been neglected. Real-time mapping of pedestrian infrastructure is crucial for path planning in transportation planning, delivery robots, and autonomous wheelchair navigation, and can provide information to mobility applications for people with disabilities. Specifically, such mapping answers important questions about *where* paths are, *how* paths are connected, and *whether paths have the amenities and landmarks* necessary to traverse them efficiently and safely. Providing street-side maps will improve accessibility for all people to navigate urban spaces. Studies of semantic segmentation in pedestrian environments showed challenges in deploying convolutional neural networks (CNNs) in real-time, low-computing environments. An *ideal model* in this application space should predict connected regions that represent pedestrian pathways (e.g., sidewalks, footways, and road crossings) in a scene, and common objects in pedestrian environments (e.g., poles, fire hydrants, and benches.). In both these cases, region-wise accuracy is more important than pixel-wise accuracy. Answering questions like "How many paths are

there and how are they connected in the current scene?" can only be achieved through segmentation results with region-wise fidelity.

5.1.2 Model Evaluation for Explainability

Semantic segmentation divides an image into segments corresponding to different objects or parts of objects. This is achieved by assigning a class label to each pixel in the image. However, as an important subtask of image understanding, it also contributes to the discovery of groups of objects and the identification of semantically meaningful distributions and patterns in input image data. In [105], segmentation is framed as a graph partitioning problem, and the normalized cut criteria is proposed for measuring the total similarity and dissimilarity of partitions. In [106], superpixel (effectively regions consisting of multiple pixels) is used to advocate for evaluating segmentation in practical applications. Semantic segmentation decomposes a surrounding scene into meaningful semantic regions that are useful for decision-making in downstream applications. For instance, when mapping in the pedestrian environment, it is important to identify which connected regions are correctly mapped to real pathways. In sum, uses of semantic segmentation are not pixel- but *region-based*, as evidenced in their application to aggregate semantically rich information about the surroundings used to assess navigation options for people with visual disabilities (e.g., Mehta et al. [107] and Yang et al. [108]).

Most quantitative evaluation metrics (e.g., mean Intersection over Union, mIoU) currently used to evaluate performance or error in semantic segmentation algorithms are based on the pixel-wise similarity between ground-truth and predicted segmentation masks. Though these metrics effectively capture certain properties about clustering or classification performance, they fail to provide a proper meter of region-based agreement with a given ground-truth or provide an expectation about broken regions that model segmentation may yield. We assert that quantitative measures of *region segmentation similarity* (rather than pixel-wise or boundary similarity) will prove useful to perform model selection and to measure the consistency of predicted regions with a ground-truth in a manner that is invariant to region refinement or boundary ambiguity. Such measures will prove practically useful in applications of semantic segmentation where performance is affected by over- and under-segmentation. For clarity, a predicted region is said to contribute to *over-segmentation* when the prediction overlaps a ground-truth region that another predicted region also

overlaps. A predicted region is *under-segmented* when the prediction overlaps two or more ground-truth regions. Examples of over- and under-segmentation are shown in Figure 5.1.

This chapter introduces quantitative evaluation metrics for semantic segmentation that provide granular accounting for over- and under-segmentation. We express the empirical probability that a predicted segmentation mask is over- or under-segmenting and also penalize model-segmentations that repeatedly over-segment the same region, causing large variations between model-prediction and ground-truth. We demonstrate how these issues affect currently used segmentation algorithms across a variety of segmentation datasets and models and that currently available metrics do not differentiate between models since they do not measure these issues directly.

5.2 Existing metrics

Currently, most deployed segmentation models, including the ones with encoder-decoder architectures [46, 48, 49, 45, 59, 101] and lightweight models for resource-constrained deployment [62, 64, 66, 109, 110], are typically trained with strong pixel-level supervision, using metrics such as Intersection over Union (IoU) error to evaluate consistency with annotated image segments [111]. However, in practice, these metrics are ill-suited for discriminating between effective and ineffective region-based segmentation consistency, which is a critical factor in real-world applications.

To address this limitation, the chapter highlights two questions to differentiate between metrics and assess their relevance as performance metrics for semantic segmentation: (1) Does the metric correlate with the level of agreement between the model segmentation and ground truth?, and (2) Is the metric robust to region refinements or small ambiguities in region boundaries that naturally arise in images?

IoU and other pixel-wise similarity metrics. IoU evaluates pixel-wise similarity irrespective of how pixels are grouped in bounded regions. It is the most common metric for evaluating classification accuracy and is the performance metric of choice for models trained with strong pixel-level supervision. Given a segmentation mask S_I and the ground-truth mask G_I for Image I, the IoU is calculated as $\frac{S_I \cap G_I}{S_I \cup G_I}$. This metric provides a measure of similarity ranging from 0 (when there is no similar pixel labeling between G_I and S_I) to 1 (when there is full concordance between the two pixel assignments G_I and S_I).

IoU gives a measure of pixel-wise assignment similarity that is associated, but not always correlated, with the accuracy of region classification in the single class case. Due to this relationship, the metric is also used as a surrogate for assessing segmentation performance in various segmentation datasets (e.g., PASCAL VOC [112] and MS-COCO [113]). For instance, IoU correctly identifies the difference in the overlap between predicted regions and ground-truth in synthetic examples in Figures 5.2c (lowest IoU error) and 5.2a (highest IoU error). Additionally, it appropriately penalizes and differentiates models that tend to false-predict regions. However, the correlation between accurate pixel classification and proper segmentation extends only in cases where segmented regions are contiguously uninterrupted, i.e., where it is appropriate to concatenate regions in which geometrically adjacent pixels are similarly labeled. The *horse* leg segmentation (second example in Figure 5.4) and the *towel* class segmentation (third example in Figure 5.4) apply exactly in this instance, where IoU does not dovetail with region agreement between the segmentation mask and ground-truth. As shown in the synthetic example in Figures 5.2b, 5.2e, and 5.2h, the metric is also not perturbed by significant region-wise refinements.

Other pixel-wise metrics similar to IoU (e.g., pixel accuracy and dice score [46, 114]) also measure pixel-level correspondence between the predicted segmentation mask and ground-truth. These metrics account for the correctness of pixel-based prediction at a global level but again fail to capture region-based agreement between predicted regions and ground-truth. They also fail to robustly capture or differentiate models that provide improvements in region boundary agreement with ground-truth (see Figure 5.2k and Figure 5.2n).

GCE. Martin et al. [115] proposed region-based metrics for evaluating segmentation performance. The global consistency error (GCE) and local consistency error (LCE) were introduced to measure the region-wise inconsistency between a prediction mask, S_I, and a ground-truth, G_I. For a given pixel x_i, let the class assignment of x_i in prediction mask S_I be denoted by $C(S_I, x_i)$. Likewise, the class assignment for x_i in the ground-truth, G_I, is represented by $C(G_I, x_i)$. The Local Refinement Error (LRE) is defined at pixel x_i as:

$$LRE(S_I, G_I, x_i) = \frac{\mid C(G_I, x_i) \setminus C(S_I, x_i) \mid}{\mid C(G_I, x_i) \mid} \tag{5.1}$$

, where $\setminus$ denote set difference, and $\mid \cdot \mid$ denote the cardinality of a set. Since the LRE is one-directional and some inconsistencies arise where the set difference $\mid C(G_I, x_i) \setminus C(S_I, x_i) \mid$ accounts for greater

divergence between segmentations, the Global Consistency Error (GCE) was defined for an entire image with N pixels as:

$$GCE(S_I, G_I) = \frac{1}{N}\min \quad \begin{matrix} \left(\sum_{i=1}^{N} LRE(S_I, G_I, x_i) \right. , \\ \left. \sum_{i=1}^{N} LRE(G_I, S_I, x_i) \right) \end{matrix} \tag{5.2}$$

GCE more stringently accounts for consistencies between the two segmentation masks than LCE. GCE, like pixel-based measures, still captures global pixel classification error (similar high and low measures are observed for Figures 5.2a and 5.2c, respectively). It also offers some additional penalties for false positive predictions (note the difference between GCE and IoU for Figures 5.2e and 5.2k). Unlike pixel-wise methods, this metric amplifies disagreement between prediction and ground-truth regions. However, GCE still fails to capture the difference between predictions that refine region segmentation (for example, the GCE for Figure 5.2j is worse than for 5.2i, but not by much).

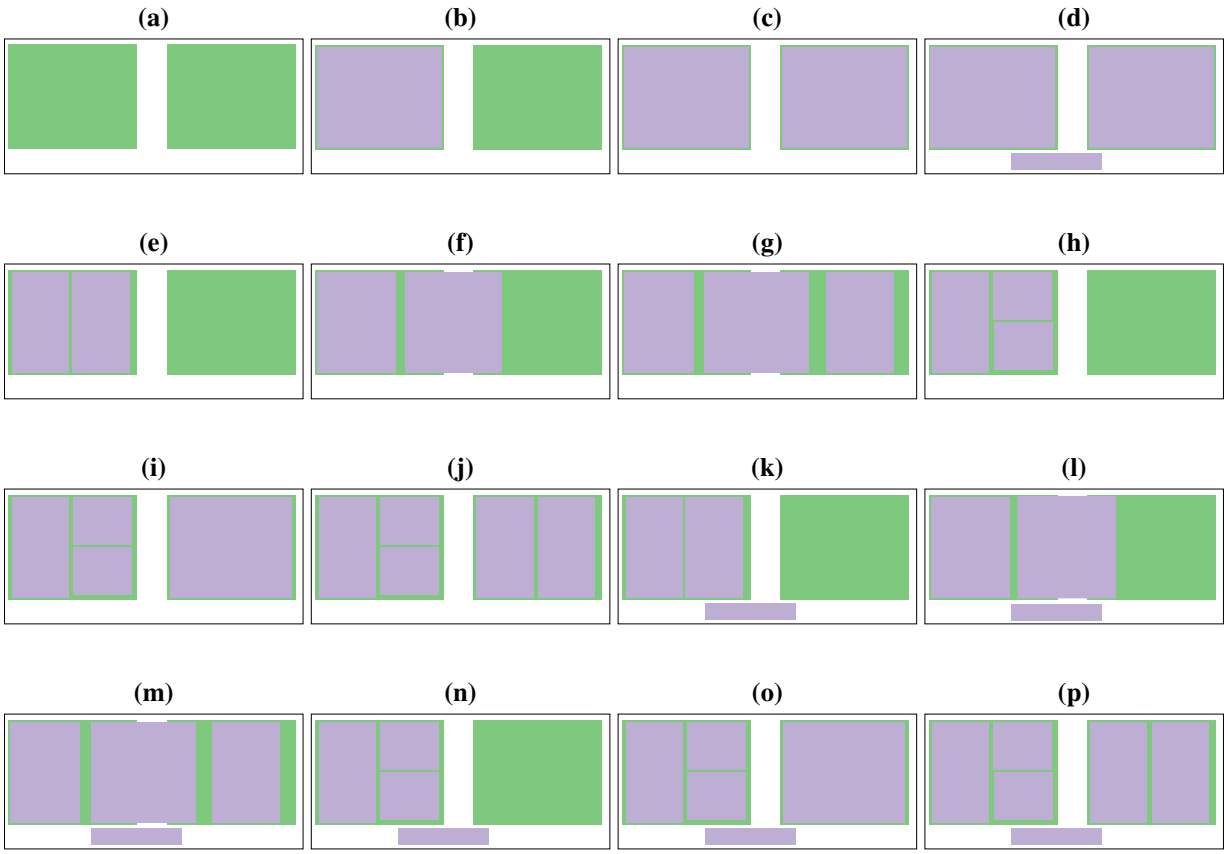
Partition distance. In [116], an error measure was proposed based on partition distance, which counts the number of pixels that must be removed or moved from the prediction for it to match the ground-truth. The partition distance penalizes for under- and over-segmentation to some extent, but partitioning a region in either the prediction or ground-truth into multiple subsets will render the same partition distance. The partition distance tracks much like the GCE for that reason, so we do not calculate it here.

Persello's error. In [117], Persello's error (PE) was proposed to measure the local error for over-segmentation (PE-OS) and under-segmentation (PE-US) based on the ratio of the area of the largest overlapping prediction region and the area of a ground-truth region. The measure identifies region agreement and penalizes for larger sizes of discrepancy. For each ground-truth region, the error is computed based on the ratio of the area of the largest overlapping predicted region and the area of the ground-truth region. PE-OS increases with over-segmentation (e.g., Figure 5.2b has a better score than either Figure 5.2e or Figure 5.2h). However, it provides the same score for Figures 5.2e and 5.2h because they have the same-sized largest predicted region for the given ground-truth region.

Average precision. Average precision (AP) is the evaluation metric used in the MS-COCO object detection and instance segmentation challenge [113]. Instead of directly measuring pixel-wise concordance, AP uses IoU as a threshold and measures region-wise concordance. If a given prediction region's IoU with a ground-

truth region exceeds the threshold, it is counted as a true positive (TP); otherwise, it is counted as a false positive (FP). The precision $\left(\frac{TP}{TP+FP}\right)$ is computed and averaged over multiple IoU values, specifically from 0.50 to 0.95 with 0.05 increments. We computed $AP^{IoU=.50}$ (same as PASCAL VOC metric) and $AP^{IoU=.75}$ (strict metric) for the synthetic examples in Figure 5.2. AP has its advantages over pure pixel-wise measures when assessing error instance-wise, but it will fail when two segmentation masks have similar precision values but suffer from different degrees of over-segmentation. For example, $AP^{IoU=.50}$ gives the same error measure for Figure 5.2h, 5.2k, 5.2l and 5.2p, but Figure 5.2h is considered a worse over-segmentation case than 5.2k and 5.2l, while 5.2p represents the worst over-segmentation case among all the examples. In terms of the stricter measure, $AP^{IoU=.75}$ indicates that the majority of these examples have an equal error of 1 without differentiating which one is worse.

Panoptic quality. Panoptic quality (PQ), which combines the semantic segmentation and instance segmentation tasks [118], is a relevant metric but requires having ground-truth annotations for both semantic and instance segmentation. It provides a unified score that measures pixel-wise concordance and penalizes false positive (FP) and false negative (FN) regions. PQ reflects the gross pixel-wise and region-wise error jointly in a predicted segmentation, but it does not differentiate between under- and over-segmentation. Although FP and FN regions indirectly relate to the over- and under-segmentation issues, PQ penalizes these errors in the same direction and may give the same score to a prediction that over-segments as to one that under-segments.



Metric	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
IOU Error	1.00	0.53	0.07	0.11	0.58	0.51	0.29	0.58	0.13	0.16	0.61	0.54	0.33	0.60	0.17	0.20
GCE	1.00	0.27	0.07	0.12	0.34	0.38	0.29	0.26	0.12	0.14	0.40	0.42	0.33	0.32	0.17	0.20
PE-OS	1.00	0.53	0.07	0.06	0.78	0.61	0.49	0.78	0.33	0.58	0.78	0.61	0.49	0.78	0.33	0.58
1-$AP^{IOU=.50}$	1.00	0.00	0.00	0.33	0.50	0.50	0.33	0.67	0.50	0.60	0.67	0.67	0.50	0.75	0.60	0.67
1-$AP^{IOU=.75}$	1.00	0.00	0.00	0.33	1.00	1.00	1.00	1.00	0.75	1.00	1.00	1.00	1.00	1.00	0.8	1.00
1-PQ	1.00	0.67	0.51	0.64	0.76	0.75	0.71	0.82	0.69	0.74	0.84	0.82	0.78	0.86	0.76	0.79
ROM (Ours)	0.00	0.00	0.00	0.00	0.46	0.46	0.96	0.76	0.64	0.99	0.32	0.32	0.91	0.64	0.54	0.98

Figure 5.2: Synthetic examples that illustrate and compare metrics. While these primarily address over-segmentation, they can be generalized to under-segmentation by inverting color interpretation of segmentation and ground-truth. Each panel represents an image overlay, with ground-truth regions (denoted in green) and predicted regions (denoted in purple). Panels are designed to represent various situations: (5.2a) no positive class prediction at all; (5.2b) near perfect prediction for one ground-truth region; (5.2c) near perfect prediction; panels (5.2e through 5.2j) different aspects of over-segmentation, with panel (5.2j) representing the worst over-segmentation case among those; panels (5.2k) through (5.2p) correspond to panels (5.2e) through (5.2j), respectively, but with an additional false-positive predicted region. These metrics include IoU error, GCE [115], Persello's error (PE) [117], average precision (AP) [113], panoptic quality (PQ) [118] and ROM (ours). Note that IoU error is the representative pixel-wise measure, while GCE, PE, AP and ROM are region-wise measures for semantic segmentation; PQ is a metric for the related instance segmentation task, not semantic segmentation.

5.3 Interpretable Region-based Measures for Over- and Under-segmentation Errors

We now introduce two measures that combine the desirable properties of: (1) accounting for model disagreement with ground-truth in over- and under-segmentation cases, and (2) displaying sensitivity for local over- and under-segmentation refinements to model predictions or refinements to ground-truth based on ambiguous semantic segmentation boundaries. These properties are relevant primarily to quantify the consistency of segmentation results (see Section 5.3.3).

5.3.1 Region-wise Metric Estimations

We now define metrics that isolate errors due to over- and under-segmentation in region-based tasks like semantic segmentation.

Region-wise over-segmentation measure (ROM). Let I be an RGB image, G_I be the ground-truth segmentation mask for I, and S_I be the model-predicted segmentation mask for I in dataset $\mathcal{D}$ (validation or test set), each with spatial dimensions of $w \times h$, where w and h correspond to width and height, respectively. Assume there are K semantic classes in $\mathcal{D}$. A valid label assignment (also referred to as a segmentation) in G_I and S_I maps each pixel $x_{r,c}$ to a single integer label $k \in [0, \cdots, K-1]$ representing the class assignment for that pixel. For simplicity, we assume that the background class label is 0. For each non-background class $k \in [1, \cdots, K-1]$, we convert G_I and S_I to their binary formats, G_{Ib} and S_{Ib} , as follows:

$$G_{Ib}[k, r, c] = \begin{cases} 1, & G_I[r, c] == k \\ 0, & \text{Otherwise} \end{cases} \tag{5.3}$$

$$S_{Ib}[k, r, c] = \begin{cases} 1, & S_I[r, c] == k \\ 0, & \text{Otherwise} \end{cases} \tag{5.4}$$

where $r \in [0, h-1]$ and $c \in [0, w-1]$ correspond to row and column indices of a pixel in the image.

Each spatial plane in G_{Ib}, i.e., $G_{Ib}[k]$, consists of $N = \|G_{Ib}[k]\|$ separate contiguous regions, where $\|\cdot\|$ is an operator that counts the number of separate contiguous regions in $G_{Ib}[k]$. Therefore, we can

represent each spatial plane $G_{Ib}[k]$ as a set of connected regions $G_{Ib}[k] = \{g_1, g_2, ..., g_N\}, k \in [1, K-1]$. We assert that $g_i \cap g_j = \varnothing, \forall i \neq j$. Similarly, we represent each spatial plane in S_{Ib}, i.e., $S_{Ib}[k]$ as a set of M connected regions $S_{Ib}[k] = \{s_1, s_2, ..., s_M\}, k \in [1, K-1]$, where $s_i \cap s_j = \varnothing, \forall i \neq j$. We refer to the complete set of binary planes that satisfy these constraints as a valid segmentation ground-truth pair and look for measures of the form $d(S_{Ib}, G_{Ib})$.

We begin by evaluating the performance of the prediction mask by making a detailed accounting of over-segmented foreground regions. A model-predicted region contributes to the over-segmentation count when the prediction region overlaps with a ground-truth region that itself overlaps with more than one model-predicted region. Fig 5.1a shows an example of over-segmentation.

We denote regions in S_{Ib} contributing to over-segmentation as $S_O^I = \{s_i \in S_{Ib}\}$, where $s_i \cap g_l \neq \varnothing \wedge s_j \cap g_l \neq \varnothing,\ i \in [1, \cdots, M],\ j \in [1, \cdots, M],\ l \in [1, \cdots, N],\ i \neq j$. S_O^I marks model-predicted region $s_i \in S_{Ib}$ as included in the over-segmentation count; it must overlap with at least one ground-truth region $g_l \in G_{Ib}$, while the ground-truth region g_l must overlap with at least one other prediction region $s_j \in S_{Ib}$. The total number of regions that contribute to over-segmentation with respect to ground-truth regions are $\|S_O^I\|$.

Similar accounting identifies ground-truth regions involved in over-segmentation (i.e., overlapping more than one predicted region). We denote regions in G_{Ib} that are involved in over-segmentation as $G_O^I = \{g_i \in G_{Ib}\}$, such that $g_i \cap s_j \neq \varnothing \wedge g_i \cap s_l \neq \varnothing$, $j \in [1, \cdots, M]$, $l \in [1, \cdots, M]$, and $j \neq l$. This definition asserts that a ground-truth region $g_i \in G_{Ib}$ is counted towards the over-segmentation count if it overlaps with at least two different model-predicted regions, $s_j \in S_{Ib}$ and $s_l \in S_{Ib}$. $\|G_O^I\|$ denotes the total number of ground-truth regions that are involved in over-segmentation.

We are interested in a measure that combines desirable statistical properties with the ability to count disagreements between S_{Ib} and G_{Ib} and accommodate model prediction refinement. Specifically, we should not use any thresholds or fixed pixel-percent requirements to determine the overlap between regions in the segmentation mask and ground-truth. This is important in the context of many methods that are used to merge or fuse segmentation results based on geometric proximity or filtering methods [119].

Given the ground-truth labeling, G_{Ib}, the probability that a model-predicted region $s_i \in S_{Ib}$ contributes to over-segmentation can be represented as $\frac{\|S_O^I\|}{\|S_b^I\|}$, and the probability that a ground-truth region $g_i \in G_{Ib}$ is

over-segmented can be represented as $\frac{\|G_O^I\|}{\|G_b^I\|}$. Our goal is to express the empirical probability that a predicted segmentation mask is over-segmenting. We define this index as the region-wise over-segmentation ratio (ROR):

$$ROR = \frac{\|G_O^I\|\|S_O^I\|}{\|G_b^I\|\|S_b^I\|}. \tag{5.5}$$

This measure takes values between 0 and 1 and indicates the percentage of elements in G_b^I and S_b^I that relate to over-segmentation. In the worst case, every single model-predicted region and ground-truth region is either contributing to or involved in over-segmentation, giving ROR a value of 1.

ROR may fail to differentiate between cases where any elements in S_O^I are further split by subsets (e.g., Figures 5.2e and 5.2h). Therefore, we seek to penalize the ROR score based on the total number of over-segmenting prediction regions. Let $S_{g_i}^I = \{s_j \in S_{Ib}\}$, such that $s_j \cap g_i \neq \varnothing$ denotes a set of all predicted regions that overlap with a given $g_i \in G_b^I$. To penalize the ROR score, we compute over-segmentation aggregate term (m_o) from $S_{g_i}^I$ for all $g_i \in G_{Ib}$ as:

$$m_o = \sum_{g_i} max\left(\|S_{g_i}^I\| - 1, 0\right) \tag{5.6}$$

Thus, m_o expresses an aggregate penalty structure accounting for each ground-truth region $g_i \in G_{Ib}$ that is overlapped with at least one model-predicted region. To compute the final region-wise over-segmentation measure (ROM), we multiply ROR by m_o. The resultant value will never be negative because both ROR and m_o are greater than or equal to 0. Since the resultant value can be very large, we scale it between 0 and 1 using a $\tanh$ function. Doing so confers the property of taking on a wider range of values over $[0, 1)$, increasing the sensitivity of the measure to classes that habitually over-segment[1] and allowing us to compare with existing metrics.

$$ROM = \tanh\left(ROR \times m_o\right) \tag{5.7}$$

A ROM value of zero indicates that there is no over-segmentation. In this case, $\|G_O^I\| = \|S_O^I\| = 0$, indicating that each $g_i \in G_b^I$ overlaps with at most one $s_j \in S_b^I$. A ROM value of 1 indicates that the predicted segmentation is worse and contains abundant over-segmented regions. In this case, $\|G_O^I\| = \|G_b^I\|$, $\|S_O^I\| = \|S_b^I\|$, and $m_o \to \infty$. This means that every single prediction/ground-truth region contributes

[1] Other scaling functions, such as sigmoid, can be applied. We choose $\tanh$ over sigmoid because it is centered at 0.

to over-segmentation, and each ground-truth region overlaps with an infinite number of prediction regions.

Region-wise under-segmentation measure (RUM). A similar argument follows for evaluating S_I from the under-segmentation perspective. A predicted region that contributes to under-segmentation (see example in Fig 5.1b) is represented as $S_U^I = \{s_i \in S_{Ib}\}$, where $\exists k, l \in [1..N], k \neq l, s_i \cap g_k \neq \varnothing \wedge s_i \cap g_l \neq \varnothing$. The total count of model-segmented regions that contribute to under segmentation is denoted by $\|S_U^I\|$. This representation identifies a model-predicted region as under-segmenting when it overlaps with at least two different ground-truth regions $g_k \in G_{Ib}$ and $g_l \in G_{Ib}$. Similarly, a ground-truth region is involved in under-segmentation when it overlaps with a prediction region that in turn overlaps with at least two ground-truth regions. This is represented as $G_U^I = \{g_i \in G_{Ib}\}$, s.t. $\exists j \in [1..N],\ \exists l \in [1..M],\ i \neq j,\ s_l \cap g_i \neq \varnothing \wedge s_l \cap g_j \neq \varnothing$. This representation counts ground-truth regions $g_i \in G_{Ib}$ that overlap with at least one prediction region $s_l \in S_{Ib}$, while the prediction region s_l overlaps with at least one other prediction region $g_j \in G_{Ib}$. The total number of ground-truth regions involved in under-segmentation is denoted by $\|G_U^I\|$.

The analog probabilistic terms for measuring under-segmentation: the region-wise under-segmentation ratio (RUR), under-segmentation multiplier (m_u), and region-wise under-segmentation measure (RUM) are defined as:

$$RUR = \frac{\|G_U^I\|\|S_U^I\|}{\|G_b^I\|\|S_b^I\|} \tag{5.8}$$

$$m_u = \sum_{s_i} max(\|G_{s_i}^I\| - 1, 0) \tag{5.9}$$

$$RUM = \tanh\left(RUR \times m_u\right) \tag{5.10}$$

5.3.2 Region-wise Confidence Measure

Many segmentation models confer a prediction class for each pixel and an associated probability (confidence) for each pixel prediction. Instead of looking at the confidence of each pixel individually (as was done in the PASCAL VOC object detection evaluation [112]), we propose to use the confidence of each predicted contiguous region, when available. We represent the confidence of a predicted region as the numerical mean of the confidence of all pixels enclosed in that region. When evaluating with ROM (or RUM), all pixels in a region whose confidence is lower than a certain threshold are mapped to the class *unknown*. We experiment with the effect of different thresholds in Section 5.5.

5.3.3 Qualitative Assessment

Here, we qualitatively demonstrate that the ROM/RUM metric can: (1) differentiate between models that provide fewer or more numerous over-/under-segmentation inconsistencies, and (2) adequately quantify improvements to model predictions while providing robust tolerance to small perturbations in ground-truth regions.

While ROM/RUM are valuable and effective in measuring over-/under-segmentation issues, they should not be used as measures of classification accuracy that account for gross error. For example, panels in Figure 5.2a through Figure 5.2d are equivalent from the perspective of over-segmentation error (ROM is 0 for all), indicating the lack of over-segmentation errors in all of these examples (whether due to mis-classification or not). As demonstrated in Section 5.2, widely available pixel-wise metrics can be used in conjunction with ROM and RUM to address these concerns. Future work might consider how to combine these metrics in a principled way.

The advantage of the ROM metric is that it isolates disagreement and correlates with discrepancies between model and ground-truth due to region-wise over-segmentation. The analogy is true for RUM with respect to under-segmentation. Specifically, accounting only for over-segmentation, panels in Figure 5.2e and Figure 5.2f are equivalent. Moreover, the metric is useful in accounting for over-segmentation in the average model-predicted region, therefore indicating higher over-segmentation in panels in Figure 5.2e and Figure 5.2f versus panels in Figure 5.2k and Figure 5.2l. This demonstrates the ability to quantify model differences that provide fewer or more numerous over-segmentation inconsistencies. Finally, as ROM error increases in panels in Figure 5.2b, 5.2e and 5.2h, we see that the ROM differentiates and grows monotonically with over-segmentation errors in model predictions.

Notably, our approach yields a supervised objective evaluation and highlights discrepancies in overall region segmentation results, comparing it to a ground-truth. Calculations for ROM and RUM specifically avoid penalizing pixel-wise non-concordance in $g_i \cap s_j = \varnothing$ because such information has been handled by other metrics (like IoU). ROM and RUM help solely with evaluation from the over-and under-segmentation perspective, respectively.

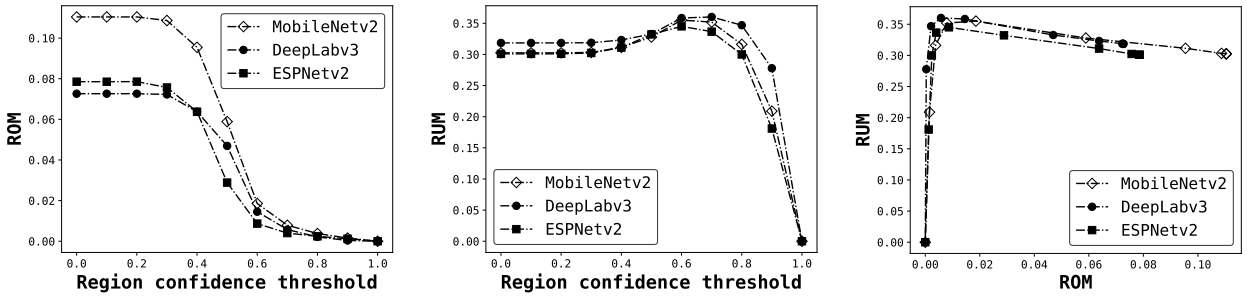
5.4 Experimental Set-up

To demonstrate the effectiveness of our proposed metrics, we evaluate and compare them with existing metrics across different semantic segmentation datasets and different segmentation models.

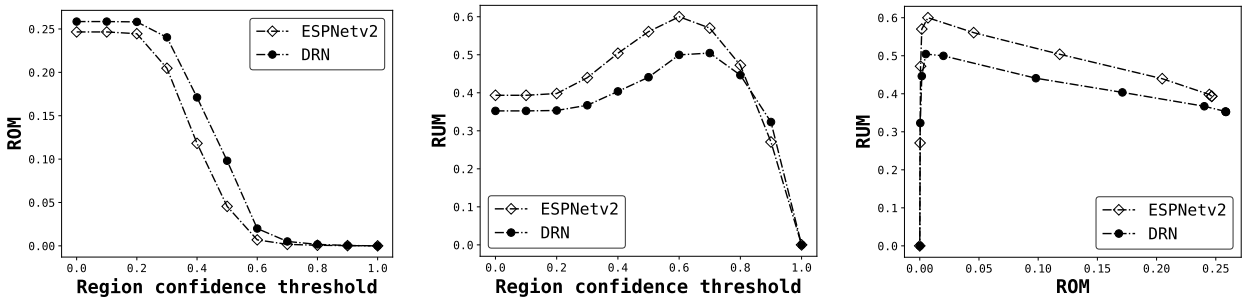
Baseline metrics: We use the following metrics for comparison: (1) mIoU error, (2) Dice error, (3) pixel error, (4) global consistency error (GCE), (5) Parsello's error (PE-OS and PE-US for over- and under-segmentation), and (6) Average precision error ($1 - AP^{IoU=.50:.05:.95}$). These metrics report a similarity score between 0 and 1, where 0 means that predicted mask S_I and ground-truth mask G_I for image I are the same, while 1 means they are not similar.

Semantic segmentation models and datasets. We study our metric using state-of-the-art semantic segmentation models on three standard datasets (PASCAL VOC 2012 [112], ADE20K [120], and Cityscapes [35]). These models were selected based on network design decisions (e.g., light- vs. heavy-weight), public availability, and segmentation performance (IoU), while the datasets were selected based on different visual recognition tasks (PASCAL VOC 2012: foreground-background segmentation, ADE20K: scene parsing, and Cityscapes: urban scene understanding). In this chapter, we focus on semantic segmentation datasets because of their wide applicability across different domains, including medical imaging (e.g., [121]). However, our metric is generic and can be easily applied to other segmentation tasks (e.g., panoptic segmentation) to assess over- and under-segmentation performance.

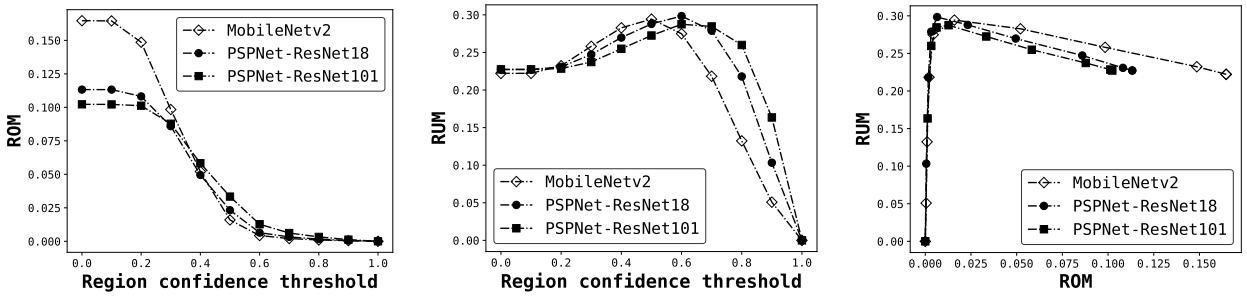
For the PASCAL VOC 2012, we chose Deeplabv3 [102] as our heavy-weight model and two recent efficient networks (ESPNetv2 [66] and MobileNetv2 [64]) as light-weight models. For ADE20K, we chose PSPNet [101] (with ResNet-18 and ResNet-101 [122]) as heavy-weight networks and MobileNetv2 [64] as the light-weight network. For Cityscapes, we chose DRN [50] and ESPNetv2 [66] as heavy- and light-weight networks, respectively.



(a) The PASCAL VOC dataset



(b) The Cityscapes dataset



(c) The ADE20K dataset

Figure 5.3: ROC curves for evaluating over- (ROM) and under-segmentation (RUM) on different datasets. Note the rightmost plot in each sub-figure is between ROM and RUM errors. Therefore, a lower area under ROC curve means better performance.

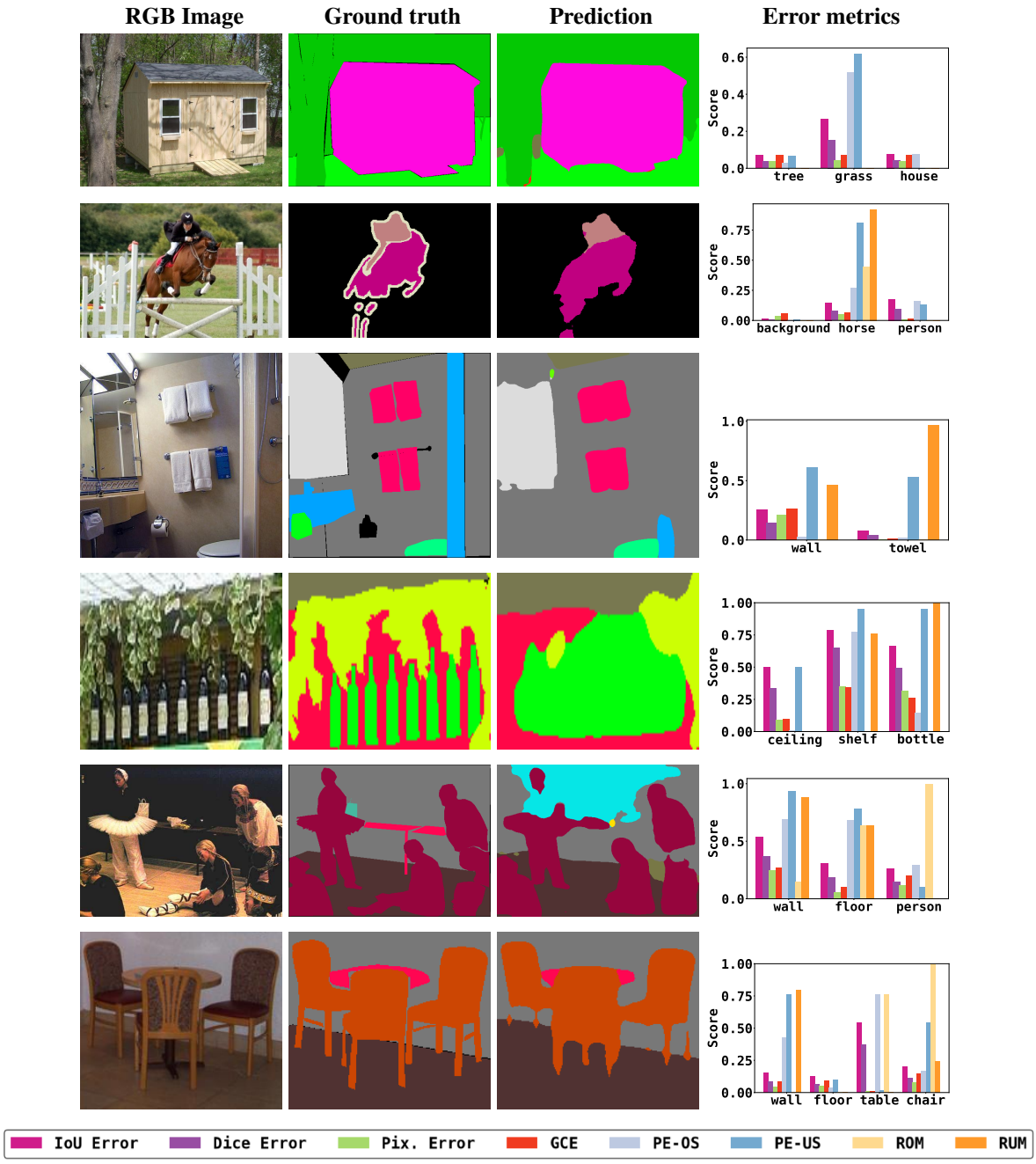


Figure 5.4: Qualitative examples illustrating different metrics for over- and under-segmentation. The example in the top row is a near-perfect segmentation. The second row shows an under-segmentation example on the *horse* class. The third row shows an under-segmentation example on the *towel* class. The fourth row shows an under-segmentation example on the *bottle* class. The fifth row shows an over-segmentation example on the *person* class. The last row shows an over-segmentation example on the *chair* class. The error metric legend is shown in the bar graphs (bottom).

Table 5.1: Performance Comparison Using Different Metrics. Our Metrics Quantify Over- and Under-segmentation Errors, Thus Explaining The Source of Error in Semantic Segmentation Performance.

(a) PASCAL VOC 2012

Error metrics	DeepLabv3	ESPNetv2	MobileNetv2
mIoU Error	0.18	0.25	0.24
Dice Error	0.13	0.20	0.18
Pixel Error	0.11	0.13	0.13
GCE	0.07	0.09	0.10
PE-OS	0.31	0.37	0.35
PE-US	0.41	0.45	0.44
AP Error	0.67	0.70	0.70
ROM (Ours)	0.07	0.08	0.11
RUM (Ours)	0.32	0.30	0.30

(b) Cityscapes

Error metrics	DRN	ESPNetv2
mIoU Error	0.30	0.38
Dice Error	0.22	0.30
Pixel Error	0.17	0.18
GCE	0.13	0.16
PE-OS	0.46	0.55
PE-US	0.50	0.62
AP Error	0.81	0.86
ROM (Ours)	0.26	0.25
RUM (Ours)	0.35	0.40

(c) ADE20K

Error metrics	PSPNet		MobileNetv2
	ResNet-18	ResNet-101	
mIoU Error	0.62	0.58	0.65
Dice Error	0.43	0.40	0.46
Pixel Error	0.21	0.19	0.24
GCE	0.19	0.17	0.21
PE-OS	0.53	0.50	0.56
PE-US	0.52	0.49	0.54
AP Error	0.73	0.69	0.76
ROM (Ours)	0.11	0.10	0.16
RUM (Ours)	0.23	0.23	0.22

5.5 Results and Discussion

We describe the utility of our approach through extensive use of examples. Figure 5.3, Figure 5.4, and Table 5.1 show results for different datasets and segmentation models. Qualitatively, our metric can quantify over- and under-segmentation issues (Figure 5.4). For instance, the example in the first row of Figure 5.4 shows a three-class task. The model prediction was almost perfect for this example, i.e., it correctly segmented each object. Our metrics correctly assigned $ROM = 0$ and $RUM = 0$, meaning there was no over- or under-segmentation in this example. However, all other metrics indicated that non-negligible errors occurred in this example. The fifth row of Figure 5.4 shows an example of over-segmentation for the *person* class. The torso of the leftmost person was over-segmented into two regions that were far apart, delivering a false prediction that two people were at the same location. A high value for ROM reflected this discrepancy. The third row of Figure 5.4 shows under-segmentation for the towel class, where accurate pixel classification and proper region segmentation disagreed. Pixel-wise measures (IoU, Dice score, and pixel error) indicated little error for the *towel* class. However, the number of model-predicted regions was only half the number of ground-truth regions. A high value for RUM reflected this disagreement.

Quantitatively, all models performed well on the PASCAL VOC dataset, as reflected by the low error scores in Table 5.1a. The light-weight models ESPNetv2 and MobileNetv2 had similar mIoU errors, but ROM shows that ESPNetv2 made fewer mistakes in terms of over-segmentation. The receiver operating characteristic (ROC) curves in Figure 5.3a indicate that all models made fewer mistakes when region confidence thresholds increased. At the threshold of 0.6, MobileNetv2 showed a reduction in over-segmentation error (achieving similarity to the other two models) without worsening under-segmentation error.

In the context of quotidian goals of weighing trade-offs between heavy- and light-weight models for particular segmentation tasks, we can make several observations about ROM/RUM. Table 5.1b compares the overall performance of different models on the Cityscapes validation set. Comparing the heavy-weight (DRN) and light-weight (ESPNetv2) models, we expected to see ESPNetv2 generally perform worse, which is reflected by an 8-point difference in IoU, Dice score, and pixel accuracy. In addition to these pixel-wise measures, our metrics help explain the source of performance discrepancy. DRN and ESPNetv2 had the same value for ROM but a significant difference of 5 points in RUM. This indicates that performance degradation in ESPNetv2 in this dataset mainly emanated from under-, not over-, segmentation. Similarly,

in Table 5.1c, we can explain that the lower performance of MobileNetv2 (a light-weight model) compared to PSPNet with ResNet-101 (a heavy-weight model) on the ADE20K dataset is primarily due to over-segmentation issues. A researcher may elect to tolerate the increase in under- or over-segmentation in exchange for computational efficiency, depending on downstream uses. Conversely, for navigation tasks, under-segmenting light poles may undercut downstream wayfinding and mapping uses, making it an undesirable outcome. Section 5.5.1 provides an in-depth analysis of using RUM/ROM for model selection.

5.5.1 Conjoint Use of RUM/ROM with mIoU to Inform Model Selection and Evaluation

As illustrated in Figure 5.4, region-based error metrics are not necessarily correlated with pixel-wise error metrics. As alternative sources of information about the model's performance in segmentation settings, and in particular, specific class performance, it is important to understand how pixel-based and region-based metrics might interact to better inform model selection or focus on improvements in any particular class.

In this section, we demonstrate our suggested metrics are orthogonal to traditional segmentation metrics and the manner in which the metrics can be used conjointly to inform model selection and evaluation. We integrate additional segmentation examples, with their class-wise score and RUM/ROM score. The examples are drawn from the ADE20K dataset [120] and the Cityscapes dataset [35]. The ADE20K dataset provides a variety of indoor/outdoor scenes with 150 object classes. The wide range of scenes in this dataset allows us to locate examples for different segmentation scenarios. The Cityscapes dataset contains urban scenes captured from a car dashboard's perspective, with 19 object classes commonly seen in the urban road environment. This dataset is important to tasks involving autonomous vehicles and outdoor environment mapping. The examples drawn from these two datasets are assorted into different categories as follows. Different metrics are represented with the same color code used in Figure 5.4.

Detailed Interrogation Into Low Pixel-wise Error and Region-based Error Metrics

Pixel-wise measures (IoU, dice score, pixel accuracy) are indeed informative when addressing questions surrounding pixel classification. However correlated these measures are, they are not surrogate metrics for near-perfect semantic segmentation. Importantly, pixel-wise measurements do not capture any region-wise information, nor indicate if there are any over-/under-segmentation issues in the prediction. The following

examples show that despite similar pixel-wise error metrics, the quality of semantic segmentation in the prediction will diverge.

Low ROM, Low RUM: Exploring the category of joint low pixel-wise errors, low over-segmentation error (ROM), and low under-segmentation error (RUM), we anticipate high model-prediction fidelity to the ground truth. Looking at these metrics conjointly, we expect that the model predicts a near-perfect segmentation. As an example, in Figure 5.5, an excellent prediction on a simple 2-class image with low pixel-wise error metrics, and no error along with ROM/RUM. Without the need to inspect the results by eye, we can expect few or no region-wise issues in this example and others like it.

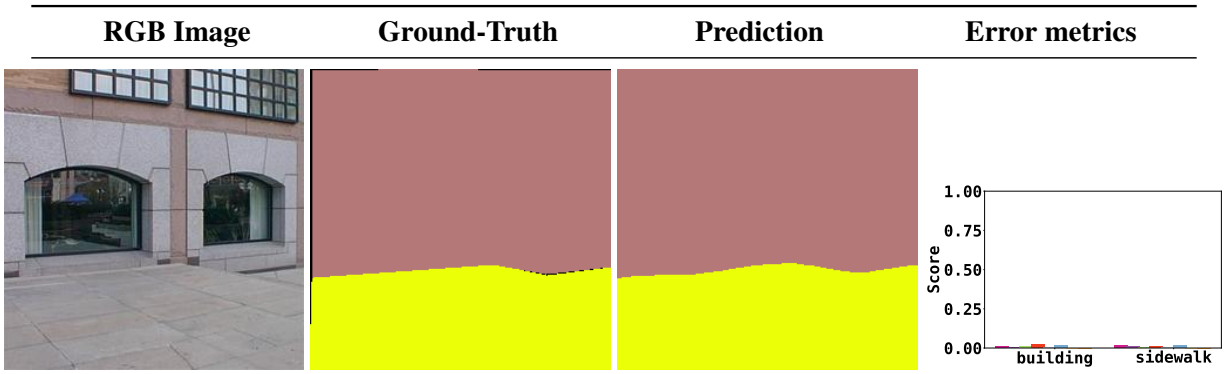


Figure 5.5: Low IoU Error, Low ROM/RUM

High RUM: Demonstrating that pixel-wise and region-wise metrics are not always correlated, it is important to note cases in which a result yields low pixel-wise errors, but high RUM, indicating that the number of regions predicted for a particular class by the model is less than the number of regions in the ground truth. As an example, in Figure 5.6, the model only predicts 2 regions of people out of 6 in the ground truth. The pixel distance between regions classified as people is small, and therefore region under-segmentation does not significantly impact pixel-wise error. Nevertheless, the region-wise discrepancy is reflected by the high RUM. By far, this is the most prevalent type of conjoint error we see in the Cityscapes dataset with the ESPNetv2 model predictions.

High ROM: High ROM in conjunction with low pixel-wise errors may also occur, indicative of model-predictions that demonstrate low per-pixel classification for an object class, despite segmenting more than

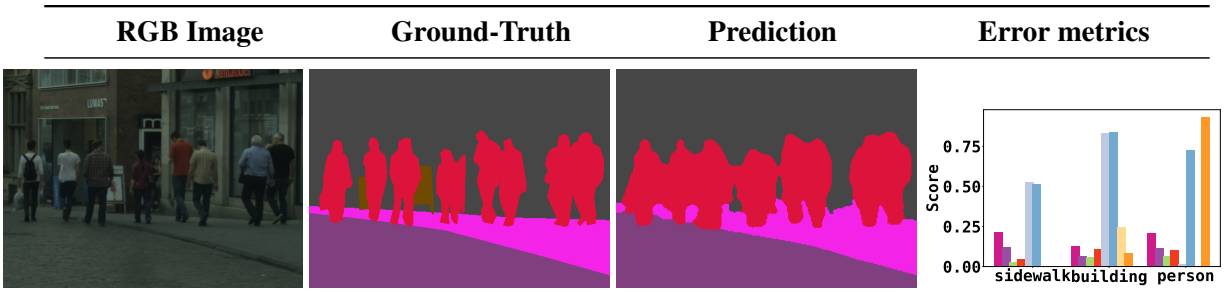


Figure 5.6: Low IoU Error, High RUM (person)

one distinct region per single class object in the ground truth. For example in Figure 5.7, the chair legs are segmented into multiple separate regions, creating over-segmentation issues and delivering false inference that there are multiple chairs at the same location. Note that this is an interesting example of the interplay between different class segmentation: when one chair is over-segmented in multiple disconnected regions, the result impacts another class' region segmentation, i.e., the chair background is then connected into one large region for the wall class. The wall is then under-segmented as a result of the chair over-segmentation. This demonstrates how models might be interrogated for intricate nuanced relationships among classes. In this case, looking at the entire model (ResNet101+PSPNet [101]) performance on this dataset (ADE20k), the overall mIoU error is 0.58, whereas for the chair class (mIoU error is 0.54, RUM is 0.09, and ROM is 0.14), interplayed with wall class (mIoU error is 0.32, RUM is 0.31, and ROM is 0.10).

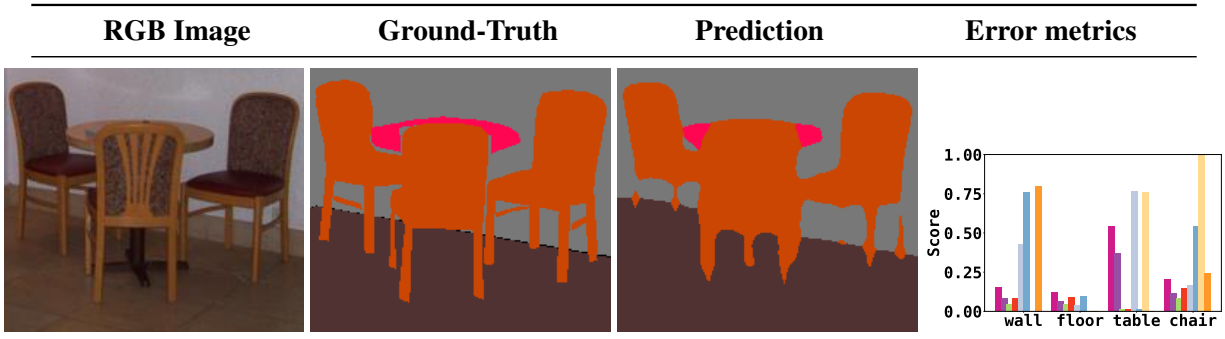


Figure 5.7: Low IoU Error, High ROM (chair)

Detailed Interrogation Into High Pixel-wise Error and Region-based Error Metrics

There is no question that significant pixel-wise classification errors cannot yield very good segmentation. However, the examples we looked at are encouraging in terms of the potential uses of light-weight models (with occasionally higher pixel-wise error) in certain semantic segmentation scenarios.

We looked at images that tended to have high pixel-wise errors in conjunction with ROM/RUM. Identifying images with specific region-wise metric attributes can provide insight on region-wise segmentation qualities and offer clues about the types of images or scenes in which pixel-wise predictions may fail.

Low ROM, Low RUM: In this class of images, the predicted pixel classification may create false-negative/false-positive regions that degrade pixel-wise accuracy. However, those classes yielding regions that are correctly predicted by the model correspond well with ground-truth, and are reflected by the low ROM/RUM metrics. For example, in Figure 5.8, although the model fails to predict some persons and poles, and consequently makes few false-positive predictions, those regions that are predicted correctly are matched to the corresponding objects in the ground truth. As no over-/under-segmentation occurred in those particular predicted regions, they are assigned a 0 value ROM/RUM. Again, this notion can be used to interrogate model predictions in a more nuanced manner to understand how the interplay between different classes (with potentially different backgrounds) may account for gross pixel-wise classification errors with certain models.

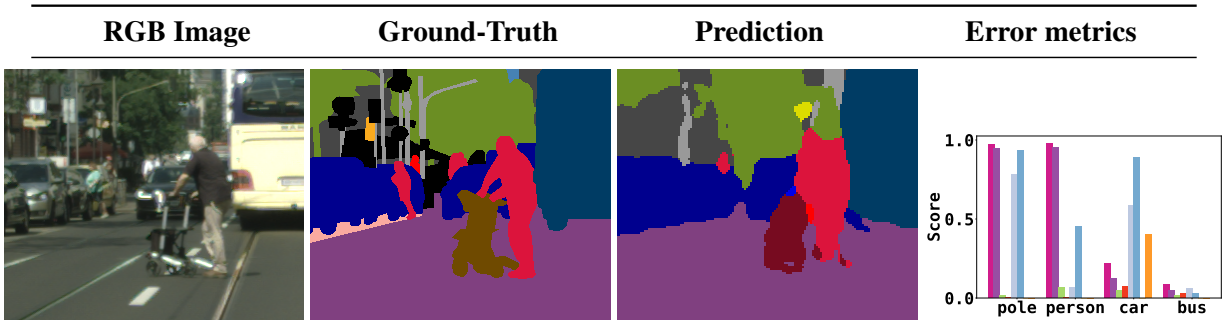


Figure 5.8: High IoU Error, Low RUM/ROM (person, pole)

High RUM: In this category, the pixel-wise accuracy errors are high, and a high RUM value indicates that these errors mainly come from under-segmentation. As shown in Figure 5.9, the bottles on the shelf are falsely grouped into one large object. Meanwhile, the pixels in between each bottle are falsely predicted

into the bottle class which impairs the overall pixel-wise accuracy.

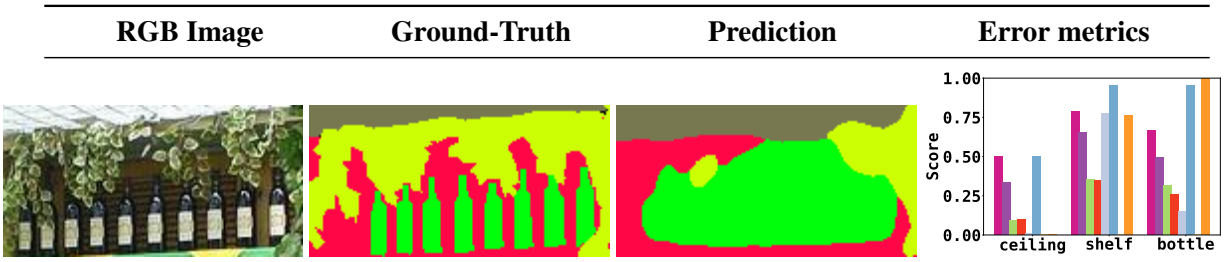


Figure 5.9: High IoU Error, High RUM (bottle)

High ROM: In this category, the pixel-wise accuracy errors are high, and a high ROM value explains that these errors occur when a single object in the ground truth is subdivided into several regions in the prediction. For example, in Figure 5.10, a chair is falsely segmented into multiple far-apart regions. This is concurrent with the majority of pixels in the middle of the chair being predicted as other classes, resulting in low pixel-wise accuracy.

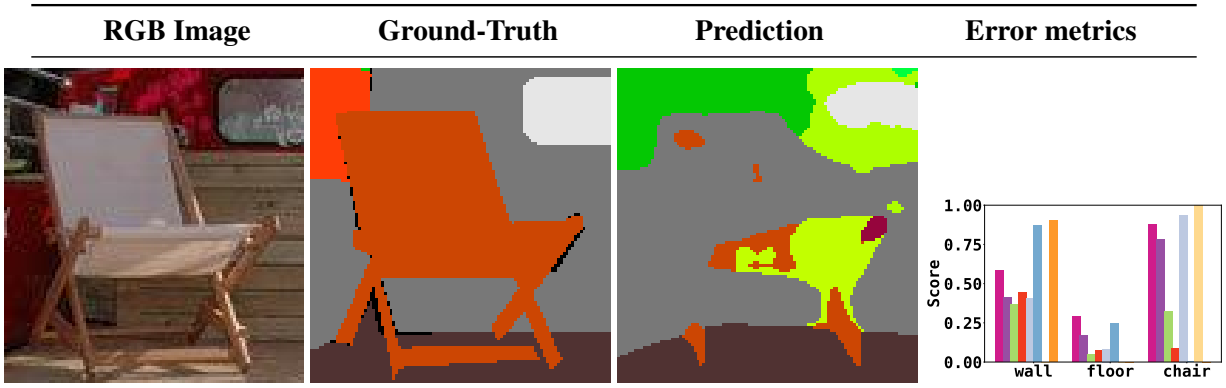


Figure 5.10: High IoU Error, High ROM (chair)

High ROM, High RUM: Rarely, there are cases where both ROM and RUM are high, e.g. the plant class in Figure 5.11. This happens when one or more regions in the ground truth are over-segmented, while there are other regions that are under-segmented.

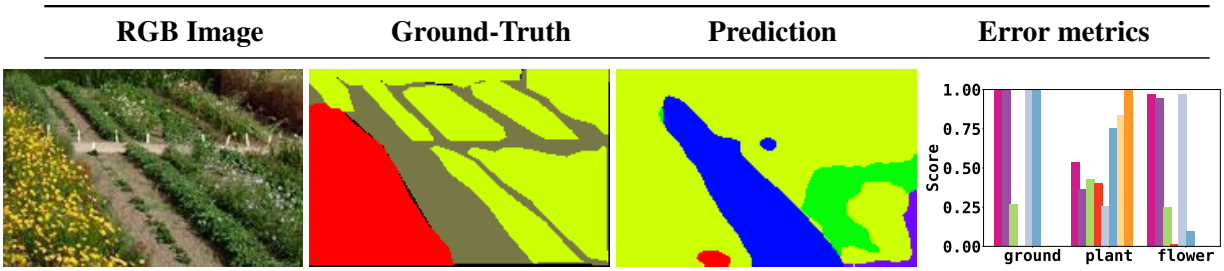


Figure 5.11: High IoU Error, High ROM/RUM (plant)

5.5.2 ROM/RUM for Greater Model Explainability

From the examples shown in Section 5.5.1, we demonstrate that ROM and RUM have the following attributes and advantages.

ROM/RUM accounts for model disagreement with ground-truth in over- and under-segmentation cases. Revealing over-/under-segmentation issues that are not reflected in other metrics. For example, as shown in Figure 5.6, the prediction may have high pixel-wise accuracy, while suffering from severe under-segmentation issues. In downstream applications where under-segmentation is the primary concern, RUM needs to be considered together with other pixel-wise metrics for model selection.

ROM/RUM can differentiate among different degrees of over-segmentation and under-segmentation issues. This is important because it allows researchers to quantify the severity of the over-/under-segmentation issues in a particular model as it pertains to a specific class or classes of objects. For example, in Figure 5.7, both the table class and the chair class are over-segmented, but the chair class receives a higher ROM because it has more over-segmented regions per ground truth region.

ROM/RUM offers additional information for evaluating and selecting a segmentation model alongside the pixel-wise metrics. When pixel-wise accuracy is high, ROM/RUM can assist in validating the model prediction in region-wise performance. All examples in Figures 5.5, 5.6, and 5.7 have relatively low pixel-wise errors for certain object classes, but the pixel-wise metrics alone do not lend for a proper nor complete interpretation of the model's region-wise performance. As illustrated in Figure 5.5, a prediction should receive 0 values for pixel-wise error, ROM, and RUM if and only if it predicts perfect region-wise segmentation. A non-zero value of ROM/RUM indicates there are over-/under-segmentation issues within the prediction even if the pixel-wise errors are low.

When pixel-wise accuracy is low, ROM/RUM can explain the source of performance degradation and provide useful information for model selection. As illustrated in Figure 5.9 and Figure 5.10, the pixel-wise accuracy of predictions can be penalized to a similar degree while creating contrasting region-wise segmentation qualities – one may have major under-segmentation issues, while the other creates over-segmentation issues.

Overall, we demonstrated certain common model interrogation scenarios in which combining ROM/RUM with other pixel-wise measures allows researchers to select the most appropriate model for specific datasets or applications. Furthermore, interrogating models while using metrics conjointly may give rise to a more nuanced understanding of model performance for certain classes alone, or for certain models in multi-class correlated scenarios. In future research, we intend to explore how ROM/RUM can be used during the learning process of a model, in order to tailor a model for a specific use case. We intend to further study the interrelation between ROM and RUM, and combined ROM/RUM with pixel-wise metrics applied in multi-class correlated settings. A principled approach in this direction can allow for comprehensive, simultaneous evaluation and interpretation of model performance in semantic segmentation.

5.6 Discussion

This chapter reviewed measures of similarity popular in computer vision for assessing semantic segmentation fidelity to a ground-truth and discussed their shortcomings in accounting for over- and under-segmentation. While IoU and similar evaluation metrics can be applied to measure pixel-wise correspondence between both model-predicted segmentation and ground-truth, the evaluation can suffer from inconsistencies due to different region boundaries and notions of significance with respect to a particular class of objects. We proposed an approach to segmentation metrics that decomposes to explainable terms and is sensitive to over- and under-segmentation errors. This new approach confers additional desirable properties, like robustness to boundary changes (since both annotations and natural-image semantic boundaries are non-deterministic). We contextualized the application of our metrics in current model selection problems that arise in practice when attempting to match the context of use to region-based segmentation performance in supervised datasets.

Chapter 6

Conclusion

Pedestrian paths are key to a healthy transportation network and building an accessible city, however the pedestrian paths and the street-side environments that serve pedestrians have not been widely mapped. Most cities do not have a geographically accurate map that shows where sidewalks and footpaths are, or how they are connected. To gather this information, typical mapping methods mainly rely on human annotations, including paratransit pathway review team surveyor's on-site collections. These methods are non-standardized, laborious, costly, unscalable, and difficult to keep current.

In this comprehensive work, we made significant contributions to the research and development of automated mapping and assessment of pedestrian path networks, urban sidewalk assessment, and model selection in practice.

The introduction of the APE dataset provides researchers, communities, and industries with an open and shared tool for addressing pedestrian path network mapping through scalable machine learning methods. The APE dataset contains images and annotations from diverse built environments and is applicable to a variety of computer vision tasks, particularly in the context of urban planning and wayfinding. By offering aerial images, street map tiles, and annotation in pedestrian environments, the APE dataset presents an essential resource for researchers to work on the challenging task of developing pedestrian path network mapping methods. Following the introduction of APE, we also present Prophet, an end-to-end process to infer connected pedestrian path networks. Prophet uses a multi-input segmentation network trained on the APE dataset to predict pedestrian path locations and integrate them with existing street network data to

complete a pedestrian path network.

The development of OASIS, a novel urban sidewalk mapping and assessment system using computer vision techniques on portable edge devices, offers many benefits. With the ability to map sidewalks and assess their connectivity quickly and accurately, OASIS provides standardized data at scale, promoting an open and shared data ecosystem for various stakeholders, including transportation agencies, municipalities, and individual travelers. The edge computing approach addresses privacy concerns and enables new applications for autonomous navigation in sidewalk environments, such as autonomous wheelchairs and self-driving delivery robots.

Furthermore, this research highlights the importance of adopting a new approach to segmentation metrics that accounts for over- and under-segmentation errors for model performance explainability and model selection in practice. The novel metrics ROM/RUM decomposes into explainable terms, making it easier for researchers to understand and address segmentation inconsistencies that may arise due to over- and under-segmentation with respect to a particular class of objects. With ROM/RUM, researchers can select the most appropriate model for specific datasets or applications and gain an understanding of model performance in various contexts.

In conclusion, this work represents an advancement in the field of pedestrian environment mapping research. The outcome of our approaches improves non-motorized transportation assessment, leads to a more accessible and sustainable urban environment, and is vital for the development of smart city technologies that address the needs of individuals with different mobility levels. The open and shared output mapping data benefits individual travelers through downstream use in routing and wayfinding applications. Altogether, this work contributes to the important yet under-addressed task of building walkable and accessible urban cities, and establishes a benchmark for future studies in mapping and assessing pedestrian environments.

Bibliography

[1] N. Bolten and A. Caspi, "Towards operationalizing the communal production and management of public (open) data: a pedestrian network case study: A pedestrian network case study in operationalizing communal open data," in *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, 2022, pp. 232–247.

[2] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3438–3446.

[3] L. Mi, H. Zhao, C. Nash, X. Jin, J. Gao, C. Sun, C. Schmid, N. Shavit, Y. Chai, and D. Anguelov, "Hdmapgen: A hierarchical graph generative model of high definition maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4227–4236.

[4] Z. Guo, Y. Huang, X. Hu, H. Wei, and B. Zhao, "A survey on deep learning based approaches for scene understanding in autonomous driving," *Electronics*, vol. 10, no. 4, p. 471, 2021.

[5] N. Lefler, Y. Zhou, D. Carter, H. W. McGee, D. L. Harkey, F. M. Council *et al.*, "Model inventory of roadway elements–mire 2.0," United States. Federal Highway Administration. Office of Safety, Tech. Rep., 2017.

[6] N. Bolten, S. Mukherjee, V. Sipeeva, A. Tanweer, and A. Caspi, "A pedestrian-centered data approach for equitable access to urban infrastructure environments," *IBM Journal of Research and Development*, vol. 61, no. 6, pp. 10–1, 2017.

[7] J. E. Froehlich, A. M. Brock, A. Caspi, J. Guerreiro, K. Hara, R. Kirkham, J. Schöning, and B. Tannert, "Grand challenges in accessible maps," *interactions*, vol. 26, no. 2, pp. 78–81, 2019.

[8] B. W. Landis, T. A. Petritsch, H. F. Huang, and A. H. Do, "Characteristics of emerging road and trail users and their safety," *Transportation research record*, vol. 1878, no. 1, pp. 131–139, 2004.

[9] K. J. Clifton, A. D. L. Smith, and D. Rodriguez, "The development and testing of an audit for the pedestrian environment," *Landscape and urban planning*, vol. 80, no. 1-2, pp. 95–110, 2007.

[10] J. F. Sallis, K. L. Cain, T. L. Conway, K. A. Gavand, R. A. Millstein, C. M. Geremia, L. D. Frank, B. E. Saelens, K. Glanz, and A. C. King, "Peer reviewed: Is your neighborhood designed to support physical activity? a brief streetscape audit tool," *Preventing chronic disease*, vol. 12, 2015.

[11] N. Bolten and A. Caspi, "The opensidewalks schema," 2022. [Online]. Available: https://github.com/OpenSidewalks/OpenSidewalks-Schema

[12] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[13] X. Liu and Y. Long, "Automated identification and characterization of parcels with openstreetmap and points of interest," *Environment and Planning B: Planning and Design*, vol. 43, no. 2, pp. 341–360, 2016.

[14] H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on openstreetmap," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 700–719, 2014.

[15] P. Mooney, M. Minghini *et al.*, "A review of openstreetmap data," *Mapping and the citizen sensor*, pp. 37–59, 2017.

[16] A. Ballatore and A. Zipf, "A conceptual quality framework for volunteered geographic information," in *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings 12*. Springer, 2015, pp. 89–107.

[17] M. A. Brovelli, M. Minghini, and G. Zamboni, "Public participation gis: a foss architecture enabling field-data collection," *International Journal of Digital Earth*, vol. 8, no. 5, pp. 345–363, 2015.

[18] S. Wu, C. Du, H. Chen, Y. Xu, N. Guo, and N. Jing, "Road extraction from very high resolution images using weakly labeled openstreetmap centerline," *ISPRS International Journal of Geo-Information*, vol. 8, no. 11, p. 478, 2019.

[19] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, "Leveraging crowdsourced gps data for road extraction from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7509–7518.

[20] K. Zhou, Y. Xie, Z. Gao, F. Miao, and L. Zhang, "Funet: A novel road extraction network with fusion of location data and remote sensing imagery," *ISPRS International Journal of Geo-Information*, vol. 10, no. 1, p. 39, 2021.

[21] X. Lu, Y. Zhong, Z. Zheng, and L. Zhang, "Gamsnet: Globally aware road detection network with multi-scale residual learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 340–352, 2021.

[22] D. Pan, M. Zhang, and B. Zhang, "A generic fcn-based approach for the road-network extraction from vhr remote sensing images–using openstreetmap as benchmarks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2662–2673, 2021.

[23] R. Ewing and R. Cervero, "Travel and the built environment: A meta-analysis," *Journal of the American planning association*, vol. 76, no. 3, pp. 265–294, 2010.

[24] R. Cervero and K. Kockelman, "Travel demand and the 3ds: Density, diversity, and design," *Transportation research part D: Transport and environment*, vol. 2, no. 3, pp. 199–219, 1997.

[25] S. Handy, X. Cao, and P. Mokhtarian, "Correlation or causality between the built environment and travel behavior? evidence from northern california," *Transportation Research Part D: Transport and Environment*, vol. 10, no. 6, pp. 427–444, 2005.

[26] H. Li, J. Cebe, S. Khoeini, Y. Xu, C. Dyess, and R. Guensler, "A semi-automated method to generate gis-based sidewalk networks for asset management and pedestrian accessibility assessment," *Transportation research record*, vol. 2672, no. 44, pp. 1–9, 2018.

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

[28] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: the international journal for geographic information and geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

[29] J. W. Brandt and V. R. Algazi, "Continuous skeleton computation by voronoi diagram," *CVGIP: Image understanding*, vol. 55, no. 3, pp. 329–338, 1992.

[30] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, "Torontocity: Seeing the world with a million eyes," *arXiv preprint arXiv:1612.00423*, 2016.

[31] S. Aksoy, B. Ozdemir, S. Eckert, F. Kayitakire, M. Pesarasi, O. Aytekin, C. C. Borel, J. Cech, E. Christophe, S. Duzgun *et al.*, "Performance evaluation of building detection and digital surface model extraction algorithms: Outcomes of the prrs 2008 algorithm performance contest," in *2008 IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2008)*. IEEE, 2008, pp. 1–12.

[32] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[33] "Isprs test project on urban classification, 3d building reconstruction and semantic labeling," 2022. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx

[34] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

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

[36] P. Mirowski, A. Banki-Horvath, K. Anderson, D. Teplyashin, K. M. Hermann, M. Malinowski, M. K. Grimes, K. Simonyan, K. Kavukcuoglu, A. Zisserman *et al.*, "The streetlearn environment and dataset," *arXiv preprint arXiv:1903.01292*, 2019.

[37] M. Haklay, "How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets," *Environment and planning B: Planning and design*, vol. 37, no. 4, pp. 682–703, 2010.

[38] P. Neis, D. Zielstra, and A. Zipf, "Comparison of volunteered geographic information data contributions and community development for selected world regions," *Future internet*, vol. 5, no. 2, pp. 282–300, 2013.

[39] J.-F. Girres and G. Touya, "Quality assessment of the french openstreetmap dataset," *Transactions in GIS*, vol. 14, no. 4, pp. 435–459, 2010.

[40] D. Zielstra and H. H. Hochmair, "Comparative study of pedestrian accessibility to transit stations using free and proprietary network data," *Transportation Research Record*, vol. 2217, no. 1, pp. 145–152, 2011.

[41] N. Bolten and A. Caspi, "Accessmap website demonstration: Individualized, accessible pedestrian trip planning at scale," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 676–678.

[42] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1013–1020.

[43] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich *et al.*, "Speeding up semantic segmentation for autonomous driving," in *MLITS, NIPS Workshop*, vol. 2, no. 7, 2016.

[44] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, pp. 1–13, 2022.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

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

[47] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[48] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[50] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

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

[52] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[53] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 297–312.

[54] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3992–4000.

[55] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3150–3158.

[56] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training," in *European Conference on Computer Vision*. Springer, 2016, pp. 381–397.

[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[58] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[59] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[60] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9167–9176.

[61] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[62] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[63] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[64] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[65] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[66] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9190–9200.

[67] H. A. Karimi and P. Kasemsuppakorn, "Pedestrian network map generation approaches and recommendation," *International Journal of Geographical Information Science*, vol. 27, no. 5, pp. 947–962, 2013.

[68] D. Ahmetovic, R. Manduchi, J. M. Coughlan, and S. Mascetti, "Zebra crossing spotter: Automatic population of spatial databases for increased safety of blind travelers," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, 2015, pp. 251–258.

[69] M. C. Ghilardi, J. C. Jacques, and I. Manssour, "Crosswalk localization from low resolution satellite images to assist visually impaired people," *IEEE computer graphics and applications*, vol. 38, no. 1, pp. 30–46, 2016.

[70] H. Ning, X. Ye, Z. Chen, T. Liu, and T. Cao, "Sidewalk extraction using aerial and street view images," *Environment and Planning B: Urban Analytics and City Science*, vol. 49, no. 1, pp. 7–22, 2022.

[71] Q. Hou and C. Ai, "A network-level sidewalk inventory method using mobile lidar and deep learning," *Transportation research part C: emerging technologies*, vol. 119, p. 102772, 2020.

[72] K. Manaugh and A. El-Geneidy, "Validating walkability indices: How do different households respond to the walkability of their neighborhood?" *Transportation research part D: transport and environment*, vol. 16, no. 4, pp. 309–315, 2011.

[73] V. Mehta, "Walkable streets: pedestrian behavior, perceptions and attitudes," *Journal of urbanism*, vol. 1, no. 3, pp. 217–245, 2008.

[74] N. Bolten and A. Caspi, "Towards routine, city-scale accessibility metrics: Graph theoretic interpretations of pedestrian access using personalized pedestrian network analysis," *PLoS one*, vol. 16, no. 3, p. e0248399, 2021.

[75] "Opensidewalks," 2021. [Online]. Available: https://www.opensidewalks.com/

[76] O. L. Sarmiento, A. F. Useche, D. A. Rodriguez, I. Dronova, O. Guaje, F. Montes, I. Stankov, M. A. Wilches, U. Bilal, X. Wang *et al.*, "Built environment profiles for latin american urban settings: The salurbal study," *PLoS One*, vol. 16, no. 10, p. e0257528, 2021.

[77] N. Bolten and A. Caspi, "The opensidewalks schema," 2021. [Online]. Available: https://github.com/opensidewalks/OpenSidewalks-Schema

[78] "Los angeles geohub sidewalks (mapped areas)," 2022. [Online]. Available: https://geohub.lacity.org/maps/lahub::sidewalks-mapped-areas/about

[79] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[80] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," *arXiv preprint arXiv:1902.11097*, 2019.

[81] E. S. Jo and T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 306–316.

[82] J. C. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," *IEEE Transactions on aerospace and electronic systems*, vol. 34, no. 3, pp. 817–823, 1998.

[83] P.-A. Quinones, T. Greene, R. Yang, and M. Newman, "Supporting visually impaired navigation: a needs-finding study," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 1645–1650.

[84] S. Zimmermann-Janschitz, S. Landauer, S. Drexel, and J. Obermeier, "Independent mobility for persons with vib using gis," *Journal of Enabling Technologies*, 2021.

[85] S. J. Bosch and A. Gharaveis, "Flying solo: A review of the literature on wayfinding for older adults experiencing visual or cognitive decline," *Applied ergonomics*, vol. 58, pp. 327–333, 2017.

[86] U.S. Code, "Americans with disabilities act," 1990.

[87] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1743–1751.

[88] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.

[89] Y. Eisenberg, A. Heider, R. Gould, and R. Jones, "Are communities in the united states planning for pedestrians with disabilities? findings from a systematic evaluation of local government barrier removal plans," *Cities*, vol. 102, p. 102720, 2020.

[90] Y. Eisenberg, A. Hofstra, S. Berquist, R. Gould, and R. Jones, "Barrier-removal plans and pedestrian infrastructure equity for people with disabilities," *Transportation research part D: transport and environment*, vol. 109, p. 103356, 2022.

[91] S. Gallo, D. Chapuis, L. Santos-Carreras, Y. Kim, P. Retornaz, H. Bleuler, and R. Gassert, "Augmented white cane with multimodal haptic feedback," in *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*. IEEE, 2010, pp. 149–155.

[92] R. Pyun, Y. Kim, P. Wespe, R. Gassert, and S. Schneller, "Advanced augmented white cane with obstacle height and distance feedback," in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2013, pp. 1–6.

[93] United States Government Accountability Office, "Transportation accessibility: Lack of data and limited enforcement options limit federal oversigh," *United States Government Accountability Office*, 2007.

[94] J. Pearlman, E. Sinagra, J. Duvall, R. Cooper, D. Stuckey, and A. Kortum, "Development and characterization of pathway measurement tool (pathmet)," Transportation Research Board, Tech. Rep., 2014.

[95] Mass Transit, "King county metro wins national award for improving data on sidewalks - commonpaths provides tools for agencies to collect and distribute high-fidelity pathway data in opensidewalks," *Mass Transit Best Practices For Integrated Mobility*, 2021. [Online]. Available: https://www.masstransitmag.com/safety-security/press-release/21247636/

[96] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[97] N. Bolten and A. Caspi, "Accessmap," 2021. [Online]. Available: accessmap.io

[98] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[99] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[100] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.

[101] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[102] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[103] S. Mehta, F. Abdolhosseini, and M. Rastegari, "Cvnets: High performance library for computer vision," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7327–7330.

[104] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[105] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[106] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[107] S. Mehta, H. Hajishirzi, and L. Shapiro, "Identifying most walkable direction for navigation in an outdoor environment," *arXiv preprint arXiv:1711.08040*, 2017.

[108] K. Yang, K. Wang, L. M. Bergasa, E. Romera, W. Hu, D. Sun, J. Sun, R. Cheng, T. Chen, and E. López, "Unifying terrain awareness for the visually impaired through real-time semantic segmentation," *Sensors*, vol. 18, no. 5, p. 1506, 2018.

[109] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 552–568.

[110] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[111] W. Shimoda and K. Yanai, "Weakly supervised semantic segmentation using distinct class specific saliency maps," in *Computer Vision and Image Understanding*, vol. 191, 2018.

[112] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[113] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[114] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, p. 29, 2015.

[115] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2. IEEE, 2001, pp. 416–423.

[116] M. Polak, H. Zhang, and M. Pi, "An evaluation metric for image segmentation of multiple objects," *Image and Vision Computing*, vol. 27, no. 8, pp. 1223–1227, 2009.

[117] C. Persello and L. Bruzzone, "A novel protocol for accuracy assessment in classification of very high resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1232–1244, 2009.

[118] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[119] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[120] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[121] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

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