Image Fusion for Misaligned Visual and Thermal Images

Aadhar Chauhan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2023

Committee:
Karen Leung
Juris Vagners
Xu Chen
Santosh Devasia

Program Authorized to Offer Degree:
Mechanical Engineering

Universiry of Washington

Abstract

Image Fusion for Misaligned Visual and Thermal Images

Aadhar Chauhan

Chair of the Supervisory Committee:
Karen Leung
Department of Aeronautics and Astronautics

The main challenge of a ***Wilderness Search and Rescue (WiSAR)*** mission is to cover a large area within a reasonable and short amount of time effectively. The recent development in the area of sensor-equipped Unmanned Aerial Vehicles (UAV) and object detection algorithms have pushed the idea of using UAVs equipped with appropriate sensors for the surveillance of the area and finding humans in an efficient and time-saving way. However, visual and thermal cameras used for surveillance can produce images with misaligned features due to varying viewpoints and sensor characteristics, leading to difficulties in accurate detection. The present work discusses a novel deep learning approach using a Generative Adversarial Network (GAN) with two Discriminators and an attention mechanism used in the Generator, which enhances the ability of the network to capture more complex features in the input images. The proposed solution addresses the challenges of diverse terrain and weather conditions by fusing visual and thermal images to produce a single composite image with complementary information. To address the misalignment issue, we propose a novel attention-based Generator that selectively extracts features from the visual and thermal images based on their similarities, while maintaining their unique characteristics. The results obtained further motivate the idea of developing an end-to-end pipeline for human detection in a wilderness environment using both, visual and thermal images as inputs.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Karen Leung, for her unwavering support, invaluable guidance, and endless encouragement throughout the entire journey of this thesis. Her expertise, insightful feedback, and mentorship have been instrumental in shaping the research direction and pushing me to strive for excellence. I am also immensely grateful to my co-advisor, Dr. Juris Vagners, for his valuable contributions, insightful discussions, and dedication to my academic and research growth. His expertise and encouragement have been pivotal in broadening my perspectives and enriching the quality of this work.

I extend my heartfelt appreciation to the members of my thesis committee for their time, expertise, and constructive feedback during the evaluation and refinement of this research. Their insights have greatly strengthened the overall quality of this thesis.

I am indebted to my lab mates and Ph.D. students, Danny and Chris, for their collaboration, camaraderie, and intellectual discussions, which have been an endless source of inspiration and motivation. Their support has made the research journey more rewarding and enjoyable. I am grateful to my lab peers for creating a vibrant and stimulating research environment. Our collective efforts and collaborations have been instrumental in propelling my research forward and fostering a sense of community within the lab.

I would like to extend a special thank you to my friends, whose unwavering support and understanding have been a constant source of encouragement during challenging times. Their presence in my life has made this academic journey more meaningful and enjoyable.

Lastly, I want to express my deepest gratitude to my parents and my brother. Their love, encouragement, and sacrifices have been the bedrock of my academic pursuits. Their unwavering belief in me has given me the strength to overcome obstacles and reach new heights in my academic journey.

To all the individuals mentioned above and to those who have supported me along the way, I offer my heartfelt thanks. This thesis would not have been possible without your support, and I am truly grateful for the impact you have had on my academic and personal growth.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

1.1 Autonomous Systems and Sensor Modalities

Autonomous systems have emerged as a transformative technology with the potential to revolutionize various industries, including transportation, manufacturing, agriculture, and healthcare. These systems, also known as self-driving or autonomous robotic systems, are designed to operate and make decisions without direct human intervention, relying on a combination of advanced algorithms, artificial intelligence, and sensor technologies.

At the core of autonomous systems lie sensors, which serve as the sensory organs, enabling these systems to perceive and interact with their environment. These sensors play a crucial role in providing real-time data about the system's surroundings, facilitating accurate perception, and enabling informed decision-making. Among the diverse range of sensors employed in autonomous systems, some of the key modalities include visual cameras, thermal cameras, Radio Detection and Ranging (RADAR), and Light Detection and Ranging (LiDAR).

Visual Cameras

Visual cameras similar to those used in everyday photography, capture images of the system's surroundings. They rely on visible light to provide a detailed representation of objects and their spatial relationships. Visual cameras are generally used as a stereo pair, consisting of synchronized cameras that mimic human binocular vision, enabling depth perception and facilitating the 3-D reconstruction of the environment.

Thermal Cameras

Thermal cameras operate based on the detection of infrared radiation emitted by objects. They excel in detecting heat signatures, making them valuable for applications such as night vision, detecting living beings in obscured environments, and monitoring thermal patterns.

Radio Detection and Ranging (RADAR)

RADAR employs radio waves to detect and track objects in the system's surroundings. By measuring the time taken for radio waves to reflect after hitting objects,

RADAR can provide information about the distance, velocity, and sometimes even the shape of objects.

Light Detection and Ranging (LiDAR)

LiDAR, on the other hand, utilizes laser beams to measure the distance to objects and create detailed 3D maps of the environment. By analyzing the reflections of laser beams, LiDAR can generate accurate and precise point cloud data, allowing for precise object detection and localization.

1.2 Importance of Sensor Fusion

In the realm of autonomous systems, compensating for the limitations of individual sensors holds immense significance. Each sensor modality possesses inherent constraints (Figure 1.1) that can hinder the system's performance and compromise its ability to operate effectively in diverse environments and scenarios. By leveraging sensor fusion techniques, these limitations can be overcome, allowing autonomous systems to achieve enhanced capabilities and improved performance.

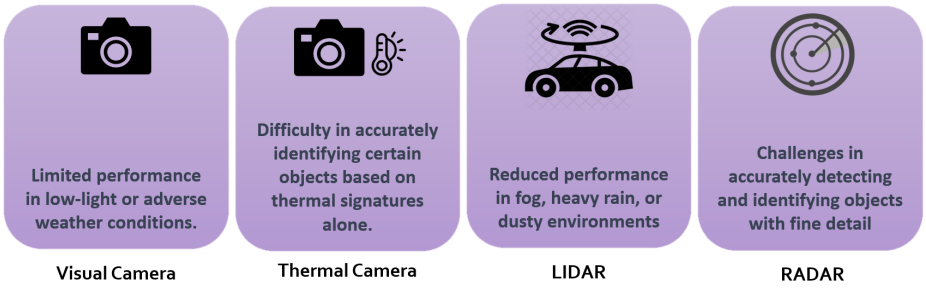

Figure 1.1: Various sensors with inherent limitations drive the necessity for sensor fusion in autonomous systems.

Compensation for sensor limitations becomes crucial when considering the vast range of environmental conditions that autonomous systems may encounter such as in the case of wilderness search and rescue. For instance, visual cameras heavily rely on sufficient lighting conditions, making them less reliable in low-light or dark environments. In such situations, the fusion of visual camera data with other sensor modalities, such as thermal cameras or LiDAR, can provide valuable supplementary information that compensates for the lack of visual clarity. Similarly, thermal cameras face challenges during thermal crossover situations [18], where the ambient temperature closely matches that of living beings of interest. Integrating thermal camera data with other sensors can help differentiate between objects based on their

thermal signatures, thus overcoming limitations imposed by thermal crossover.

sensor fusion enhances the robustness and reliability of autonomous systems. By incorporating redundant sources of information, the system becomes less susceptible to failures or inaccuracies in individual sensors. If one sensor malfunctions or encounters limitations in a particular scenario, the other sensor modalities can compensate for the shortfall, ensuring continuous and reliable operation. This redundancy and fault tolerance are particularly critical in safety-critical applications where reliable perception is vital.

1.3 Image Fusion and Its Applications

The concept of image fusion involves integrating images captured by various sensors, such as visual cameras, thermal cameras, or multispectral cameras, to leverage the strengths of each modality and compensate for their limitations. By fusing complementary data from different sensor modalities, image fusion aims to produce a composite image that enhances the overall perception and understanding of the scene.

Image fusion techniques utilize algorithms that analyze the input images, extract relevant features or information, and combine them to generate the fused image. These algorithms can operate at different levels, including pixel-level fusion, feature-level fusion, or decision-level fusion, depending on the specific requirements and characteristics of the application [12].

Practical Applications of Image Fusion

Image fusion finds numerous practical applications in various domains [16], benefiting from the enhanced information provided by the fusion process. Some notable applications include:

- **Remote Sensing and Earth Observation**

  Image fusion is extensively used in satellite imagery and aerial imaging applications. By fusing data from different sensors, such as visible, infrared, panchromatic, and multispectral sensors, it becomes possible to obtain comprehensive and multi-dimensional information about the Earth's surface, enabling tasks such as land cover classification, environmental monitoring, and disaster management [23]. An illustration of image fusion for remote sensing application is depicted in Figure 1.2, where the fusion of 'Panchromatic' and 'Multispectral' images is demonstrated.

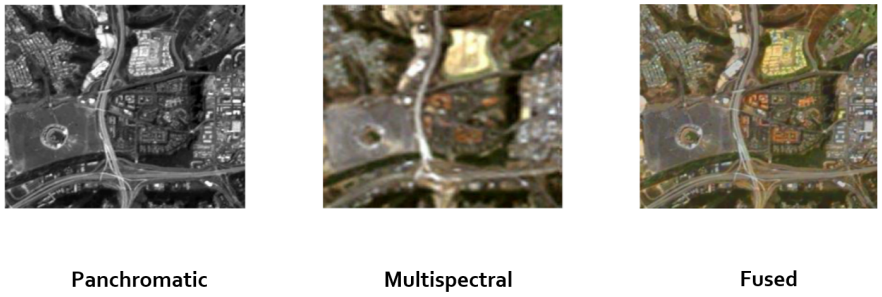

Figure 1.2: In the provided illustration, we observe the fusion of a panchromatic image (on the left) and a multispectral image (in the middle), resulting in the fused image (on the right).

- **Medical Imaging**

  Image fusion plays a crucial role in medical imaging applications, where it combines information from multiple imaging modalities, such as MRI Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), and Ultrasound. This fusion enhances the accuracy of diagnosis improves image-guided interventions and facilitates more precise localization and characterization of diseases [2, 24]. An illustration of image fusion in the field of medical imaging applications is depicted in Figure 1.3, where the fusion of Magnetic Resonance Imaging (MRI) and Computed Tomography scan (CT-scan) images is demonstrated.

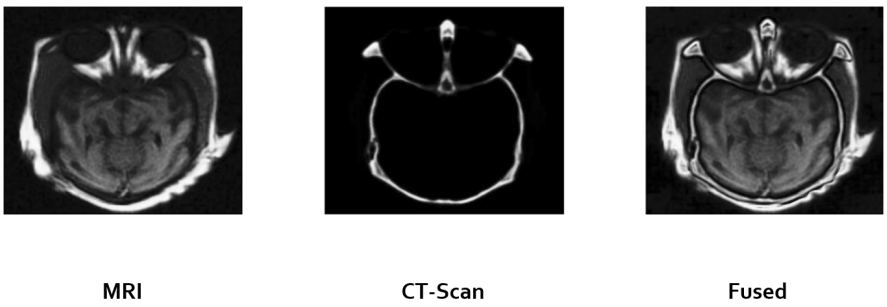

Figure 1.3: In the provided illustration, we observe the fusion of an MRI (on the left) and an MS image (in the middle), resulting in the fused image (on the right).

- **Autonomous Systems**

  Image fusion is particularly relevant to autonomous systems, such as self-driving cars or uncrewed aerial vehicles (UAVs). By fusing visual and other sensor data, such as LiDAR or RADAR, autonomous systems can enhance

their perception capabilities, improve object detection and recognition, and enable robust decision-making in complex and dynamic environments. An illustration of image fusion in the field of medical imaging applications is depicted in Figure 1.4, where the fusion of Visual and Thermal images is demonstrated.

Figure 1.4: In the provided illustration, we observe the fusion of a visual image (on the left) and a thermal image (in the middle), resulting in the fused image (on the right).

In summary, image fusion plays a vital role in autonomous systems by enabling a more comprehensive and accurate perception of the environment. By combining images from different sensors, it enhances the capabilities of autonomous systems, facilitating safer and more efficient operations in diverse and challenging scenarios.

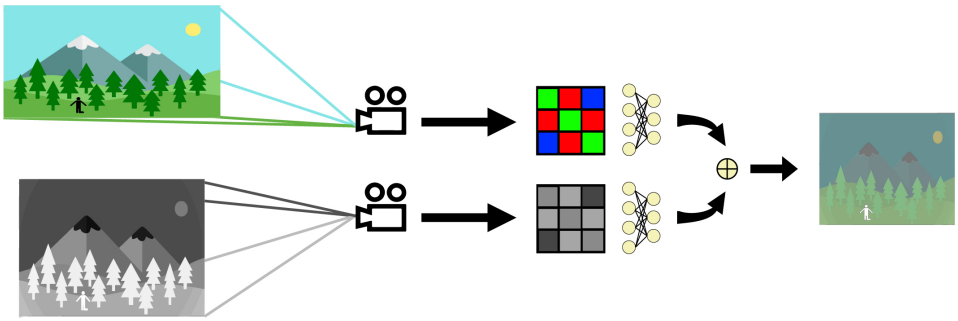

Figure 1.5: **System Conceptual Overview.** Images from visual (RGB) and thermal (IR) cameras capturing the same scene are misaligned due to physical properties. Our proposed methodology resolves this misalignment while emphasizing useful features from both modalities in the fused output.

1.4 Motivation and Objectives

The motivation behind this research is to address the problem of misaligned visual and thermal image fusion using an unsupervised deep learning method. The main

objective is to develop a methodology that can generate high-quality fused images by combining salient features from both modalities. By achieving this, we aim to enhance the effectiveness and efficiency of Wilderness Search and Rescue (WiSAR) operations.

WiSAR missions involve locating and rescuing individuals in remote and challenging outdoor environments such as forests, mountains, or deserts. These missions are often characterized by rugged terrains, limited visibility, adverse weather conditions, and the urgency to find missing persons in a timely manner. WiSAR operations require a coordinated effort involving search teams, drones, and advanced technologies to ensure successful outcomes.

One of the critical components in WiSAR operations is the ability to accurately detect and locate humans in the wilderness. In recent times, the utilization of UAVs equipped with a range of sensors, including visual cameras, thermal cameras, and various other sensor types, has become increasingly prevalent in supporting these operations. However, these sensors have inherent limitations that can hinder the effectiveness of search and rescue operations.

Visual cameras, for instance, rely on sufficient lighting conditions and may struggle to capture useful information in low-light or dark environments, which are common in wilderness settings. On the other hand, thermal cameras can detect heat signatures, making them valuable for detecting humans even in obscured environments. However, thermal cameras face challenges during periods of thermal crossover, where the ambient temperature is similar to that of living beings, impacting their accuracy in human detection.

The aforementioned limitations and challenges faced by WiSAR missions have motivated the problem of misaligned visual and thermal image fusion. By developing a methodology that effectively fuses misaligned visual and thermal images, we aim to enhance the effectiveness and efficiency of WiSAR operations. The use of unsupervised deep learning methods provides a promising approach to tackle this problem, as it can learn to combine salient information from both modalities without relying on explicit image registration or ground truth fused images. By generating high-quality fused images, the proposed methodology aims to improve the perception capabilities of autonomous systems in WiSAR missions, facilitating the timely detection and rescue of individuals in wilderness environments.

1.5 Contribution

This thesis makes a significant contribution by addressing the challenging problem of fusing misaligned visual and thermal image pairs through the implementation of advanced unsupervised deep learning methods. The primary focus is to develop a robust methodology for generating high-quality fused images while considering the practical considerations of WiSAR operations. The major contributions of this study are as follows:

- The proposed method revolutionizes the traditional approaches to misaligned image fusion by eliminating the need for image registration, a time-consuming and often error- prone process. Instead, through the utilization of unsupervised deep learning, the method overcomes the need to align the features of the visual and thermal images, ensuring accurate and efficient fusion.
- A distinguishing factor of the proposed methodology is its independence from ground truth fused images during model training. Traditional supervised methods often require annotated ground truth data, which can be challenging WiSAR scenarios. By opting for unsupervised learning, the proposed approach overcomes this limitation, making it highly applicable in real-world settings.
- Moreover, the fused outputs generated by the proposed method are inherently human-interpretable. This characteristic is crucial in WiSAR operations, where human operators actively monitor video feeds to aid in the detection and rescue of individuals. The human-interpretable fused images enable better situational awareness and decision-making during critical missions.
- In addition to its practical advantages, the proposed methodology ensures a balanced representation of both visual and thermal features in the fused images. This balance is vital for enhancing the effectiveness and reliability of autonomous systems in WiSAR operations. By preserving the salient information from each modality, the fused images provide a more comprehensive and informative view of the environment, enabling accurate human detection and aiding in the successful execution of rescue missions.

1.6 Problem Statement

The problem addressed in this research is to fuse misaligned visual and thermal image pairs, $(I_{\text{RGB}}, I_{\text{IR}})$, into a single image I_{fus}, where salient information from

each image is combined. Refer to figure 1.5. The challenge lies in the fact that the image pairs are not aligned due to different optical characteristics and the absence of ground truth fused images. Furthermore, the fused image must remain human-interpretable, considering the involvement of human operators in WiSAR missions.

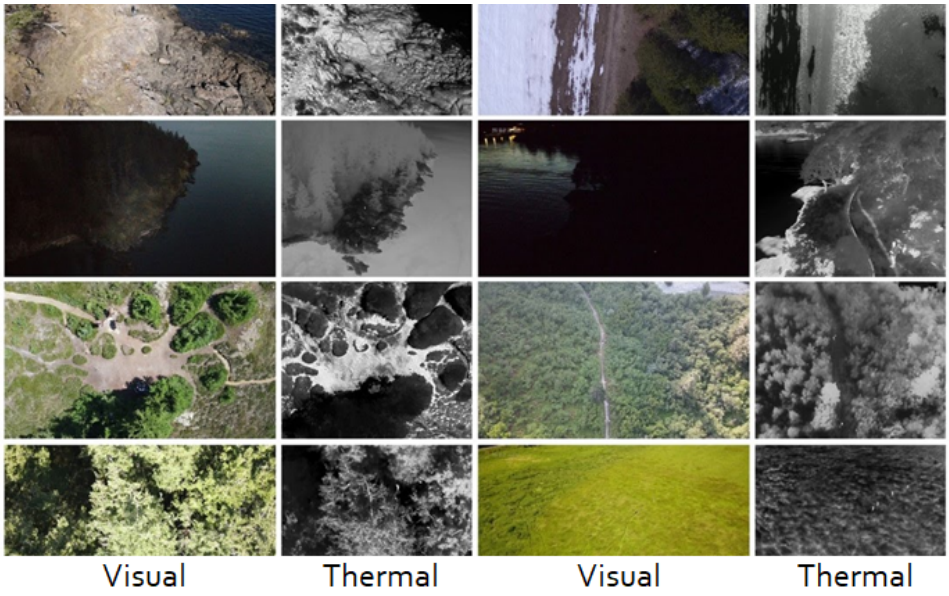

Figure 1.6: Examples from the WiSARD dataset demonstrate the inherent misalignment between visual and thermal images, highlighting the challenging scenarios where one modality excels while the other performs sub-optimally.

Chapter 2

PRELIMINARIES

2.1 Homography in Computer Vision

Homography stands as a cornerstone within the realm of computer vision, serving as a vital bridge that brings together two images captured from distinct viewpoints but depicting the same scene. This seemingly simple concept, with its underpinning of geometric transformation, holds a position of paramount importance across a range of real-world applications. Imagine seamlessly stitching together multiple images to create panoramic vistas that extend beyond a single camera's field of view. Think of the process that ensures a virtual object convincingly interacts with the real world in augmented reality experiences. Visualize a camera being calibrated, its intrinsic and extrinsic parameters fine-tuned with accuracy. All of these scenarios owe their functionality to the versatile and adaptable concept of homography.

Mathematical Representation of Homography

Homography is a projective transformation that maps points in one image to their corresponding points in another image. It can be represented using a 3x3 matrix, denoted as (**H**), that relates the coordinates of points in one image (**P**) to their coordinates in another image (**P'**):

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = H \begin{bmatrix} x \\ y \\ w \end{bmatrix} \tag{2.1}$$

Here, (x, y, w) are the homogeneous coordinates of a point (**P**) in the first image, and (x', y', w') are the transformed homogeneous coordinates of the point (**P'**) in the second image. The transformation is normalized by dividing the coordinates by w to obtain the Cartesian coordinates.

Estimation of Homography

The estimation of homography involves finding the matrix H that best aligns corresponding points in two images. This is typically achieved using a set of point correspondences between the images. One common method for homography estimation is the Direct Linear Transform (DLT) algorithm. Given at least four point

correspondences, the DLT algorithm formulates an over-determined linear system of equations, which is then solved using techniques such as Singular Value Decomposition (SVD).

Application of Homography

- **Image Stitching**

  Homography is widely used in image stitching to combine multiple overlapping images into a seamless panorama. By estimating the homography between adjacent images, we can warp and align them to a common reference frame, enabling the creation of a panoramic view.

- **Camera Calibration**

  Homography is essential for camera calibration, where the intrinsic and extrinsic parameters of a camera are determined. By capturing images of a known calibration pattern from different perspectives, homography can be used to calculate the transformation between the calibration pattern and the camera.

- **Object Recognition**

  Homography plays a role in object recognition by enabling the transformation of object features from one viewpoint to another. This allows for robust matching and recognition of objects across different images or video frames.

- **Augmented Reality**

  In augmented reality applications, homography is used to overlay virtual objects onto the real world by aligning them with the captured scene. By estimating the homography between the camera view and the virtual object, realistic augmentations can be achieved.

2.2 Autoencoder

In the rapidly evolving domain of artificial intelligence, autoencoders have emerged as a prominent and captivating paradigm [14]. An autoencoder is a neural network architecture that functions as a data compression algorithm, capable of capturing intricate underlying patterns within the input data. It represents a powerful tool that not just facilitates data reduction, but also facilitates the revelation of the underlying patterns within the input data, thereby exposing its latent structure. This, in turn,

aids in comprehending the data and manipulating it according to our needs. In this section, we embark on a comprehensive exploration of autoencoders their conceptual framework, diverse variants, and the vast spectrum of applications they empower.

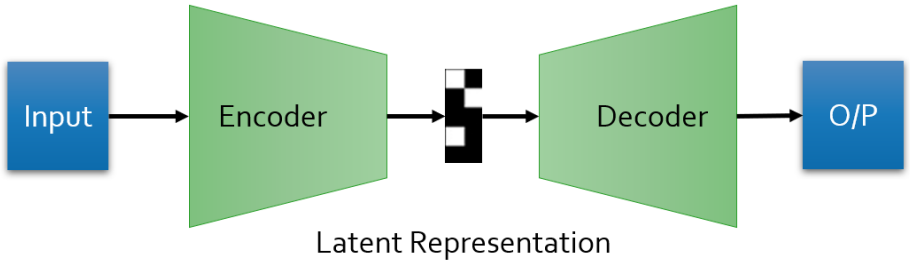

Figure 2.1: This image illustrates the fundamental components of an autoencoder. Starting with the input data (left), the encoder compresses the information into a latent representation (middle). The decoder then reconstructs the original input from this compressed form, resulting in the output data (right). Autoencoders are like data wizards, transforming complex data into a simpler representation and back again.

Conceptual Framework of Autoencoders

An autoencoder is composed of two primary components: an encoder and a decoder as depicted in Figure 2.1. The encoder is responsible for transforming the input data into a compact representation, often referred to as a latent space or representation. This compact representation captures essential features and patterns present in the input data. Subsequently, the decoder takes this compact representation and reconstructs the original input data. In essence, the autoencoder learns to compress the input data into a lower-dimensional representation and then reconstruct it, aiming to minimize the reconstruction error. This process encourages the autoencoder to capture meaningful features and underlying structures in the data. The intertwining of these two components allows an autoencoder to extract salient information and uncover hidden patterns within the input data.

Variations and Specializations

Autoencoders have various types to address specific challenges or extract certain types of features from data. Denoising autoencoders, for instance, equip themselves to denoise corrupted input data, sharpening their ability to learn robust and salient features. Sparse autoencoders introduce sparsity constraints on the hidden layer activations during training. By encouraging only a subset of neurons to activate

for a given input, sparse autoencoders tend to learn more robust and salient features. Variational autoencoders (VAE), introduce probabilistic modeling into the autoencoder framework. They not only learn to map inputs to a latent space but also generate new data points by sampling from this space. VAEs enable smooth interpolation in the latent space and can generate diverse data samples.

Applications Across Domains

Autoencoders have found a multitude of practical applications across diverse domains. In image processing, they excel in tasks such as denoising images, reconstructing damaged or missing parts of images, and even generating new images. They also play a pivotal role in anomaly detection by learning the normal patterns of data and flagging deviations from those norms. Autoencoders are vital in dimensionality reduction, which is crucial for visualizing high-dimensional data and speeding up subsequent computations. In the realm of natural language processing, autoencoders aid in language translation, summarization, and sentiment analysis. Furthermore, they are instrumental in recommender systems, where they learn users' preferences and suggest relevant products or content.

2.3 Generative Adversarial Network (GAN)

In recent years, Generative Adversarial Networks (GANs) have emerged as an intriguing and powerful concept in the realm of artificial intelligence [1]. GANs are a class of artificial intelligence algorithms that consist of two neural networks, a generator and a discriminator, engaged in a competitive process to generate realistic data. GANs have revolutionized various domains by enabling the creation of realistic content, such as images, music, and even text, with profound implications for creativity and innovation.

Deconstructing GAN Components: The Generator and the Discriminator

- **Generator**

  In the world of GANs, the generator is a neural network responsible for creating data. It takes random noise as input and gradually refines it to generate data that should resemble the target data distribution. The generator essentially learns to produce samples that are indistinguishable from real data by transforming noise into meaningful patterns. As it evolves through training, the generator becomes proficient at generating data that progressively becomes more convincing to the discriminator.

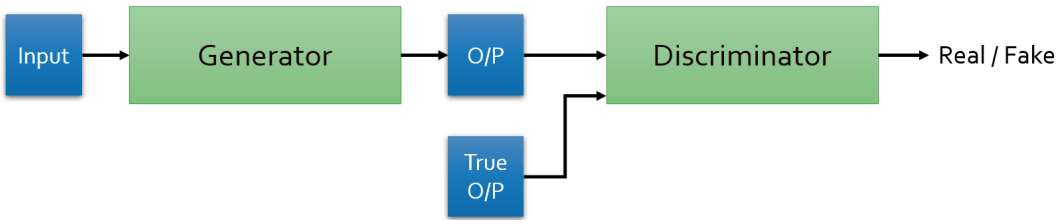

Figure 2.2: This image provides an overview of the Generative Adversarial Network (GAN) architecture, showcasing its fundamental components: the Generator and the Discriminator. The Generator, depicted on the left, takes random noise as input and progressively refines it to produce synthetic data (O/P) that resembles real samples (True O/P). On the right, the Discriminator receives both real data and generated data, aiming to distinguish between the two. The interaction between the Generator and the Discriminator forms an adversarial game, driving the Generator to create increasingly convincing data that can deceive the Discriminator. This competitive process results in the refinement of the Generator's ability to produce highly realistic outputs, making GANs a powerful tool for generating complex and high-quality data.

- **Discriminator**

  On the other side of the GAN competition stands the discriminator. This neural network's role is to distinguish between real data and artificially generated data from the generator. The discriminator's training involves learning to classify input data as either real or fake. Over time, it becomes increasingly skilled at differentiating authentic data from the generated data. The discriminator's feedback serves as a crucial signal for the generator to refine its output to produce data that is progressively more authentic and challenging for the discriminator to classify.

Working of GAN

A GAN operates as an adversarial game where the generator and the discriminator engage in a dynamic interplay. The generator crafts data from random noise, aiming to create samples that resemble real data as closely as possible. Simultaneously, the discriminator evaluates these generated samples along with actual real data, striving to accurately distinguish between the two. As training progresses, the generator refines its output to create more convincing data that the discriminator struggles to differentiate. This adversarial dynamic propels both networks to improve iteratively, with the generator producing increasingly realistic data and the discriminator becoming more adept at differentiation. This back-and-forth process continues until the generator's output is virtually indistinguishable from real data. The result is a

generator capable of producing data that holds a striking resemblance to the actual target distribution.

Applications of GAN

In the realm of computer vision, GANs have revolutionized image synthesis, enabling the creation of realistic images that are often indistinguishable from real photographs. These generative capabilities have been harnessed for tasks like image-to-image translation, style transfer, and super-resolution. GANs have also made substantial contributions to medical imaging, aiding in tasks such as image denoising, segmentation, and even generating synthetic medical images for training deep learning models.

In conclusion, the remarkable capacity of Generative Adversarial Networks to create realistic and novel data has established them as a pivotal innovation in the field of artificial intelligence. By fostering a dynamic interplay between the generator and discriminator, GANs have unlocked new avenues for image synthesis, data augmentation, style transfer, and countless other applications.

2.4 Attention Mechanism

In the ever-evolving landscape of artificial intelligence, attention mechanisms [22] have emerged as a pivotal tool for enhancing the efficiency and accuracy of various tasks. These mechanisms mimic the human cognitive process of focusing on specific elements within a vast sea of information. In this section, we embark on a journey to explore attention mechanisms, with a keen focus on multihead attention and cross attention, unraveling their mathematical underpinnings and their transformative impact across diverse applications.

Attention

At the heart of attention lies a simple yet profound concept: the ability to selectively emphasize and weigh different parts of input data. This enables models to allocate resources effectively to the most relevant information, facilitating more informed decisions. Mathematically, the attention mechanism can be conceptualized as a function that takes in a set of input values, often referred to as "values," and assigns weights to these values based on their relevance to a specific context. This weighting process is carried out using two additional sets of data, known as "queries" and "keys." The queries represent what the model is specifically looking for in the data, while the keys provide a means to map the input values against the queries.

This relationship between queries, keys, and values is governed by a parameterized function that computes attention scores. These scores dictate the emphasis placed on each value, effectively shaping the final output. The mechanism's adaptability is manifested through the transformation of these attention scores into normalized weights using a softmax function. The weighted values are then linearly combined to produce the context vector, a summary representation that captures the most important aspects of the input data with respect to the given queries.

In the Attention module as explained in [22], the input representation (encoded features) is transformed into three vectors (i.e., query(Q), key(K), value(V)). Q is multiplied by K to generate an attention map between vectors. And V, which represents the value of the input representation, is multiplied by the attention map to get the result of the Attention module. The operation can be defined mathematically as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.2}$$

where d_k is the dimension of the q and k vectors.

Multihead Attention

While a single attention mechanism can provide valuable insights, multihead attention (see Figure 2.3) takes this concept a step further. In multihead attention, the model employs multiple instances of the attention mechanism in parallel, each operating with a different set of learned queries, keys, and values. This diversity enhances the model's ability to capture a wider range of relationships and patterns within the data. The outputs from these parallel attention heads are then concatenated and linearly transformed to generate the final multihead attention output.

Cross Attention

Cross attention introduces a fascinating extension to the attention paradigm by allowing information from one modality to influence the processing of another. In scenarios involving multiple modalities, such as images and text, cross attention enables the model to understand the interplay between these modalities more effectively. Mathematically, this is achieved by using the queries from one modality to determine the relevance of the keys and values from another modality. This way, cross attention bridges the semantic gap between different data types, promoting a more holistic understanding of the input.

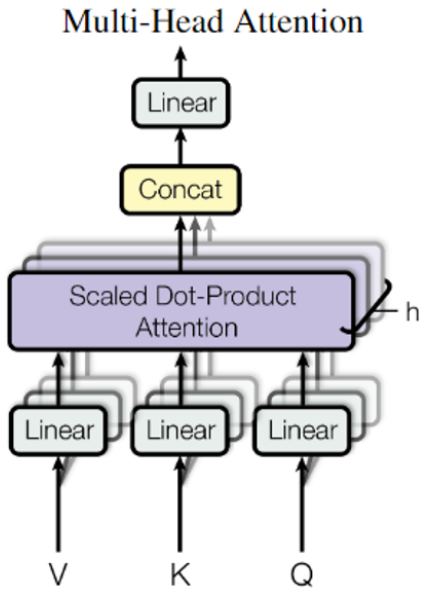

Figure 2.3: This image illustrates the concept of multihead attention, a key component introduced in the "Attention is All You Need" paper [22]. Multihead attention enables the model to focus on different parts of the input data, capturing various patterns and features simultaneously. The process involves multiple sets of queries, keys, and values, each contributing to different aspects of the attention mechanism. This results in a more comprehensive and nuanced understanding of the input data, leading to improved performance in various natural language processing and computer vision tasks.

In conclusion, attention mechanisms, multihead attention, and cross attention represent not just algorithms but cognitive strategies that allow AI models to focus, learn, and comprehend data in a more nuanced and context-aware manner. Their mathematical foundations, as well as their applications, provide us with a glimpse into the remarkable synergy between human cognitive processes and machine intelligence.

2.5 Thesis Organization

This thesis is organized as follows:

Related Work

This chapter provides a comprehensive review of existing research and approaches related to sensor fusion, image fusion, and deep learning-based techniques. It highlights the gaps in the literature that this research aims to address.

Methodology

In this chapter, we present the details of the proposed MISFIT-V approach for fusing misaligned visual and thermal images. The architecture, components, and training process of the model will be discussed in depth.

Experiment and Results

Here, we present the results of extensive experiments conducted to evaluate the performance of the proposed methodology. We compare the results with existing state-of-the-art methods and analyze the advantages and limitations of our proposed method.

Conclusion

This chapter offers a comprehensive discussion of the research findings, their implications, and potential future directions. We conclude by summarizing the contributions of this study to the field of sensor fusion and autonomous systems.

Chapter 3

RELATED WORK

The problem of image fusion has been extensively explored in the literature (see [16] for a review.), and various methods have been proposed to tackle this challenge. Many standard image fusion methods for thermal and visual images, as presented in [3]

3.1 Deep Learning Based Fusion Approaches

Pixel-Level Alignment for Thermal and Visual Images

Deep Learning Based image fusion methods, such as those proposed in [3, 9, 15, 20] for thermal and visual images have historically relied on the assumption that the input images are precisely aligned at the pixel level. The process of aligning visual and thermal images typically involves intricate hardware-level calibration and image processing techniques to achieve precise registration. In this context, **Homography** is a commonly used method to estimate the relationship between two images of the same scene and align them accordingly. Homography relies on identifying common features in both images and finding correspondences between them to establish the transformation needed for alignment. Recent work proposed in [8] which is based on [25], uses an unsupervised learning approach to estimate a homography to align the two images.

However, in real-world scenarios, especially in the case of two different modalities such as Thermal & Visual and dynamic environments where images are captured from different viewpoints or under varying conditions, achieving pixel- level alignment can be challenging. One significant obstacle is the lack of common features between thermal and visual images, as they operate based on different principles and exhibit distinct features. Consequently, using traditional Homography directly on the raw visual and thermal images does not work well and more often than not fails to find suitable correspondences, leading to misaligned features in the fused images.

To address this issue, one approach we explored was blurring the visual image to reduce its level of detail, attempting to create common features that could be matched with the thermal image using Homography. While this method yielded

some success in finding correspondences on specific images, it was not a scalable solution. (Refer to Figure 3.1 and Figure 3.2). The inherent differences in thermal and visual image features, such as ORB (Oriented FAST and Rotated BRIEF) and SIFT (Scale-Invariant Feature Transform) [19], posed significant challenges for the Homography calculation, limiting its effectiveness overall.

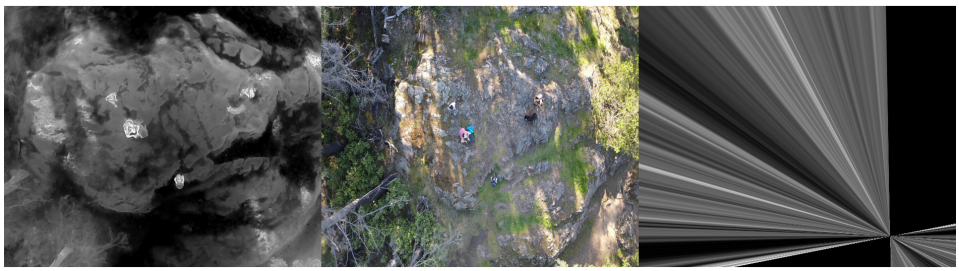

Figure 3.1: An example of bad homography estimation due to lack of features from the thermal image. The presented images consist of a thermal image (left), and a visual image (middle), which are used to calculate homography and generate a homographed thermal image (right).

Figure 3.2: An example of good homography estimation. Although the given two examples are taken from a similar scene, the lack of thermal features in the example presented in figure 3.1 leads to a bad homography estimation which suggests the inconsistency in estimating homography for a larger dataset.

As a result, while Homography is a powerful technique for aligning images with common features, its application in aligning thermal and visual images proved to be less effective due to the lack of inherent common features between the two modalities. This limitation has motivated the exploration of alternative approaches, such as deep-learning-based image-to-image translation and cross-attention mechanisms, to address the challenges of misalignment and enhance the performance of image fusion algorithms for autonomous systems.

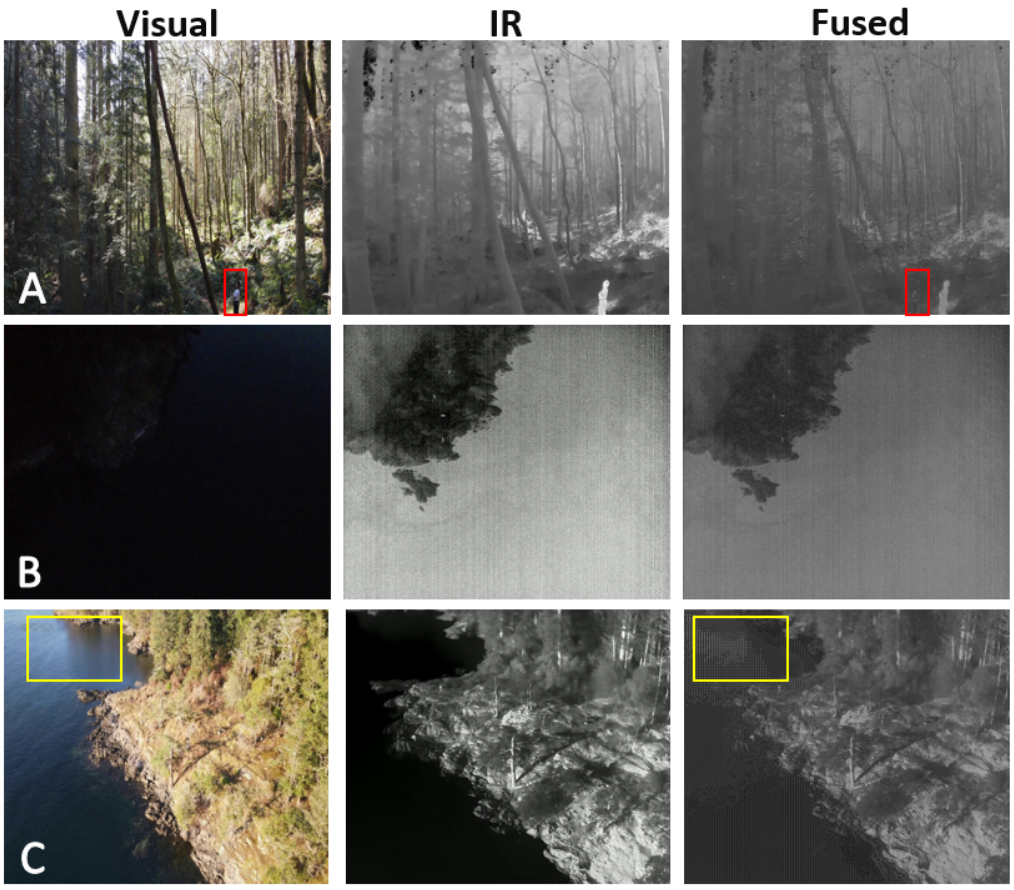

Figure 3.3: **Unwanted Artifacts**, The problem of the appearance of unwanted artifacts while fusing the misaligned images together. In Ex. 'A' the visual features appear in the fused image without changing their location with respect to the original visual image (highlighted with the red box), in Ex. 'B' the visual image has no useful information as against its thermal counterpart and the fused image has more of the thermal features, in Ex. 'C' the fused image has the visual image features highlighted in yellow

Challenges of Misalignment in Practical Scenarios

In real-world applications, visual and thermal imaging sensors operate with distinct characteristics, resulting in inherent misalignment between the two modalities. Factors such as differences in resolution, field of view, lens effects, and noise characteristics can lead to misalignment, making it difficult to obtain perfectly aligned visual and thermal images. In scenarios where autonomous systems use both visual and thermal cameras, the misalignment becomes a critical challenge to address. Misaligned features in the fused images can affect the system's ability to accurately interpret the environment and make informed decisions (see Figure 3.3), motivating the exploration of alternative approaches that can effectively handle misaligned

visual and thermal image pairs.

3.2 Image-to-Image Translation Methods for Fusion

Using Generative Adversarial Networks (GANs) for Translation

To address the lack of common features between visual and thermal images, recent research has explored image-to-image translation [11, 24, 26] techniques as a potential solution. Image-to-image translation involves converting images from one domain to another, adapting their style or characteristics to resemble images from the target domain. In the context of image fusion, these methods leverage the power of Generative Adversarial Networks (GANs) to learn the mapping between visual and thermal images (see Figure 3.4). GANs consist of a generator and a discriminator network, where the generator learns to generate images that are indistinguishable from real images, and the discriminator learns to differentiate between real and generated images. By training the GAN on pairs of visual and thermal images, the generator can learn to translate images from one modality to the other, effectively fusing the information from both modalities.

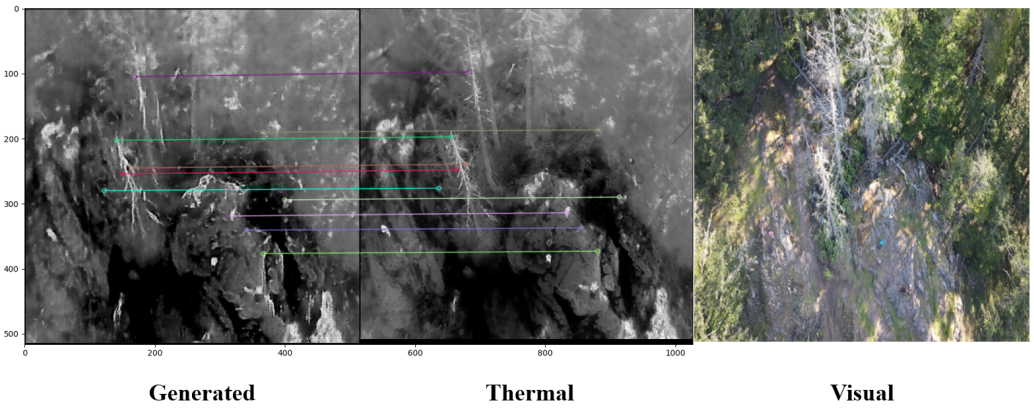

Figure 3.4: In this illustration, a Pix-to-Pix [11] based Generative Adversarial Network (GAN) network is employed to generate a Thermal Image (on the left) using the original thermal image (in the middle) and the visual image (on the right) of the same scene. The GAN network translates the visual image into a corresponding thermal image, and feature matching, using ORB features in this case, is performed to estimate the homography between the generated thermal image and the original thermal image.

Limitations of Pre-registered and Pixel-Aligned Datasets

Despite the promise of image-to-image translation using GANs, many existing datasets (for example a dataset called TNO [21]) used for training these models are

pre-registered and aligned at the pixel level. While this alignment may be suitable for certain applications, it may not be practical in real-world scenarios such as wilderness search and rescue (WiSAR), where the alignment between visual and thermal sensor pairs is unknown or subject to variations. For instance, in aerial imagery captured by UAVs, misalignments between visual and thermal images are common due to differences in camera viewpoints and other environmental factors. Consequently, approaches that rely on pixel-aligned datasets may not be applicable to these real-world scenarios where pixel-level alignment cannot be assumed. For example, if we were to follow the conventional approach of using pixel-level aligned datasets for image-to-image translation using GANs in our case, we would observe that the generated thermal image, translated from the visual image, would closely resemble the original thermal image in terms of alignment as shown in Figure 3.4 and will fail to capture the misalignment from the visual image. Hence, for our WiSAR applications, where pixel-level alignment cannot be assumed or is subject to variations, we must seek alternative approaches that can effectively handle the misalignment between visual and thermal images.

3.3 Cross-Attention Mechanism in Multi-Modal Data Processing

Advancements in Natural Language Processing (NLP)

The attention mechanism has emerged as a significant advancement in the field of natural language processing (NLP). It was introduced in the Transformer model [22], a groundbreaking architecture for sequence-to-sequence tasks. The attention mechanism allows the model to establish meaningful relationships between different segments of input sequences, enabling the model to focus on relevant information from one segment while processing features from another segment. The success of the attention mechanism in NLP has motivated researchers to explore its application in other domains, including sensor fusion tasks that involve multiple sensor modalities. The cross-attention mechanism is the extension of the attention mechanism where data from more than one modality is involved.

Application of Cross-Attention in Image Fusion

In recent years, cross-attention has been successfully applied to multi-modal data processing tasks. For instance, [6] combined visual data and text data to classify skin diseases, [7] combined multi-scale picture characteristics using cross-attention, while [4] combined visual and LiDAR image features using self-attention and cross-attention. By incorporating the cross-attention mechanism into deep learning mod-

els, researchers have been able to effectively leverage the complementary information from different modalities and establish connections between different modality features. This approach allows the model to attend to relevant features from one modality while analyzing features from another, facilitating the fusion of misaligned information. As a result, cross-attention-based approaches show promise in handling the challenges posed by misalignment, improving the overall performance and interpretability of autonomous systems in image fusion applications.

By exploring the concepts and challenges presented in these subsections, this thesis aims to contribute to the development of advanced image fusion techniques for autonomous systems. The integration of image-to-image translation and cross-attention mechanisms seek to overcome the limitations of traditional approaches, making strides towards a more robust and accurate fusion of visual and thermal images in real-world scenarios.

Chapter 4

METHODOLOGY

In this chapter, we present a detailed description of the methodology employed. The model is designed to address the challenging problem of fusing misaligned visual and thermal image pairs using unsupervised deep learning methods. Our approach utilizes a Generative Adversarial Network (GAN) architecture in corporated with a cross-attention mechanism in the generator, which has shown promising results in various image synthesis tasks.

4.1 Problem Formulation

The problem of image fusion arises when attempting to combine misaligned thermal and visual images to generate a single fused image that retains the salient features and characteristics of both modalities. Given the distinct optical properties and imaging mechanisms of thermal and visual sensors, the captured images often suffer from misalignment due to differences in resolution, field of view (FOV), lens distortions, and noise characteristics. This misalignment hinders the direct fusion of the images and necessitates the development of advanced techniques to address this challenge.

Let I_{IR} and I_{RGB} represent the misaligned thermal and visual images, respectively. The goal is to design a fusion method F that takes these misaligned images as input and generates a fused image I_{fus} that enhances the informative content from both modalities while compensating for their misalignment. Mathematically, the fusion process can be represented as:

$$I_{\text{fus}} = F(I_{\text{IR}}, I_{\text{RGB}}) \tag{4.1}$$

To achieve this, several challenges need to be addressed:

- **Feature Extraction**

  Extract meaningful and complementary features from both modalities that capture the unique information present in each image. The extracted features should contribute to enhancing the overall quality and informative content of the fused image.

- **Information Fusion**

  Design an effective fusion strategy that combines the extracted features from both modalities in a manner that highlights the strengths of each while minimizing the impact of misalignment-induced artifacts. This step aims to create a fused image that represents a harmonious blend of thermal and visual information.

The proposed methodology should leverage advanced techniques such as unsupervised deep learning, cross-attention mechanisms, and generative adversarial networks to address the challenges posed by misaligned thermal and visual images. The ultimate objective is to devise a fusion method that produces high-quality fused images, with improved visual interpretability and informative content, suitable for applications such as search and rescue missions, surveillance, and autonomous systems.

In the following sections, we present a comprehensive exploration of the proposed approach, its architecture, training process, and evaluation against various metrics and variations to demonstrate its efficacy in solving the misaligned image fusion problem.

4.2 Proposed Architecture

The architecture of the proposed method is depicted in Figure 4.1. Our design is motivated by the idea of leveraging two discriminators, inspired by previous works in the field [13, 17]. This choice enables the model to effectively handle the fusion of misaligned visual and thermal images while preserving the unique characteristics of each modality. Below, we provide a comprehensive explanation of each component of the GAN-based architecture.

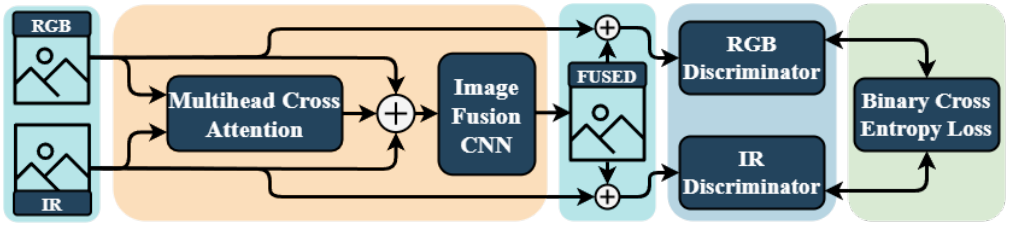

Figure 4.1: **Training Pipeline.** Thermal (IR) and visual (RGB) images are fed into a generator (orange block) consisting of a cross-attention module and CNN to produce a fused image. The fused image is fed into both discriminator networks, encouraging a balanced set of features from both images.

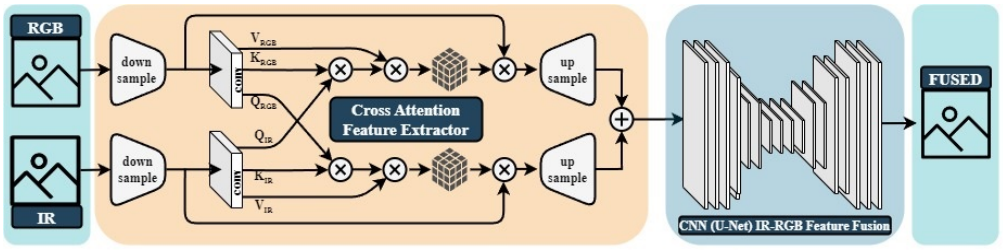

Figure 4.2: **Generator Architecture.** Input images are fed into a downsampling CNN ('Down') separately to retrieve their features. These features are then fed into the cross-attention network to calculate the cross-attention map between the modalities, which are multiplied with the downsampled features and concatenated to form the input to the U-Net CNN, which generates a fused image.

4.3 Generator with Cross-Attention

The generator is the central component of the proposed method, responsible for generating a fused image from a given misaligned visual-thermal image pair. Our generator network (refer Figure 4.2) comprises two separate Convolutional Neural Networks (CNNs) for downsampling and feature extraction from the input images. Additionally, we introduce a Cross-Attention Network, a powerful mechanism that allows the model to capture meaningful and unique features from both the visual and thermal modalities.

The Cross-Attention Network plays a pivotal role in addressing the fusion of misaligned input images. Unlike traditional image fusion methods that rely on pixel-level alignment, the Cross- Attention Network eliminates the need for explicit alignment by learning to attend to relevant regions in both modalities. By considering the diverse aspects that thermal and visible images focus on in the same scene, the Cross-Attention Network enables the model to capture complementary and relevant information from each modality.

Cross-Attention Mechanism

In the pursuit of achieving high-quality visual-thermal image fusion, we strategically incorporated a cross-attention mechanism within our generator network. This innovative addition enables the generator to selectively focus on significant features from both image modalities during the fusion process, thereby enhancing the overall quality of the generated output. To better understand the underlying working principle, it is essential to delve into the attention mechanism of the transformer architecture [22], which plays a fundamental role in our cross-attention approach.

In the context of attention, there are three key components: queries, keys, and values. Queries represent the specific information that the model seeks or searches for within the data, while keys and values hold the essential information that can be queried. In our case, we have two image modalities: visual and thermal. During the cross-attention process, the queries from one modality are exchanged with the queries from the other modality, resulting in a powerful information exchange mechanism. Specifically, queries from the visual modality are interchanged with queries from the thermal modality (see Figure 4.2). This strategic exchange enables the model to effectively capture and leverage the complementary information present in both modalities.

Figure 4.3: **Cross Attention**. The heatmaps on the right illustrate the regions of the images on the left that receive higher attention, with warmer colors indicating a greater degree of focus. In particular, areas with a reddish hue indicate heightened attention and prioritization.

Through this ingenious cross-attention approach, the generator gains the ability to attentively focus on the crucial features in the thermal modality using visual queries and vice versa. By effectively combining and synthesizing the relevant information from each modality, the generator achieves an accurate and precise fusion of the visual-thermal image pair. This ensures that the fused image retains the distinctive characteristics and essential features from both modalities, resulting in a comprehensive representation of the scene.

To visually illustrate the effectiveness of our cross-attention mechanism, Figure 4.3 provides an example of the attention process. The attention maps reveal the specific

regions of interest that are prioritized during the fusion process. By leveraging the power of cross-attention, our generator can precisely align and combine the salient features from the visual and thermal modalities, leading to enhanced fusion results and improved image quality.

In summary, the incorporation of cross-attention within our generator network is a pivotal aspect of our proposed methodology. This mechanism allows the model to selectively attend to crucial information from both visual and thermal modalities, ultimately facilitating accurate and effective image fusion. By effectively exploiting the complementary information and combining it in a coherent manner, our approach achieves remarkable results, showcasing its potential for various real-world applications, particularly in the context of WiSAR missions.

The outputs of the Cross-Attention Network are a tuple of cross-attention maps with respect to both modalities. These maps are multiplied with the downsampled features and fed into an upsampling CNN. The resulting outputs from the upsampling CNNs are concatenated to form the input to the U-Net CNN. The U-Net CNN integrates the information from both modalities and generates the final fused image. By fusing information at multiple levels and leveraging cross-attention, the proposed method can produce a comprehensive and meaningful fused image that retains critical details from both visual and thermal images.

4.4 Dual Discriminators

Due to the absence of ground truth data for the fused images, the proposed method utilizes a unique approach with two discriminator networks. As shown in Figure 4.1, each discriminator is responsible for classifying a concatenated image consisting of the original (visual or thermal) and the fused image. The two discriminators work collaboratively to assess the authenticity of the fused image with respect to both modalities. The architecture of each discriminator is explained in Table 4.1.

By employing dual discriminators, the proposed method is encouraged to learn and retain valuable information from both the visual and thermal modalities during the image fusion process. This ensures that the generator produces a balanced set of features from each modality in the fused output, leading to high-quality thermal-visual fused images.

Table 4.1: The architecture of the discriminator network is as follows: The input has I number of input channels, and the output has O number of output channels. The convolutional layers have kernel size K, stride size S, and padding size P. 'IN2D' represents InstanceNorm2D, and 'Conv' represents convolutional layers.

Layer	Discriminator Network
L1	Conv(I4, O64, K4, S2, P1), LeakyReLU
L2	Conv(I64, O128, K4, S2, P1), IN2D, LeakyReLU
L3	Conv(I128, O128, K4, S1, P0), IN2D, LeakyReLU
L4	Conv(I128, O256, K4, S2, P0), IN2D, LeakyReLU
L5	Conv(I256, O512, K4, S1, P0), IN2D, LeakyReLU
L6	Conv(I512, O1, K4, S1, P0), LeakyReLU

4.5 Loss Function

The loss function used to train the proposed method is carefully designed to optimize the generator's performance in generating high-quality fused images. The GAN framework is employed, which incorporates adversarial loss to encourage the generator to produce images that are indistinguishable from real images to the discriminators. Here, we describe the loss function used to train the proposed model. We employ an adversarial loss function to train the discriminators and generator in order to generate high-quality fused images. The adversarial loss for the discriminator, which corresponds to a specific modality X (either thermal/IR or visual/RGB), is defined as follows:

$$\mathcal{L}_{\text{adv,X}} = -\log D_{\text{X}}(I_{\text{X}}) - \log(1 - D_{\text{X}}(I_{\text{fus}})), \tag{4.2}$$

where $D_{\text{X}}(I_{\text{X}})$ represents the probability that I_{X} is classified as modality X (Thermal [IR] or Visual [VIS]), and $D_{\text{X}}(I_{\text{fus}})$ represents the probability that I_{fus} is classified as modality X by the discriminator.

The generator loss is defined as the sum of the adversarial losses from both discriminators, weighted by hyperparameters λ_{IR} and λ_{RGB}, respectively:

$$\mathcal{L}_{\text{gen}} = \lambda_{\text{IR}} \cdot \mathcal{L}_{\text{adv,IR}} + \lambda_{\text{RGB}} \cdot \mathcal{L}_{\text{adv,RGB}} \tag{4.3}$$

where λ_{IR} and λ_{RGB} control the relative importance of the respective losses in the overall generator loss. Here, the adversarial losses encourage the generator to create fused images that are indistinguishable from thermal images and visual images by training against two discriminators that try to classify them.

In addition to the adversarial losses, a Kullback-Leibler (KL) Divergence loss is used to compare the fused image generated by the generator with the original visual and thermal images in terms of their distribution. The KL Divergence loss is defined as:

$$\mathcal{L}_{\text{KL}} = \text{KL}(I_{\text{fus}}||I_{\text{IR}}) + \text{KL}(I_{\text{fus}}||I_{\text{RGB}}) \tag{4.4}$$

where $\text{KL}(I_{\text{fus}}||I_{\text{IR}})$ and $\text{KL}(I_{\text{fus}}||I_{\text{RGB}})$ represent the KL Divergence between the fused image I_{fus} and the thermal image I_{IR}, and between the fused image I_{fus} and the visual image I_{RGB}, respectively.

Furthermore, an L1 loss is utilized to calculate the pixel-wise differences between the fused image and each of the original images. This loss can be expressed as:

$$\mathcal{L}_{\text{L1}} = \|I_{\text{fus}} - I_{\text{IR}}\|_1 + \|I_{\text{fus}} - I_{\text{RGB}}\|_1, \tag{4.5}$$

where $\|\cdot\|_1$ denotes the L1 norm.

The L1 loss measures the absolute pixel-wise differences between the fused image and the original images, encouraging the generator to minimize these differences. By including the L1 loss in the overall generator loss, the model is incentivized to produce fused images that closely resemble both the thermal and visual images in terms of their pixel values. This helps preserve important details from each modality during the fusion process.

The overall loss,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} \tag{4.6}$$

is the sum of the generator loss, the KL divergence loss, and L1 loss weighted by hyperparameters λ_{KL}, and λ_{L1} that control the relative importance of the KL divergence loss and L1 loss.

The generator aims to generate fused images that can fool the discriminators, while the discriminators aim to correctly classify the fused images and the original images. The KL Divergence loss encourages the generator to produce fused images that are similar in distribution to the original images, helping to ensure the preservation of meaningful features from both modalities in the fused image.

Chapter 5

EXPERIMENT AND RESULTS

5.1 Dataset and Training Details

In our study, the proposed model was trained using a total of 2752 pairs of thermal and visual images, sourced from the WiSARD dataset [5]. For the purpose of training and validation, we employed an 80:20 split, ensuring that 80% of the dataset was used for training, while the remaining 20% was reserved for validation. Additionally, to evaluate the performance of the trained model and gauge its generalization capabilities, we curated a separate test dataset, comprising 200 pairs of images.

Throughout the training process, we executed 20 epochs, a substantial number of iterations necessary to optimize the model's parameters effectively. The learning rate was set to 1×10^{-4}, an empirically determined value that facilitates the convergence of the optimization algorithm while minimizing the risk of overshooting the optimal solution.

As for the hyperparameters employed during the training procedure, we carefully selected and fine-tuned their values to achieve optimal performance and a balanced training process. Specifically, we assigned weights to different loss components to steer the model's behavior effectively. To be precise, we utilized $\lambda_{\mathrm{KL}} = 10$, $\lambda_{\mathrm{L1}} = 100$, $\lambda_{\mathrm{IR}} = 1$, and $\lambda_{\mathrm{RGB}} = 1$ as the hyperparameters, each of which serves a crucial role in guiding the model's learning process.

The hyperparameter λ_{KL} represents the weight assigned to the Kullback-Leibler (KL) divergence term, which encourages the model to generate fused images that align better with the distribution of real-world images. A higher value for λ_{KL} emphasizes the significance of this term in the overall loss function, promoting the generation of more realistic and authentic fused images.

On the other hand, λ_{L1} serves as the weight for the L1 loss term, which fosters a higher-fidelity fusion process by penalizing the discrepancy between the fused image and the input images. By assigning a larger value to λ_{L1}, we prioritize the minimization of the L1 loss, thereby ensuring the preservation of essential visual and thermal characteristics in the final fused output.

Furthermore, λ_{IR} and λ_{RGB} correspond to the weights applied to the infrared and

RGB (visible) modalities, respectively. These hyperparameters play a crucial role in balancing the influence of each modality in the fusion process. By adjusting their values, we can control the model's emphasis on each modality, ensuring that both thermal and visual information contribute meaningfully to the generation of the fused image.

By employing carefully tuned hyperparameters and a well-structured training pipeline, we aimed to achieve a model that not only accurately fuses misaligned thermal and visual images but also retains the salient features of each modality, resulting in high-quality fused images that are both visually appealing and informative.

In summary, the training process involved the use of a substantial number of image pairs from the WiSARD dataset, employing an appropriate split for training, validation, and testing. We ensured adequate training iterations to converge the model effectively and carefully selected hyperparameters to optimize the loss function and guide the model's learning process. These efforts were geared towards creating a robust and high-performing model capable of addressing the challenges of fusing misaligned visual and thermal images in the context of WiSAR missions

5.2 Qualitative Analysis

In our study, we conducted an extensive evaluation of the fused images generated by the proposed model, aiming to showcase their superiority in representing the environment compared to the individual modalities. The results, as depicted in Figure 5.1, provide compelling evidence of the effectiveness of our method in producing well-fused images that successfully preserve essential terrain features while accurately extracting the bright human silhouettes from the thermal image. The fused images exhibit a remarkable combination of visual and thermal information, providing a comprehensive and informative representation of the scene.

Despite the remarkable performance of our proposed model, it is essential to acknowledge its inherent limitations, which became evident in certain scenarios. For instance, when objects of interest are relatively small in size, the attention mechanism employed in our generator network may encounter challenges in accurately determining the essential features from both modalities. As a result, this limitation can manifest in the form of ghost artifacts in the fused image, as observed in the fourth row of Figure 5.1.

These limitations are essential to consider in practical applications, especially in situations where precise identification of small objects plays a critical role. Un-

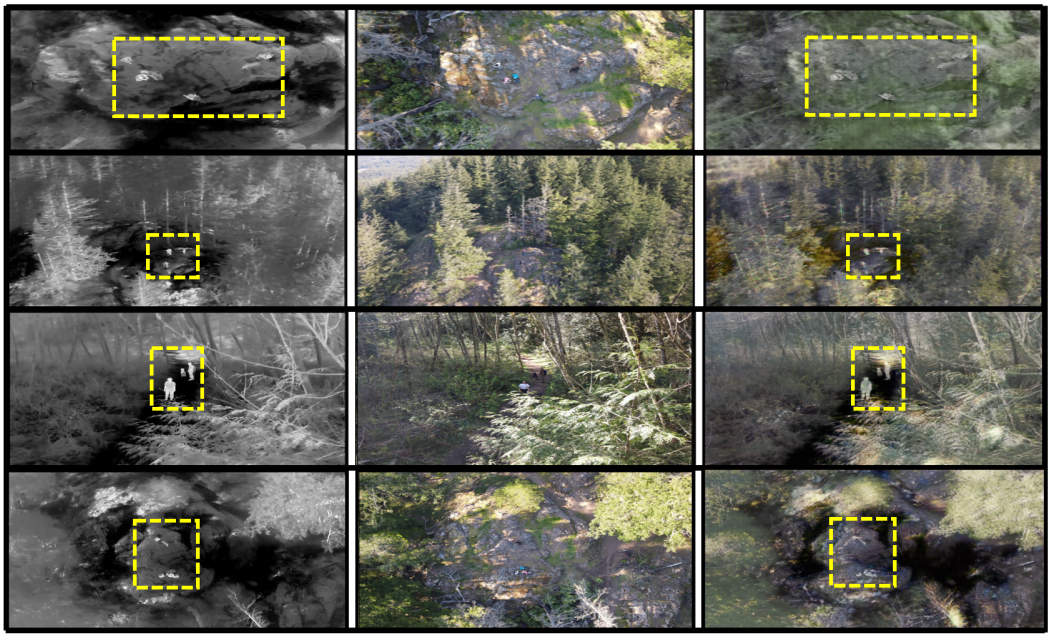

Figure 5.1: **Results.** Each row, from left to right, shows a scene's thermal representation, its visual representation, and finally the resulting fused image via the proposed method, in a wilderness environment. The yellow bounding boxes highlight the locations of humans.

derstanding such limitations pave the way for future improvements and potential refinements of the proposed model, leading to more robust and versatile image fusion techniques in the context of Wilderness Search and Rescue (WiSAR) operations.

Nonetheless, the promising results obtained from our model attest to its ability to enhance the quality and informativeness of the fused images, surpassing the capabilities of the individual modalities alone. The clearer representation of the environment offered by the fused images showcases the potential of the proposed method in bolstering the effectiveness and efficiency of WiSAR missions, where accurate detection and understanding of the surroundings are of paramount importance. By highlighting both the successes and the limitations of our proposed methodology, we contribute valuable insights to the field of image fusion, laying the foundation for further advancements and innovation in this domain.

5.3 Quantitative Analysis

To comprehensively assess the fusion results and quantify the extent to which the information from each modality is preserved, we conduct an in-depth analysis

using specific metrics tailored for thermal and visual images. This evaluation process is essential in gauging the efficacy of the fusion process and understanding the capabilities of the proposed MISFIT-V model compared to the state-of-the-art SeAFusion [20], which has also demonstrated high performance in visual-thermal image fusion.

To facilitate a fair and meaningful comparison, we evaluate both the proposed method and SeAFusion on an autonomous driving dataset [10], which offers more structured and complex scenes compared to WiSAR settings, and crucially, provides ground truth labels for the images. This the dataset enables us to conduct a rigorous evaluation, where we employ five widely used metrics for image quality assessment: Mean-squared-error (MSE), universal quality index (UQI), multi-scale structural similarity (MSSSIM), normalized mutual information (NMI), and the peak signal-to-noise ratio (PSNR). These metrics provide a robust and comprehensive evaluation of the fusion performance, capturing different aspects of the image quality.

Mean Squared Error (MSE)

The Mean-squared-error (MSE) is a common metric used to measure the average squared difference between the fused image and the ground truth image. It quantifies the overall pixel-wise difference between the two images, with lower MSE values indicating better fusion performance. The MSE is computed using the following formula:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_{\text{fus}}(i) - I_{\text{GT}}(i))^2, \tag{5.1}$$

where $I_{\text{fus}}(i)$ represents the pixel intensity of the fused image at pixel location i, $I_{\text{GT}}(i)$ is the corresponding pixel intensity of the ground truth image, and N is the total number of pixels in the images.

Universal quality index (UQI)

The Universal quality index (UQI) is a metric that evaluates the structural similarity between the fused image and the ground truth image. It takes into account both the luminance and contrast information and ranges from 0 to 1, with higher values indicating better fusion results. The UQI is calculated using the following equation:

$$\text{UQI} = \frac{4\sigma_{I_{\text{fus}}}\sigma_{I_{\text{GT}}}}{(\sigma^2_{I_{\text{fus}}} + \sigma^2_{I_{\text{GT}}})} \times \frac{(2\mu_{I_{\text{fus}}}\mu_{I_{\text{GT}}} + C_1)(2\sigma_{I_{\text{fus}}I_{\text{GT}}} + C_2)}{(\mu^2_{I_{\text{fus}}} + \mu^2_{I_{\text{GT}}} + C_1)(\sigma^2_{I_{\text{fus}}} + \sigma^2_{I_{\text{GT}}} + C_2)} \tag{5.2}$$

where $\mu_{I_{\text{fus}}}$ and $\mu_{I_{\text{GT}}}$ are the mean pixel intensities of the fused image and the ground truth image, $\sigma_{I_{\text{fus}}}$ and $\sigma_{I_{\text{GT}}}$ are their respective standard deviations, and $\sigma_{I_{\text{fus}}}\sigma_{I_{\text{GT}}}$ is the cross-covariance between the fused image and the ground truth image. C_1 and C_2 are small constants added to avoid division by zero.

Multi-Scale Structural Similarity (MSSSIM)

The Multi-Scale Structural Similarity (MSSSIM) is an extension of the structural similarity index (SSIM) that incorporates multi-scale information. It assesses the structural similarity between the fused image and the ground truth image at different scales and provides a more detailed evaluation of fusion performance. The MSSSIM is calculated using the following equation:

$$\text{MSSSIM} = \frac{1}{M}\sum_{i=1}^{M} SSIM_i^{\alpha} \tag{5.3}$$

where M is the number of scales considered, $SSIM_i$ is the SSIM[1] value at scale i, and α is a weight that determines the importance of each scale.

Normalized Mutual Information (NMI)

The Normalized mutual information (NMI) measures the amount of mutual information shared between the fused image and the ground truth image, normalized by their individual entropies. It ranges from 1 to 2, with higher values indicating better fusion performance. The NMI is computed using the following equation:

$$\text{NMI} = \frac{MI(I_{\text{fus}}, I_{\text{GT}})}{\sqrt{H(I_{\text{fus}}) \cdot H(I_{\text{GT}})}} \tag{5.4}$$

where $MI(I_{\text{fus}}, I_{\text{GT}})$ is the mutual information between the fused and ground truth image, and $H(I_{\text{fus}})$ and $H(I_{\text{GT}})$ are their individual entropies.

[1]
$$\text{SSIM}(I_1, I_2) = \frac{(2\mu_{I_1}\mu_{I_2} + C_1) \cdot (2\sigma_{I_1 I_2} + C_2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + C_1) \cdot (\sigma_{I_1}^2 + \sigma_{I_2}^2 + C_2)}$$

where:

- I_1 and I_2 are the images to be compared.
- μ_{I_1} and μ_{I_2} are the mean values of I_1 and I_2, respectively.
- σ_{I_1} and σ_{I_2} are the standard deviations of I_1 and I_2, respectively.
- $\sigma_{I_1}\sigma_{I_2}$ is the cross-covariance of I_1 and I_2.
- C_1 and C_2 are small constants added to avoid division by zero.

Peak Signal-to-Noise Ratio (PSNR)

The Peak signal-to-noise ratio (PSNR) is a commonly used metric to evaluate the quality of the fused image compared to the ground truth image. It measures the ratio of the maximum possible pixel intensity to the mean squared error between the two images. Higher PSNR values indicate better fusion results. The PSNR is calculated using the following equation:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{max intensity}^2}{MSE} \right) \tag{5.5}$$

where max intensity is the maximum pixel intensity in the images, and MSE is the mean squared error calculated between two images.

Figure 5.2 showcases the comparison results, with the y-axis representing the numerical values corresponding to each metric. For the sake of brevity, we present the results for three of the normalized metrics in this section, and the plots for the remaining metrics are available in separate figures, Figure 5.3 and Figure 5.4. The evaluation demonstrates several interesting trends in the performance of the proposed method and SeAFusion.

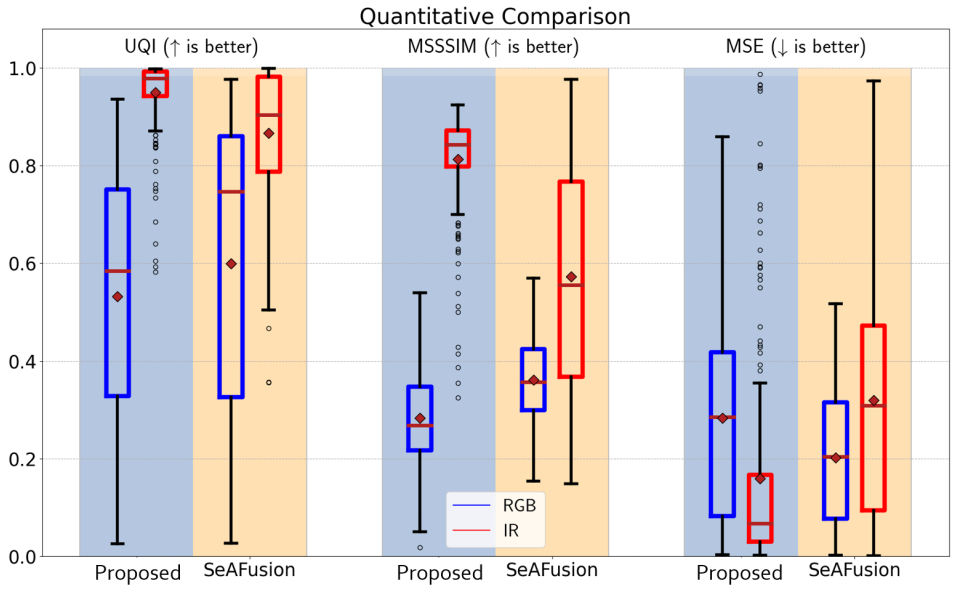

Figure 5.2: **Method Comparison.** This plot shows the results of three performance metrics comparing thermal and visual images against fused images generated by the proposed method and SeAFusion.

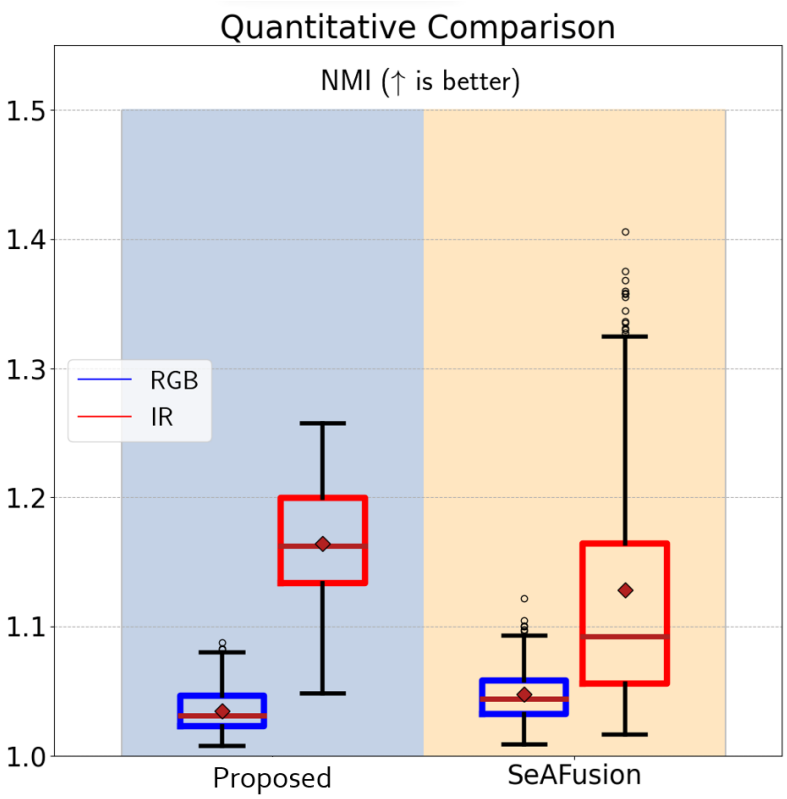

Figure 5.3: Normalized Mutual Information Comparison

One notable observation is that this method exhibits a tendency to prioritize information from the thermal modality to a greater extent while still performing commendably in retaining visual information, albeit with only marginal differences compared to SeAFusion. In certain aspects, the proposed method even outperforms SeAFusion, whereas, in other metrics, it demonstrates only slightly inferior performance. Overall, the proposed method showcases its strength and versatility by delivering competitive fusion results, and more importantly, by providing a significant advantage over SeAFusion.

A remarkable distinction lies in the proposed method's ability to eliminate the requirement for semantic labeling and ground truth data during the training process. This characteristic empowers the model with enhanced scalability and generalizability, enabling it to be readily applied to diverse datasets and real-world scenarios. The independence from ground truth data not only streamlines the training procedure but also enhances the practical utility of the proposed method in settings where obtaining labeled data may be challenging or impractical.

The comprehensive evaluation and comparison with SeAFusion, along with the

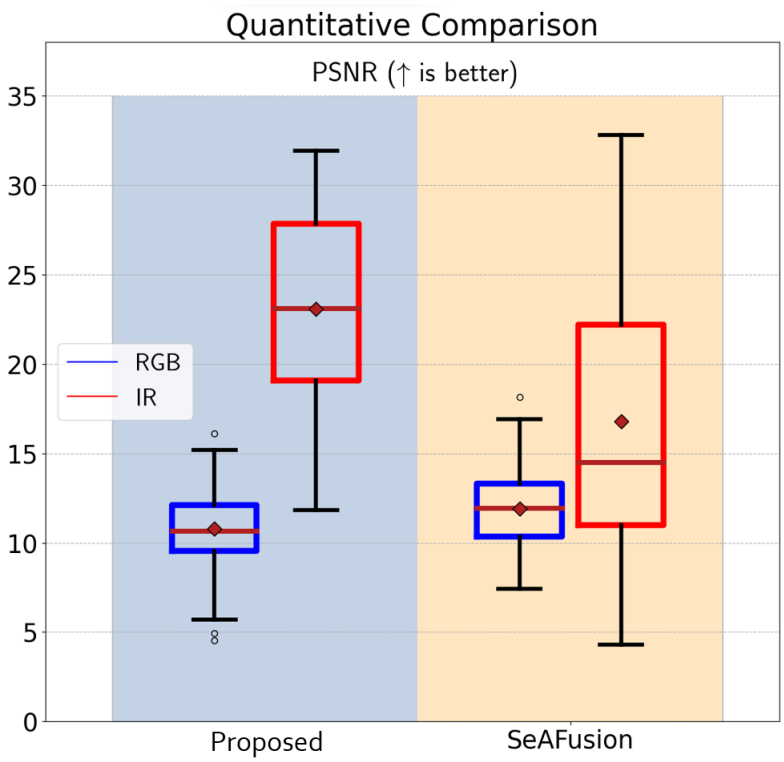

Figure 5.4: Peak Signal Noise Ratio Comparison

consistent and favorable fusion outcomes, reinforce the efficacy and applicability of the proposed method as an effective solution for visual-thermal image fusion. The results obtained through this rigorous evaluation provides compelling evidence of the potential benefits and practical advantages of our proposed model, emphasizing its significance in advancing the field of image fusion for autonomous systems, particularly in the context of Wilderness Search and Rescue (WiSAR) operations.

5.4 Ablation Study

In the pursuit of refining and optimizing our proposed methodology, we conducted an ablation study to meticulously analyze the effects of various modifications on the performance of our model. Through a series of controlled experiments, we aimed to dissect the contribution of specific components and choices within the architecture and total loss function. Here, we present the findings of our ablation study, comparing the original method with distinct variations: one involving the adjustment of the weightage of certain loss functions (Figure 5.5) and excluding the cross attention mechanism between thermal and visual data.

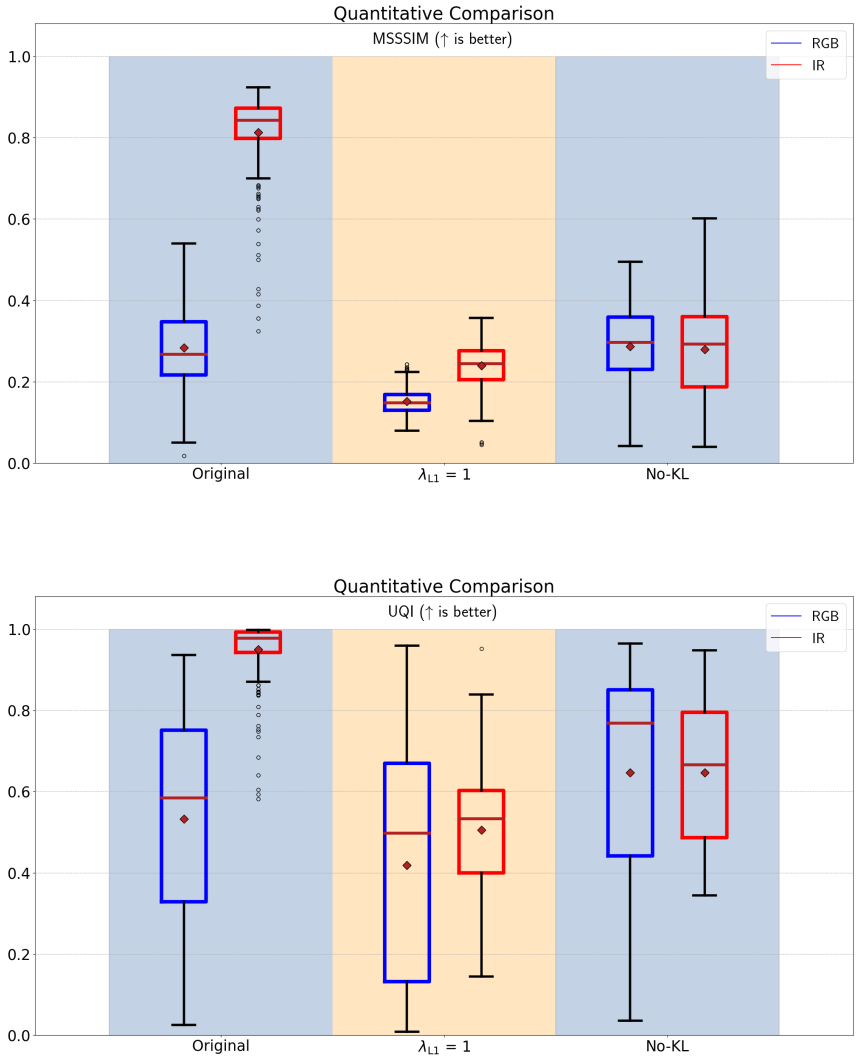

Figure 5.5: Comparison of image fusion performance using metrics 'MSSSIM' ans 'UQI'. The first column, labeled as 'Original', presents the scores achieved by the original method. In the second column, labeled as '$\lambda_{L1} = 1$' the fusion performance is shown when the weightage of L1 loss is adjusted to 1. The third column displays the results obtained when the KL loss term is omitted. Notably, the quality of the fused image is observed to decrease when the KL loss is omitted, and this degradation is further exacerbated when the λ_{L1} is set to 1. This visually emphasizes the significance of the KL loss term and the weightage of L1 loss in maintaining the quality of the generated fused images.

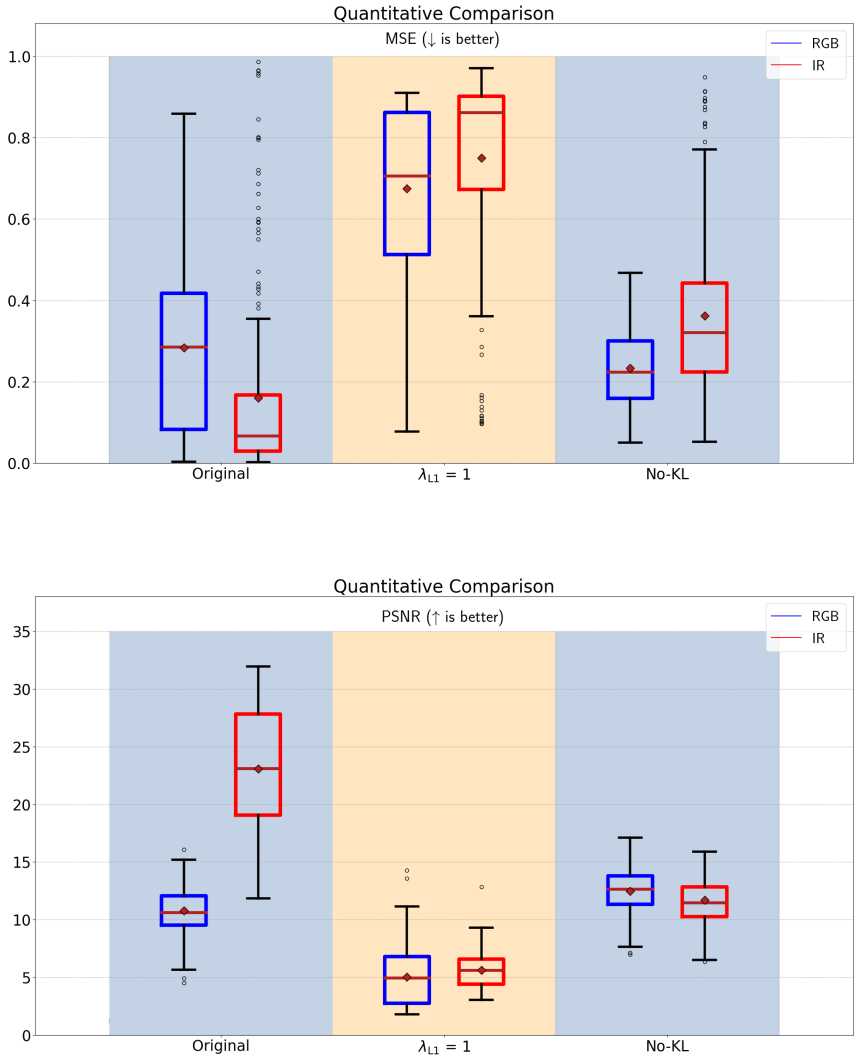

Figure 5.6: Comparison of image fusion performance using metrics 'MSE' and 'PSNR'. The first column, labeled as 'Original', presents the scores achieved by the original method. In the second column, labeled as '$\lambda_{\mathrm{L1}} = 1$' the fusion performance is shown when the weightage of L1 loss is adjusted to 1. The third column displays the results obtained when the KL loss term is omitted. Notably, the quality of the fused image is observed to decrease when the KL loss is omitted, and this degradation is further exacerbated when the λ_{L1} is set to 1. This visually emphasizes the significance of the KL loss term and the weightage of L1 loss in maintaining the quality of the generated fused images.

Impact of Loss Function Variations on Fused Image Quality

An essential component of our proposed method revolves around the integration of L1 loss within the comprehensive loss function, aimed at optimizing the fusion procedure. To assess the significance of this specific loss term, we embarked on an experiment by modifying the weightage attributed to the L1 loss. In particular, we reduced the weightage from its original value of 100 to a much lower value of 1. The rationale behind this manipulation was to ascertain whether diminishing the emphasis on the L1 loss would result in discernible alterations in the quality of the generated fused images. See Figure 5.5 and 5.6.

The outcomes of this experiment unveiled an intriguing insight. By reducing the L1 loss weightage, we observed a distinct deterioration in the comprehensibility of the resultant fused images. In other words, when the weightage was lowered from 100 to 1, the fused images exhibited a decrease in their interpretability and coherence. This phenomenon suggests that the L1 loss component indeed plays a pivotal role in shaping the clarity and visual coherence of the fused images. As such, its significance as a contributing factor to the overall loss function is highlighted, reinforcing the critical importance of its weightage within the fusion process.

In another variation, we examined the consequences of omitting the Kullback-Leibler (KL) loss term while maintaining the weightage of L1 loss at the original value of 100. This omission aimed to explore the repercussions of excluding the KL loss term on the final quality of the fused images. The subsequent analysis, as evidenced by the diverse metric plots presented below, offers valuable insights into the outcomes of these variations and their implications on the fused image quality. See Figure 5.5 and 5.6.

The experimental results shed light on an intriguing phenomenon. When the KL loss was removed from the loss function, we observed a discernible reduction in the quality of the fused images. This reduction was evident across various metrics that assess the image quality, underscoring the importance of the KL loss in enhancing the fusion process. By omitting the KL loss, which serves as a vital bridge between the latent space and the generated image, the model's ability to capture and reproduce intricate visual features was compromised. Consequently, the fused images exhibited a lower level of fidelity and coherence.

Impact of Attention Mechanism on Fusion Quality

The cross attention mechanism serves as a pivotal bridge between thermal and visual data, enabling the model to capture distinct yet complementary information from both modalities. In this ablation variation, we removed the cross attention mechanism entirely from the architecture to evaluate its influence on the fusion process.

The comparison is presented in Figure 5.7, where the right image displays the fused images generated without the attention mechanism, and the left image showcases the fused images produced with the attention module. Notably, the fused image generated without attention exhibited ghost artifacts and the inclusion of visual features, which can be attributed to a lack of emphasis on the essential characteristics of both modalities during the fusion process. In contrast, the fused images generated with the attention module demonstrated a distinct improvement in terms of quality and coherence. The attention mechanism effectively identifies and prioritizes significant features while minimizing the impact of less relevant visual features. This results in a more balanced and comprehensive fused image that accurately represents the salient information present in both thermal and visual modalities.

Figure 5.7: A comparison between fused images generated with and without the attention mechanism is presented. The right side showcases the fused image generated without attention, where the inclusion of visual features leads to the emergence of ghost artifacts. On the left side, the fused image produced with the attention module is displayed, demonstrating that the attention mechanism effectively omits less important visual features, resulting in a more coherent and comprehensible image.

The visual comparison is further supported by quantitative assessments, where metrics such as Mean-Squared Error (MSE), Multi Spectral Structural Similarity Index (MSSSIM), Normalised Mutual Information (NMI), Universal Quality Index (UQI) and Peak Signal-to-Noise Ratio (PSNR) are employed to quantify the improvement in image quality achieved through the attention mechanism. See Figure 5.8 and 5.9.

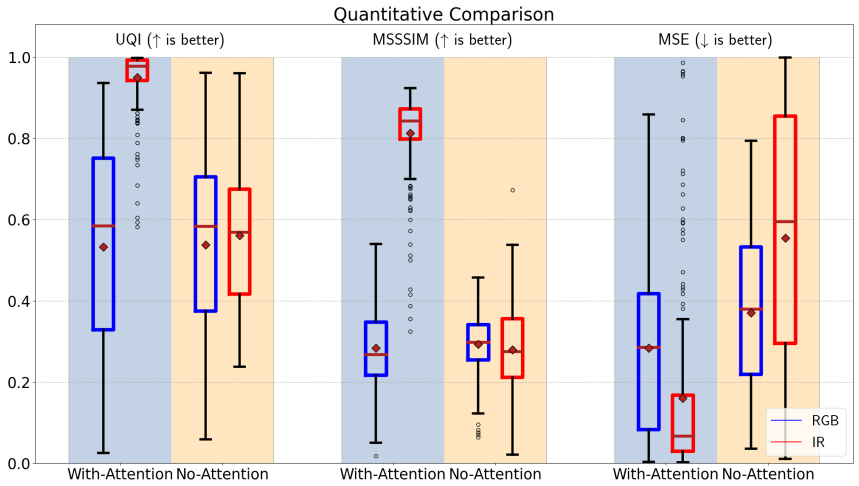

Figure 5.8: The plot illustrates the effect of utilizing the attention mechanism in the image fusion process. In the columns named 'With-Attention', fusion results obtained with the attention mechanism are displayed, while in the columns named 'No-Attention', results without the attention mechanism are presented. Evidently, the quality of the fused images noticeably decreases when the attention mechanism is not employed, underscoring its vital role in enhancing the fusion process by selectively focusing on significant features from both modalities.

The results of this experiment underscore the significance of cross attention. Without cross attention, the fused images exhibited a noticeable misalignment of thermal and visual features, leading to a diminished level of information integration. The absence of cross attention hindered the model's ability to harmoniously combine the distinctive aspects of thermal and visual data, reaffirming its critical role in achieving meaningful fusion.

Synthesis and Implication

Through this comprehensive ablation study, we gain deeper insights into the intricate interactions and dependencies that shape the performance of our proposed method. The variations explored in the study shed light on the delicate balance between L1 Loss and feature alignment, as well as the indispensable role of cross attention in harmonizing modalities. These findings underscore the careful design considerations that underpin the effectiveness of our model and emphasize the interplay of its components in achieving superior fusion outcomes.

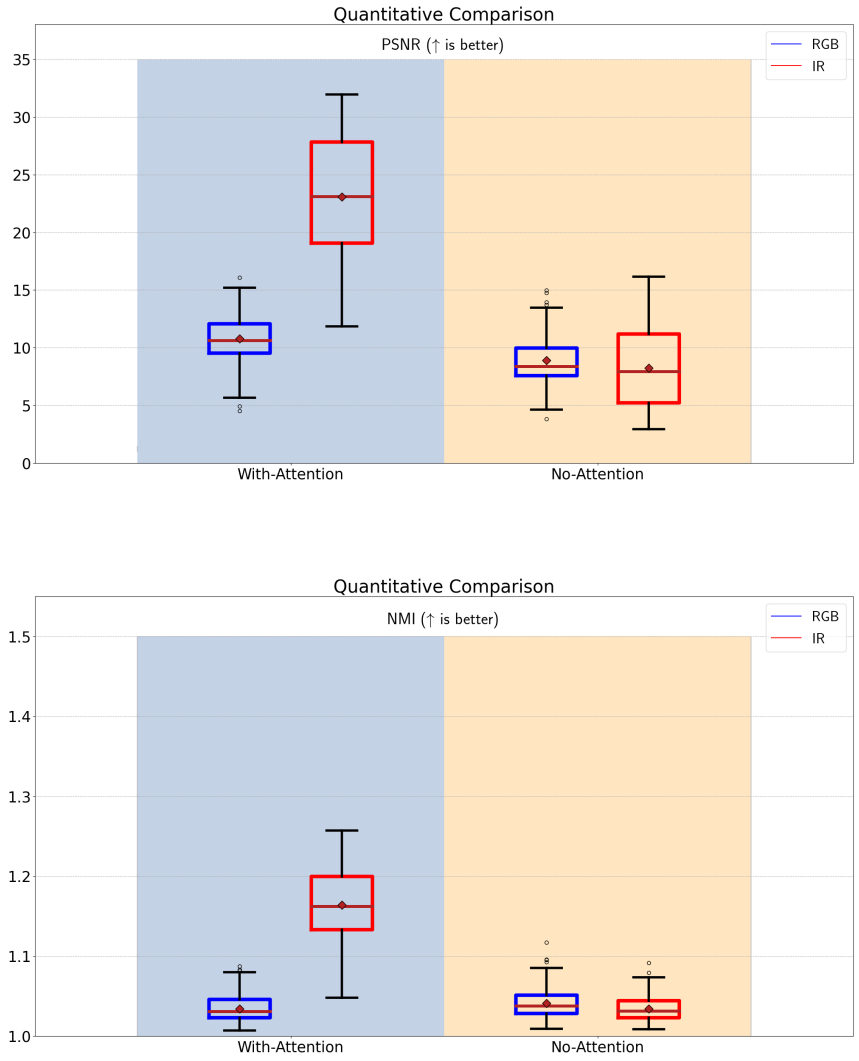

Figure 5.9: The plot illustrates the effect of utilizing the attention mechanism in the image fusion process. In the columns named 'With-Attention', fusion results obtained with the attention mechanism are displayed, while in the columns named 'No-Attention', results without the attention mechanism are presented. Evidently, the quality of the fused images noticeably decreases when the attention mechanism is not employed, underscoring its vital role in enhancing the fusion process by selectively focusing on significant features from both modalities.

Chapter 6

CONCLUSION

In this research, we propose a novel approach for visual-thermal image fusion, which we refer to as the proposed method. Our methodology introduces a sophisticated architecture, namely a Generative Adversarial Network (GAN), with the incorporation of two discriminators and a cross-attention mechanism. This unique combination enables the model to effectively fuse essential information from both visual and thermal modalities without the need for explicit image alignment. By leveraging the capabilities of the GAN framework, our proposed method learns to generate high-quality fused images that provide a comprehensive representation of the scene, combining the distinct strengths of each modality.

The experimental evaluations conducted on the WiSAR dataset highlight the robustness and superior performance of the proposed method, as compared to a state-of-the-art baseline method. Notably, the proposed method exhibits remarkable efficiency in handling misaligned input images, a common issue faced in real-world scenarios where visual and thermal cameras may not be perfectly aligned. The fusion results clearly demonstrate the ability of our model to produce well-fused images that retain the terrain features while effectively extracting bright human silhouettes from the thermal images.

Through a rigorous evaluation using widely-used image quality metrics, including Mean-squared- error (MSE), universal quality index (UQI), multi-scale structural similarity (MSSSIM), normalized mutual information (NMI), and the peak signal-to-noise ratio (PSNR), we quantitatively demonstrate the competence of our proposed method. The comprehensive assessment of image quality from different perspectives validates the effectiveness and reliability of the fusion process performed by the proposed method.

Furthermore, the proposed method offers the distinct advantage of eliminating the need for semantic labeling and ground truth data during training. This characteristic significantly enhances the scalability and adaptability of our model across diverse datasets and real-world applications, making it a practical and versatile solution for various image fusion tasks.

The potential implications of our proposed method extend beyond the realm of image

fusion. By providing a clearer and more complete representation of the environment, our model has the potential to significantly enhance the effectiveness and efficiency of WiSAR missions. Moreover, it has the ability to alleviate the cognitive load on human operators, as the generated fused images offer valuable insights and aid in human detection tasks.

In conclusion, the proposed method demonstrates a pioneering approach to visual-thermal image fusion, leveraging advanced deep-learning techniques to address the challenges of misaligned images and produce superior fused images. The experimental results and comprehensive evaluations highlight the competence and potential of our method, making it a promising solution for various real-world applications, including WiSAR missions and beyond.

BIBLIOGRAPHY

[1] Aggarwal A., Mittal M., and Battineni G. Generative Adversarial Network: An Overview of Theory and Applications. *International Journal of Information Management Data Insights*, 2021.

[2] M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry, and A. H. Gandomi. A Review on Multimodal Medical Image Fusion: Compendious Analysis of Medical Modalities, Multimodal Databases, Fusion Techniques and Quality Metrics. *Computers in Biology and Medicine*, 2022.

[3] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz. Convolutional Autoencoder-Based Multispectral Image Fusion. *IEEE Access*, 2019.

[4] R. Bose, S. Pande, and B. Banerjee. Two Headed Dragons: Multimodal Fusion And Cross Modal Transactions. In *IEEE International Conference on Image Processing*, 2021.

[5] D.* Broyles, C.* Hayner, and K. Leung. WiSARD: A Labeled Visual and Thermal Image Dataset for Wilderness Search and Rescue. In *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, 2022.

[6] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang. A Multimodal Transformer to Fuse Images and Metadata for Skin Disease Classification. *The Visual Computer*, 2022.

[7] C. R. Chen, F. Quanfu, and R. Panda. Crossvit: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *IEEE Int. Conf. on Computer Vision*, 2021.

[8] B. Debaque, H. Perreault, J. P. Mercier, M. A. Drouin, R. David, B. Chatelais, N. Duclos-Hindié, and S. Roy. Thermal and Visible Image Registration Using Deep Homography. In *International Conference on Information Fusion*, 2022.

[9] F. Farahnakian and J. Heikkonen. Deep Learning Based Multi-Modal Fusion Architectures for Maritime Vessel Detection. *Remote Sensing*, 2020.

[10] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, 2017.

[11] P. Isola, J. Zhu, T. Zhou, and A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[12] H. Kaur, D. Koundal, and V. Kadyan. Image Fusion Techniques: A Survey. *Archives of computational methods in Engineering*, 2021.

[13] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng. AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. *IEEE Transactions on Multimedia*, 2021.

[14] C. Liou, W. Cheng, J. Liou, and D. Liou. Autoencoder for Words. *Neurocomputing*, 2014.

[15] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang. Deep Learning for Pixel-Level Image Fusion: Recent Advances and Future Prospects. *Information Fusion*, 2018.

[16] W. Ma, K. Wang, J. Li, S. X. Yang, J. Li, L. Song, and Q. Li. Infrared and Visible Image Fusion Technology and Application: A Review. *Sensors*, 2023.

[17] D. Rao, T. Xu, and X. Wu. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. *IEEE Transactions on Image Processing*, 2023.

[18] S. J. P. Retief, C. J. Willers, and M. S. Wheeler. Prediction of Thermal Crossover Based on Imaging Measurements Over the Diurnal Cycle. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2003.

[19] P. Ricaurte, C. Chilán, C. A. Aguilera-Carrasco, B. X. Vintimilla, and A. D. Sappa. Feature Point Descriptors: Infrared and Visible Spectra. *Sensors*, 2014.

[20] L. Tang, J. Yuan, and J. Ma. Image Fusion in the Loop of High-Level Vision Tasks: A Semantic-Aware Real-Time Infrared and Visible Image Fusion Network. *Information Fusion*, 2022.

[21] A. Toet, J. K. IJspeert, A. M. Waxman, and M. Aguilar. Fusion of Visible and Thermal Imagery Improves Situational Awareness. *Displays*, 1997.

[22] A. Vaswani, N. Shazeer, N. Parmar, J Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Conf. on Neural Information Processing Systems*, 2017.

[23] D. Xiu, Z. Pan, Y. Wu, and Y. Hu. MAGE: Multisource Attention Network With Discriminative Graph and Informative Entities for Classification of Hyperspectral and LiDAR Data. *IEEE Transactions on Geosciences and Remote Sensing*, 2022.

[24] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, Y. Xu, et al. MRI Cross-Modality Image-to-Image Translation. *Scientific reports*, 2020.

[25] J. Zhang, C. Wang, S. Liu, L. Jia, N. Ye, J. Wang, J. Zhou, and J. Sun. Content-Aware Unsupervised Deep Homography Estimation. In *European Conf. on Computer Vision*, 2020.

[26] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE Int. Conf. on Computer Vision*, 2017.

INDEX