

MULTI-MODAL IMAGE REGISTRATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
ENGINEERING AND NATURAL SCIENCES
OF ISTANBUL MEDIPOL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By

Mohamed Sabbah Mohamed Mesbah Elsaeidy

July, 2023

MULTI-MODAL IMAGE REGISTRATION

By Mohamed Sabbah Mohamed Mesbah Elsaedy

28 July 2023

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

Prof. Dr. Bahadır Kürşat Güntürk (Advisor)

Prof. Dr. Hasan Fehmi Ateş

Assistant Prof. Cihan Bilge Gürbüz

Approved by the Graduate School of Engineering and Natural Sciences:

Prof. Dr. Yasemin Yüksel Durmaz

Director of the Graduate School of Engineering and Natural Sciences

I hereby declare that all information in this document has been obtained and presented in accordance with the academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Signature :

Name, Surname: MOHAMED SABBAH
MOHAMED MESBAH ELSAEIDY

ACKNOWLEDGEMENT

Firstly, I want to extend my deepest gratitude to my advisor, Prof. Dr. Bahadır Kürşat Güntürk, for his continuous support, invaluable guidance, unwavering encouragement, and inspiration. Working under his mentorship has been a significant and unforgettable chapter of my life.

I'm immensely appreciative to Prof. Dr. Hasan Fehmi Ateş and Assistant Professor Dr. Cihan Bilge KAYASANDIK for reviewing this dissertation. Their invaluable insights and time devoted to this work have significantly enriched it.

My time at Istanbul Medipol University has blessed me with remarkable friendships, and for this, I'm truly thankful. A heartfelt thanks to everyone who stood by me throughout this journey.

I wish to dedicate this work to my family. Despite the physical distances that separate us, their unwavering love and support have always been my cornerstone. Their influence on my life is immeasurable, and words often fall short when I try to express my profound gratitude. It's certain that without their presence in my life, my achievements would have been minimal.

Lastly, this thesis is supported in part by TÜBİTAK 1003 Project 118E891, entitled “Smart Camera System for Wide Area Surveillance”.

Mohamed Sabbah Mohamed Mesbah Elsaedy

July, 2023

CONTENTS

	<u>Page</u>
ACKNOWLEDGEMENT.....	iv
CONTENTS	v
LIST OF FIGURES	vii
LIST OF TABLES	ix
LIST OF SYMBOLS	x
ABBREVIATIONS.....	xi
ÖZET	xii
ABSTRACT.....	xiii
1. INTRODUCTION	1
2. LITERATURE SURVEY.....	4
2.1 Image Registration.....	4
2.1.1. SuperPoint.....	6
2.1.2. MultiPoint	8
2.2. Image-to-Image Translation	9
2.2.1. Generative Adversarial Networks.....	9
2.2.2. Pix2Pix.....	10
2.3. Visual Transformers	11
2.3.1. Attention Mechanism.....	12
2.3.2. Swin V2 Transformer	13
3. EVALUATION OF METHODS	16
3.1. Datasets.....	16
3.1.1. Multispectral ImagePair Dataset.....	16
3.1.2. VEDAI Dataset	17
3.2. Translating images with Pix2Pix	19
3.3. Generating labels for training	21
3.4. Training losses	22
3.4.1. Pix2Pix losses	22
3.4.2. MultiPoint	23
3.4.2.1. Detector loss	23
3.4.2.2. Descriptor loss	23
3.4.3. Cosine Similarity loss	24
3.5. Evaluation Metrics.....	25
3.5.1. Fréchet inception distance	25
3.5.2. Repeatability	25
3.5.3. Nearest neighbor mean average precision	26
3.5.4. Homography estimation.....	26
3.5.5. Matching accuracy	27
4. OUR APPROACH.....	28
4.1. Image Translation With Pix2Pix.....	28
4.2. SwinV2 Encoder and Cosine Loss.....	29
5. RESULTS AND DISCUSSION	31
5.1. Training Details	31
5.1.1. Image translation using Pix2Pix	31
5.1.2. Swin V2 Transformer	32
5.2. Image Translation Approach Results.....	33
5.3. Visual Transformer and Cosine Loss Approach Results	36
6. CONCLUSIONS AND FUTURE WORK.....	38

BIBLIOGRAPHY.....	40
CURRICULUM VITAE.....	42



LIST OF FIGURES

Figure 1.1: The multispectral image matching problem. The matched IR (left) and optical (right) images due to the different capturing sensors, they significantly differ in appearance as well.	2
Figure 2.1: SuperPoint Network Architecture	7
Figure 2.2: Visual Transformers architecture overview where we divide the image into fixed-size patches, linearly embedding each, adding position embeddings, and feeding the vector sequence to a transformer encoder.....	12
Figure 2.3: Attention Output Example	13
Figure 2.4: a) The Swin V2 Transformer merges image patches into hierarchical feature maps with linear computational complexity, making it suitable for image classification and dense recognition tasks. b) vision Transformers generate low-resolution feature maps and suffer from a quadratic computational complexity because of their global self-attention computation.....	14
Figure 2.5: shifted window approach for self-attention computation. From one layer to the next, window partitioning shifts, allowing new windows to cross previous boundaries and create interconnections.	15
Figure 3.1: a) IR Image, b) Visible Optical Image.....	17
Figure 3.2: Aligned ImagePair from VEDAI dataset where the left is optical and right is infrared.	18
Figure 3.3: U-Net Generator Architecture	19
Figure 3.4: Pixel GAN (Discriminator)	20
Figure 3.5: Homographic adaptation technique used to improve the consistency of keypoints of multi-modal image matching in MultiPoint, where homographies N were applied to increase the number of keypoints detected across both of the images.	21
Figure 3.6: MultiPoint Architecture with Cosine Similarity loss added where x_1 and x_2 are the output vectors from thermal and optical encoders respectively.	24
Figure 4.1: Our new architecture where image translation network applied before feature matching.	28
Figure 4.2: Block diagram of the network and its training process.....	29
Figure 5.1: Training structure of image translation network.....	31
Figure 5.2: Image translation network training loss curve	32
Figure 5.3: Descriptor losses training curves for second approach where it is observed that Swin transformer made Descriptor loss converges more and better than the other models.....	33
Figure 5.4: Visual results for Pix2Pix with U-Net 256 backbone a) thermal image b) generated optical from thermal c) ground truth optical image.	34
Figure 5.5: Qualitative results on the multi-modal <i>ImagePair</i> dataset. Correct matches are highlighted in green. The first row shows results from original methods (left image: optical, right image: IR). The second row shows results obtained using generated optical images in the corresponding methods (left image: optical, right image: generated). Where in a) matching is done using MultiPoint, b) using SURF and in c) using SIFT Matching performance is best when generated image is used in MultiPoint.	35
Figure 5.6: Qualitative results on the multi-modal <i>VEDAI</i> dataset. Green lines highlight correct matches (left image: Optical, right image: IR). a,b) shows results from original MultiPoint encoder w/ and w/o cosine loss. c,d) shows results obtained using Swin transformer w/ and w/o cosine loss. Matching performance is best when Swin transformer w/ Cosine loss is used. b,d) is results w/o cosine loss.....	37

Figure 5.7: Matching with homographic adaptation where a) and c) are the matching with multipoint and swin transformer respectively and b) and d) are their homograph estimation.....	37
--	----



LIST OF TABLES

Table 5.1: FID score results	33
Table 5.2: Detector and descriptor metrics comparison	35
Table 5.3: Keypoint detector and descriptor metrics evaluation of Swin transformer against multipoint encoder.....	36



LIST OF SYMBOLS

$\hat{\mathcal{H}}$	: Homography ground truth
$\mathcal{H}$	: Homography estimated
λ	: Weighting factor
μ	: Mean
ε	: Distance in pixels



ABBREVIATIONS

MLP	: Multi-Layer Projection
CNN	: Convolutional Neural Network
IR	: Infra-Red
SIFT	: Scale-Invariant Feature Transform
SURF	: Speeded up robust features
ORB	: Oriented FAST and Rotated BRIEF
RANSAC	: Random sample consensus
GAN	: Generative Adversarial Network
cGAN	: conditional Generative Adversarial Network
ViT	: Visual Transforms
Swin	: Shifted Window



ÇOK MODLU GÖRÜNTÜ ÇAKIŞTIRMA

ÖZET

Mohamed Sabbah Mohamed Mesbah Elsaeidy

Bilgisayar Mühendisliği, Yüksek Lisans

Tez Danışmanı: Prof. Dr. Bahadır Kürşat Güntürk

Temmuz, 2023

Çoklu modlu görüntü çakıştırma, uzaktan algılama ve görsel navigasyon gibi birçok uygulamanın temelini oluşturur. Tek modlu görüntüler için tasarlanmış geleneksel yöntemler, çoklu modlu görüntü çakıştırma problemlerinde yeterli başarı gösterememektedir. Diğer taraftan, çoklu modlu görüntü eşleştirme için geliştirilmiş teknikler genellikle gerçek zamanlı uygulamaların isterleriyle uyumlu değildir. Bu alandaki boşluk, bu tezin temel motivasyonunu oluşturur: farklı görüntüleme modlarının, özellikle kızılötesi ve optik (görünür aralık) görüntülerin gerçek zamanlı olarak ve yüksek hassasiyette çakıştırılmasını sağlayan yöntemlerin geliştirilmesi. Bu amaçla iki farklı yaklaşım önerilmiştir. Birinci yaklaşımda, kızılötesi görüntülerin optik eşdeğerlerine dönüştürülmesi için görüntüden görüntüye çeviri sağlayan derin yapay sinir ağı uygulanmıştır. Bu dönüşüm, daha gülbüz ve doğru anahtar nokta belirleme ve tanımlamayı sağlayarak görüntü çakıştırmanın başarımını artırır. İkinci yaklaşımda, karmaşık görüntü verilerinin özniteliklerinin etkin bir şekilde çıkarılmasını sağlayan Swin Transformer (Dönüştürücü) kodlayıcıyı kullanılmıştır.

Bunlara ek olarak, kodlanmış görüntülerin arasındaki mesafeyi en aza indirmeyi hedefleyen kosinüs kayıp fonksiyonunun da eklenmesiyle görüntü çakıştırmanın doğruluk ve hassasiyetinin daha da artırılması hedeflenmiştir. Bu sayede çok modlu görüntü çakıştırmada yüksek potansiyelli yeni yaklaşımlar sunulmuştur. Bu teknolojilerin tamamlayıcı doğası, çoklu modlu görüntü çakıştırma ile ilgili zorluklara entegre, çok yönlü ve verimli çözüm sağlayabilir. Kapsamlı simülasyonlarla, her iki yaklaşımın da deneysel doğrulamasını sağlanmış, dönüştürücü tabanlı bir Siyam ağı kullanarak çoklu modlu eşleştirmede belirgin iyileştirmeler gösterilmiştir. Bulgularımız, hizalama doğruluğu ve verimlilikteki önemli gelişmeler sağlamış ve önerdiğimiz yaklaşımların etkinliğini teyit etmiştir. Sonuç olarak, bu tez sadece çoklu modlu görüntü çakıştırma konusundaki mevcut sınırlamalara çözümler sağlamakla kalmaz, aynı zamanda gelecekteki yöntemler için potansiyel olan dönüştürücü tabanlı yaklaşımlara da öncülük etmektedir. Tezin, yenilikçi yaklaşımlar ile görüntü çakıştırma tekniklerinin gelişimine sağlam bir temel oluşturduğunu ve karmaşık bilgisayarlı görüş görevlerini çözmede derin öğrenmenin potansiyelini vurguladığını söyleyebiliriz.

Anahtar sözcükler: Görüntü çakıştırma, çok modlu görüntüler, görüntüden görüntüye çevirme, dönüştürücü ağ.

MULTI-MODAL IMAGE REGISTRATION

ABSTRACT

Mohamed Sabbah Mohamed Mesbah Elsaeidy

MSc in Computer Engineering

Advisor: Prof. Dr. Bahadır Kürşat Güntürk

July, 2023

Multi-modal image registration underpins a multitude of applications, with remote sensing and visual navigation being prominent examples. However, traditional methods, primarily designed for single modality images, falter when confronted with multi-modal registration tasks. Moreover, techniques explicitly developed for multi-modal image registration often lack compatibility with the demands of real-time applications. This gap within the field informs the primary motivation for the present thesis: the development of methods capable of real-time operation and highly precise alignment of different imaging modalities, specifically infrared and optical (visible range) images. In this challenge, we introduce and explore two distinct but promising approaches. Firstly, we implement an image-to-image translation deep network for the conversion of infrared images into their optical counterparts. This conversion facilitates a more robust and accurate keypoint identification and registration, enhancing the overall effectiveness of the image registration process. Secondly, we employ the Swin Transformer encoder, a powerful tool renowned for its superior handling of complex image data. This is further coupled with a cosine loss function to optimize the alignment of images by minimizing angular disparity, further enhancing the accuracy and precision of the registration process. While these approaches are implemented separately in this study, their combined potential represents an exciting prospect for future exploration. The complementary nature of these technologies could provide an integrated, versatile, and efficient solution to the challenges inherent in multi-modal image registration.

Our comprehensive simulation studies provide empirical validation of both methods, demonstrating marked improvements in multi-modal registration using a transformer-based Siamese network. Our findings underscore the significant enhancements in alignment accuracy and efficiency, attesting to the effectiveness of our proposed methodologies. In conclusion, this thesis not only provides solutions to the existing limitations of multi-modal image registration but also pioneers separate transformative approaches with the potential for future integration. By presenting innovative methodologies, this thesis establishes a solid foundation for the evolution of image registration techniques and highlights the potential of deep learning in solving complex computer vision tasks.

Keywords: Image registration, multi-modal images, image to image translation, transformer network.

CHAPTER 1

1. INTRODUCTION

The challenge of matching feature points across varied images remains a fundamental problem in the field of computer vision and image processing, with wide-ranging applications in areas such as image registration, dense image matching, 3D reconstruction, Simultaneous Localization and Mapping (SLAM), Structure-from-Motion (SfM), and remote sensing. These applications require the identification and matching of salient feature points - the centers of image patches - across multiple images of the same scene, even when captured under different poses and lighting conditions. A descriptor performs the vital task of detecting these points and encoding the surrounding patches [1].

The problem intensifies when dealing with multi-modal images, that is, multiple images of the same scene captured using diverse sensors like optical and near-infrared (NIR) imaging. This fusion of images from different modalities provides a wealth of synergistic information that is critical to areas such as remote sensing and medical imaging. However, it presents a complex set of challenges, such as significant appearance variations, including non-monotonic intensity mappings and contrast reversal. In some multimodal scenarios, even the usually consistent image edges and textures might not align correctly.

Over time, a variety of methodologies have been deployed to overcome these hurdles. The earliest solutions revolved around using unimodal descriptors such as SIFT and Daisy, which were extended to cater to the multimodal case. These were later supplanted by strategies leveraging Convolutional Neural Networks (CNNs). These networks learn a data-driven multimodal similarity measure, which has proven beneficial for feature point matching. A diverse set of techniques and methods are used to train these neural networks, including those optimizing a scalar similarity metric using the Hinge Loss

function or binary cross entropy, as well as those focusing on measuring similarity between image patches by optimizing L2-based representations.

Despite the significant advancements brought about by the use of CNNs, the pursuit of a universal feature representation that can effectively describe keypoints of images from different modalities is still ongoing. This challenge becomes increasingly pertinent when considering the computational complexity involved in matching sets of feature points detected in a pair or set of images.

Recent studies have turned to metric learning techniques and deep learning-based algorithms, which have yielded notable improvements in the repeatability of keypoint locations in multimodal images, thereby enhancing matching accuracy. For instance, the LIFT method was introduced as a deep learning-based replacement for SIFT, following the classic detect-then-describe protocol, while the SuperPoint network emerged as a self-supervised interest point detection and description model.

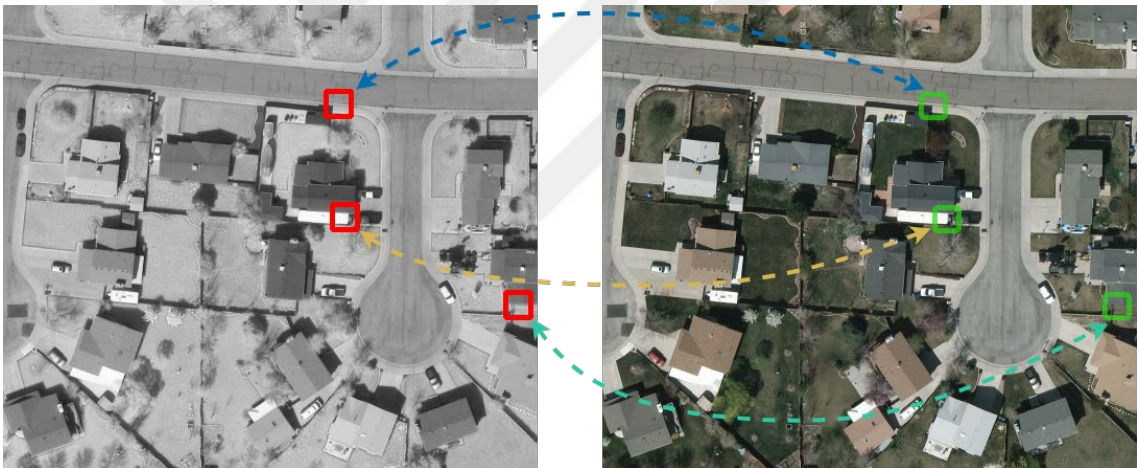


Figure 1.1: The multispectral image matching problem. The matched IR (left) and optical (right) images due to the different capturing sensors, they significantly differ in appearance as well.

Building on these advancements, the MultiPoint network was proposed for multi-modal image matching, utilizing a framework similar to that of the SuperPoint architecture. While these methods have superseded the performance of classical techniques for image matching, there is still significant room for improvement, particularly for multi-modal images.

This study aims to address these challenges by proposing a novel approach to reducing the dissimilarity in keypoint representations across multi-modal images. We leverage two methods first an image-to-image translation method via a deep network model whereby

an infrared image is first translated to an optical image, this translation is performed prior to keypoint extraction and matching, aiming to augment the keypoint matching scores between the images of different modalities. Second by incorporating Swin V2 Transformer encoder as Siamese network which processes both modes of images, divides those images into patches and generates highly discriminative descriptors for each patch using the

self-attention mechanism within the transformer.

We also employ cosine loss in training as the similarity function to generate similar feature representations. These features are subsequently used by the decoder for keypoint and descriptor extraction.

In summary, this research ventures into the domain of image registration, with a specific focus on enhancing the process of feature point matching in multi-modal images. By bridging the existing gaps, we aspire to significantly contribute to numerous computer vision applications, as well as lay the groundwork for future developments in the field.

CHAPTER 2

2. LITERATURE SURVEY

2.1. Image Registration

In the realm of image matching and registration, various methods have been devised, and a key traditional approach includes algorithms such as the Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF) [2], and the Oriented FAST and Rotated BRIEF (ORB) [3]. These algorithms are extensively utilized due to their ability to detect transformation and viewpoint invariant locations, which are vital for many vision tasks.

Fundamentally, these traditional methods perform image registration through a series of steps. Initially, distinctive features or keypoints are detected in each image. For instance, in SIFT, these keypoints are usually corners or blobs that are invariant to scale and orientation. Following detection, a descriptor is computed for each keypoint, capturing its local appearance. Then, keypoints between pairs of images are matched based on the similarity of their descriptors. Once matching keypoints have been identified, a geometrical transformation is estimated that aligns the keypoints in one image to their matches in the other. This is commonly performed using methods like RANSAC to handle outliers in the matches. Finally, this estimated transformation is applied to one image (or its corresponding points or features), effectively registering it to the other image.

However, these traditional methods have limitations, especially when faced with multi-modal image registration. They heavily rely on pixel intensity information for detecting features and extracting descriptors. This reliance poses a problem when dealing with images from different modalities, as the pixel intensity can drastically vary across them. Furthermore, these methods may lack the necessary robustness to handle variations in

image quality, distortions, and varying noise levels inherent in multi-modal images. These constraints limit their performance in multi-modal image registration tasks.

The complexities involved in multi-modal image registration present a unique set of challenges, which are difficult to tackle with traditional algorithms. Predominantly, these methods have a strong dependence on pixel intensity information for the detection of features and extraction of descriptors. However, when it comes to dealing with images from different modalities, this dependence can be a source of significant trouble, given the vast disparities that may exist in pixel intensities across varying modalities. In addition to this, traditional algorithms may not possess the robustness required to tackle the diverse range of image quality, distortions, and noise levels that are often inherent in multi-modal images. These factors substantially hamper their performance in tasks related to multi-modal image registration, leading to less than satisfactory results.

Moreover, traditional image registration techniques, including intensity-based and feature-based methods, are further hindered by a range of issues. For instance, intensity-based methods, which match images based on intensity similarity metrics, often fail in multi-modal scenarios due to the inherent intensity distribution differences across modalities. On the other hand, feature-based methods, which rely on distinctive image features for registration, may struggle with the task due to the absence of shared features across modalities.

Addressing these limitations demands innovative approaches that are capable of not only effectively handling the disparities in pixel intensities and the variations in image quality, distortions, and noise levels, but also bridging the semantic gap between different modalities.

In recent years, the emergence of deep learning-based methods has started to revolutionize the field of image registration. Among these, SuperPoint and MultiPoint have gained considerable attention due to their superior performance across various image-related tasks, including image registration. These methods leverage the ability of deep learning models to learn complex feature representations from extensive data.

However, the challenges of multi-modal image registration are far from being fully conquered. Deep learning methods, despite their advanced capabilities, still struggle with the pronounced contrast in appearance between images from different modalities.

Problems such as spatial alignment disparities, differences in intensity distribution, and presence of artifacts specific to each modality further compound these difficulties.

In essence, despite the significant advancements in the field of image matching and registration, notable challenges remain to be overcome, particularly in the context of multi-modal image registration. The development of methodologies capable of surmounting these challenges will pave the way for a significant expansion in the application spectrum of computer vision.

As mentioned previously SuperPoint and MultiPoint, two deep learning-based methods, perform significantly better than their traditional counterparts on image registration.

2.1.1. SuperPoint

The SuperPoint network, as delineated in **Figure 2.1**, represents a paradigm shift in the field of interest point detection and description. It brings forth a self-supervised learning architecture that is particularly equipped to handle multi-view geometric and photometric transformations, thereby making it ideal for tasks involving multi-view challenges.

The unique SuperPoint network's distinguishing feature lies in its strategy, known as 'homographic adaptation'. This method operates by identifying key features or 'interest points' from multiple homographic transformations of the same image. These interest points are then aggregated to generate a comprehensive and robust set of interest points. By converting these aggregated points into a heatmap, the network generates a 'ground truth' representation for each image [4].

A cornerstone of the SuperPoint architecture is its base detector, known as MagicPoint. MagicPoint is trained primarily using synthetic data and exhibits superior performance when dealing with synthetic images. In comparison to conventional methods, MagicPoint's accuracy and precision outshine, rendering it a powerful tool for synthetic image processing.

However, when confronted with real-world images, MagicPoint reveals a significant limitation. Its performance dips markedly when interpreting these real-world images, failing to replicate the high accuracy and precision it demonstrates with synthetic ones.

This discrepancy between the performance on synthetic and real-world images highlights a substantial area of improvement within the SuperPoint network architecture.

To address this limitation and ensure that SuperPoint's effectiveness extends to real-world scenarios, further enhancements and refinements to the network architecture are warranted. This could involve the integration of more complex feature extraction methods, inclusion of higher levels of noise and distortion that better mimic real-world scenarios or training the network with more diverse and extensive sets of real-world images.

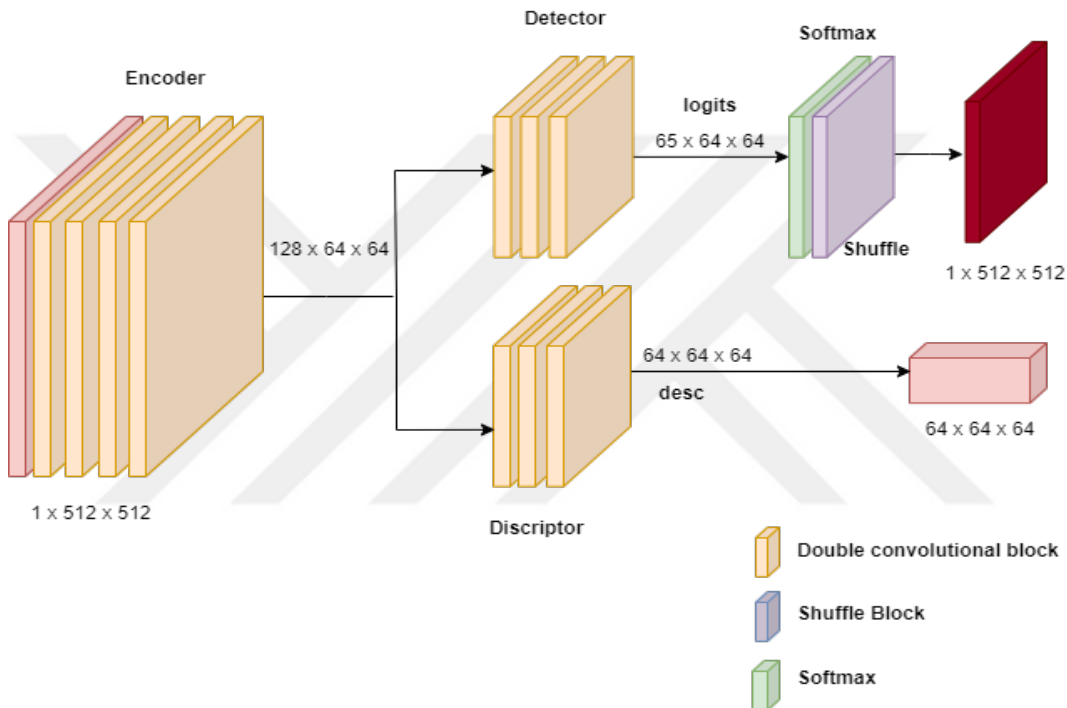


Figure 2.1: SuperPoint Network Architecture

Furthermore, the development of a robust evaluation protocol that goes beyond synthetic benchmarks is necessary. It should account for various types of distortions, lighting conditions, and other variables prevalent in real-world scenarios. By doing so, the network's performance can be accurately measured, thereby identifying its strengths and areas for improvement.

Overall, the SuperPoint network, despite its current limitations, represents a significant advancement in the field of interest point detection and description. Its further refinement and optimization will not only enhance its performance on real-world images but will also broaden its applicability and utility in a variety of computer vision tasks. The evolution

of SuperPoint, thus, continues to be an exciting area of exploration in the realm of computer vision.

2.1.2. MultiPoint

On the other hand, MultiPoint [1] generalizes SuperPoint’s framework to solve the multi-spectral image matching problem. MultiPoint introduces a novel approach for aligning multi-spectral images using homographic adaptation to enhance cross-spectral consistency of interest point detections.

The homographic adaptation method used in the multi-point offers a powerful solution to address the challenges of interest point detection and description in real-world scenarios. By leveraging multiple homographies to warp the original image, the method creates a diverse set of viewpoints, effectively simulating different perspectives of the same scene.

This process of viewpoint diversification is vital, as it allows the interest point detector to learn and generalize across various image transformations and scene variations. By extracting keypoints from each homography-applied version of the image and combining them, the method generates a comprehensive set of interest points. These aggregated keypoints form the ground truth keypoint heatmap, providing valuable information for training the interest point detector and descriptor in a self-supervised manner. The self-supervised learning approach is advantageous as it reduces the need for large annotated datasets, which can be costly and time-consuming to create. Instead, the homographic adaptation method leverages the inherent geometry of the scene to generate the ground truth keypoint annotations, making the training process more efficient and scalable.

During the multi-point training, the trained interest point detector and descriptor are rigorously evaluated using real-world images. This evaluation serves to validate the effectiveness of the homographic adaptation technique in enhancing the model's performance across different scenes, lighting conditions, and viewpoints. By testing the model on challenging and diverse datasets, researchers can gain insights into the strengths and limitations of the method and identify potential areas for further improvement.

Furthermore, the adaptability of the homographic adaptation method extends beyond interest point detection and description tasks. Its ability to handle multi-view challenges

makes it a valuable asset in various computer vision applications, such as image matching, 3D reconstruction, and visual localization.

In conclusion, the homographic adaptation method applied in the multi-point not only advances the field of interest point detection and description but also demonstrates the potential of self-supervised learning and viewpoint diversification in addressing real-world challenges. As researchers continue to refine and optimize this technique, it holds the promise of contributing to more robust and versatile computer vision systems in the future.

2.2. Image-to-Image Translation

2.2.1. Generative Adversarial Networks

GANs (Generative Adversarial Networks) have been a game changer in the field of machine learning. Their ground-breaking ability to generate high-fidelity and coherent synthetic data has far-reaching implications in a variety of fields. GANs provide a mechanism for training complex generative models that would be difficult, if not impossible, to achieve with traditional methods such as maximum likelihood estimation or variational approximations [5]. GAN architecture consists of two competing models that interact with each other: the generator and the discriminator. Starting with a random noise vector as input, the generator generates synthetic data. The discriminator, on the other hand, must distinguish between real-world data and synthetic data generated by the generator [6]. The rivalry between these two entities creates a distinct learning dynamic. The generator aims to generate data that is so realistic that the discriminator cannot tell it apart from real-world data, enhancing its generative capabilities. The discriminator, on the other hand, aims to reduce the classification error between synthetic and real data, thereby improving its discriminative abilities. In GANs, the interaction between the generator and the discriminator is the essence of adversarial learning. The generator is driven to produce increasingly high-quality synthetic data as the discriminator strives to improve its classification accuracy. This reciprocal learning cycle results in both networks improving their performance over time, resulting in more sophisticated synthetic data generation. In essence, GANs have significantly expanded the boundaries of what is possible in the generation of synthetic data through their adversarial learning framework. The ability to generate high-quality synthetic data has a wide range of potential applications, including image synthesis, text generation, and novel drug discovery.

2.2.2. Pix2Pix

Pix2Pix is a type of Conditional Generative Adversarial Network (cGAN), and because it introduces a well-defined control on the process of data generation, it is an appropriate choice for specific tasks such as translating images from one format to another [5].

An image is fed into the Pix2Pix model's generator, which is referred to as 'U-Net,' and it then makes an attempt to output a corresponding image that is rendered in a different art style. The U-Net generator uses an encoder-decoder structure with skip connections between mirrored layers in the encoder and decoder stacks. These skip connections improve the quality of the output image by allowing low-level information to get around the bottleneck that is located at the core of the network [6].

The discriminator, also known as the 'PatchGAN' classifier, works on image patches in order to determine whether or not each patch within an image is authentic. It takes into account $N \times N$ patches from both the synthesized image and the target image and attempts to determine whether or not the patches are genuine or generated [6]. PatchGAN, in contrast to full image discriminators, employs a reduced number of parameters, runs more quickly, and is able to model high frequencies accurately.

The loss function that Pix2Pix uses is an important differentiating factor that the program possesses. The GAN loss is combined with a more conventional loss known as the L1 loss in Pix2Pix. The sharpness of the generated output can be guaranteed thanks to the GAN loss, while the structural similarity of the generated image and the target image can be protected thanks to the L1 loss [5]. This particular combination of loss functions is helpful in producing synthetic images that are both high-quality and aware of their context.

It has been demonstrated that the Pix2Pix network is capable of performing a wide range of image translation tasks, including the transformation of semantic label maps into photo-realistic images, the conversion of Google maps to Google earth, and even applications such as colorization, edge-enhancement, and super-resolution [5].

Pix2Pix has some shortcomings, despite the fact that it has some strong points. It is necessary to have paired training data, which can be challenging to obtain for certain tasks. It also does not model multi-modal outputs, which means that given the same input, the model always produces the same output. This may not be desirable in certain circumstances, so it is a limitation of the model.

Pix2Pix, which is a conditional GAN model, possesses a number of unique qualities that distinguish it as an excellent option for performing image-to-image translation work. Its performance is one of a kind and highly effective, thanks in large part to its architecture as well as its loss function and approach to discriminator design.

2.3. Visual Transformers

Visual Transformers, an intriguing amalgamation of advances in natural language processing and computer vision, have emerged as a game-changer in the sphere of complex computer vision tasks [7]. Their genesis owes much to the noteworthy success of Transformer models in the realm of natural language processing. Offering a sharp deviation from the standard convolutional neural networks (CNNs), which typically employ a rigid approach in handling visual data, Visual Transformers bring to the table a more adaptive, flexible strategy that has the potential to revolutionize the field.

Visual Transformers are premised on the notion of a holistic understanding of the interrelationships among various regions of an image. Unlike conventional CNNs that emphasize local features using convolution operations within a limited receptive field, Visual Transformers herald a significant paradigm shift. They harness the power of self-attention mechanisms to encapsulate long-range dependencies and interactions that are prevalent throughout the image [7]. This shift in focus allows the model to move beyond the myopic local context and consider the entire image scene, integrating meaningful information from disparate regions when making predictions or when tasked with feature extraction.

At the core of Visual Transformers lie the self-attention layers. These layers are tasked with interpreting and processing an input sequence composed of image patches or tokens. They compute attention weights, a quantifiable reflection of the importance of each patch relative to others [8]. These weights serve as a guide, instructing the model to dynamically shift focus to regions of interest and effectively manage the influence of different patches during computational processes.

Visual Transformers are characterized by their unique architecture, comprising a succession of self-attention layers. These are often supplemented with feed-forward neural networks that carry out non-linear transformations. This integrated attention mechanism fosters effective information exchange and engenders rich contextual

understanding among the different patches. As a result, Visual Transformers are capable of generating robust and comprehensive image representations that reflect a deep understanding of the image content.

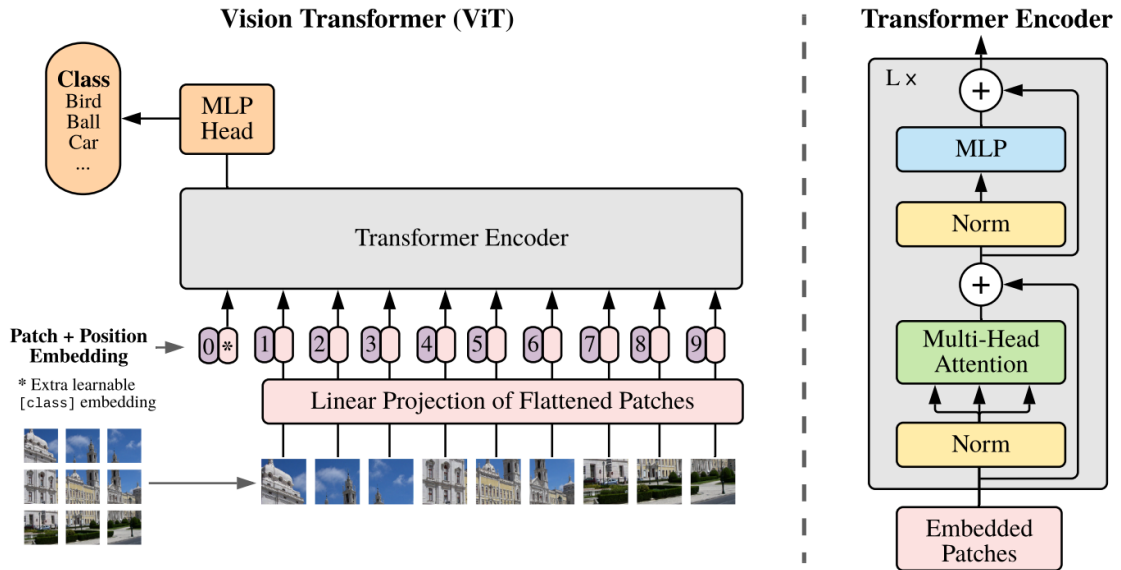


Figure 2.2: Visual Transformers architecture overview where we divide the image into fixed-size patches, linearly embedding each, adding position embeddings, and feeding the vector sequence to a transformer encoder.

Figure 2.2 provides a graphical representation of the Visual Transformers architecture, a helpful aid for understanding how different components interact and contribute to the unique capabilities offered by these models. It serves as a testament to the promise held by Visual Transformers in pushing the boundaries of what's possible in the field of computer vision. Their adaptive approach to understanding visual data makes them an impressive tool in the toolbox of modern computer vision tasks, bridging the divide between transformer-based models and vision-based models. The remarkable success of Visual Transformers across a wide array of tasks underscores their potential in shaping the future of the field.

2.3.1. Attention Mechanism

Attention mechanisms serve as a fundamental component of Visual Transformers, facilitating the modeling of spatial relationships and capturing contextual dependencies within the image [8]. By dynamically adjusting the importance of different regions, attention mechanisms enable the model to selectively attend to relevant image content and suppress irrelevant information.

The attention mechanism operates through a series of steps. Firstly, it calculates the attention weights by comparing the compatibility or similarity between query and key vectors associated with each patch. This step determines how much attention each patch should receive from other patches. Subsequently, the attention weights are applied to the corresponding value vectors, which contain the information associated with each patch. The weighted sum of the values produces the attended representation, capturing the influence of different patches on the final output. The self-attention mechanism can be enhanced through various modifications, such as multi-head attention. This technique employs multiple sets of queries, keys, and values, enabling the model to capture different types of relationships and attend to diverse aspects of the image simultaneously. Additionally, positional encodings are often incorporated to provide spatial information to the model, ensuring it can differentiate between patches based on their relative positions, an illustration is given in **Figure 2.3**.

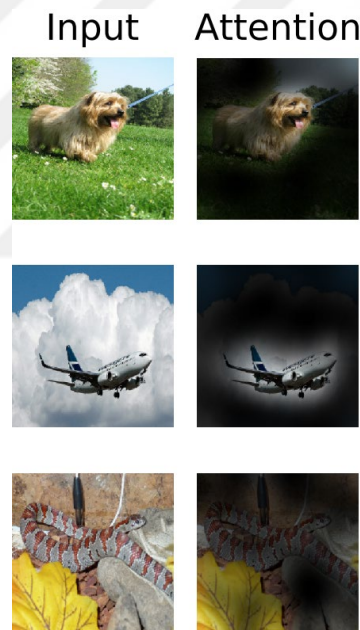


Figure 2.3: Attention Output Example

2.3.2. Swin V2 Transformer

Substantial increases in computational complexity. With linear computational complexity relative to input resolution and the number of channels in the transformer layers, the Swin V2 Transformer facilitates the efficient training of large models on high-resolution images—an undertaking that frequently challenges CNN-based approaches.

The Swin V2 Transformer [9] is a recent advancement in the field of computer vision, offering a promising alternative to conventional convolutional neural networks (CNNs).

Building upon the success of the Vision Transformer (ViT) [7], which employs the transformer architecture to reshape input images into a sequence of flattened patches for image classification tasks, the Swin V2 Transformer introduces several key improvements to enhance its suitability for a wider range of computer vision tasks. One of the primary innovations of the Swin V2 Transformer lies in its hierarchical feature representation. By dividing input images into non-overlapping patches and applying a shifted window-based approach, the Swin V2 Transformer efficiently aggregates both local and global context. This is accomplished by partitioning input patches into smaller windows and utilizing a local self-attention mechanism within each window. Neighboring windows then overlap through shifting, ensuring the capture of global context. Another notable characteristic of the Swin V2 Transformer is its capacity to scale up to larger models without incurring.

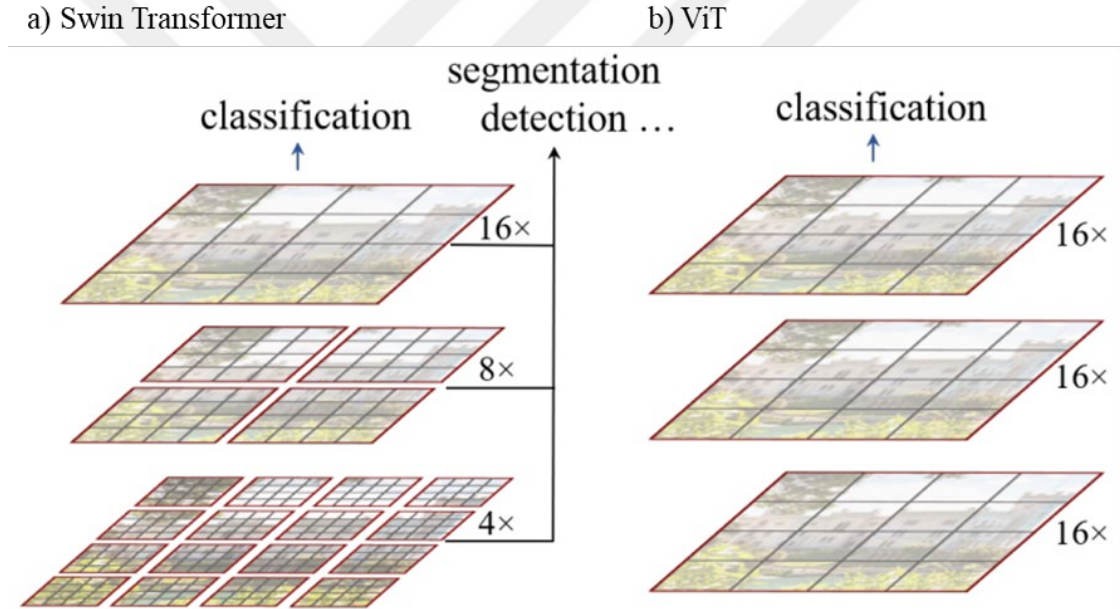


Figure 2.4: **a)** The Swin V2 Transformer merges image patches into hierarchical feature maps with linear computational complexity, making it suitable for image classification and dense recognition tasks. **b)** vision Transformers generate low-resolution feature maps and suffer from a quadratic computational complexity because of their global self-attention computation.

The Swin V2 Transformer, depicted in **Figure 2.4a**, produces a tiered representation by beginning with small-sized patches (grey outline) and progressively amalgamating neighboring patches within the deeper Transformer layers. This hierarchical organization of feature maps allows the Swin V2 Transformer model to easily integrate advanced methods for dense prediction such as Feature Pyramid Networks (FPN) [10] or U-Net [6]. Linear computational complexity is maintained by calculating self-attention within non-

overlapping windows that separate an image (red outline). Given the constant number of patches per window, complexity becomes linear relative to the image size. These advantages position the Swin V2 Transformer as a versatile backbone for various vision tasks, unlike earlier Transformer-based architectures [7] that generate feature maps of a single resolution and exhibit quadratic complexity.

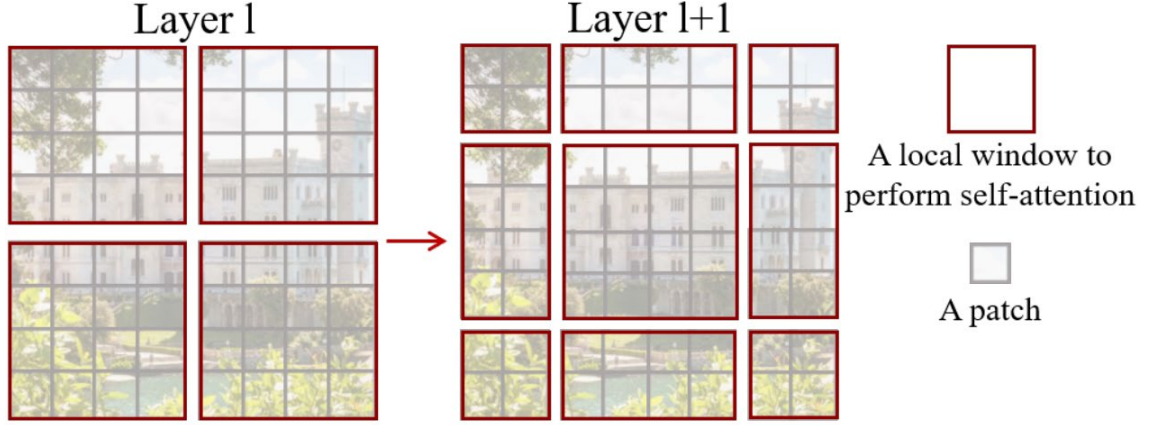


Figure 2.5: shifted window approach for self-attention computation. From one layer to the next, window partitioning shifts, allowing new windows to cross previous boundaries and create interconnections.

A defining aspect of the Swin V2 Transformer's design is the transition of the window partition between successive self-attention layers, as demonstrated in **Figure 2.5**. These shifted windows create links between the windows of the preceding layer, which significantly bolsters the model's capacity. This approach is also beneficial in terms of real-world latency: all query patches within a window share the same key set, optimizing memory access in hardware. This contrasts with prior sliding window based self-attention methodologies [11], [12] that exhibit low latency on standard hardware due to unique key sets for different query pixels.

In summary, the Swin V2 Transformer presents several advantages over traditional CNN encoders. It offers efficient local and global context aggregation via shifted window-based self-attention, scalability to large models and high-resolution images due to its linear computational complexity, and enhanced performance across various computer vision tasks in comparison to CNN-based methods. Incorporating the encoder part of Swin Transformer into the encoder segment of the MultiPoint network seeks to capitalize on these benefits for multispectral image registration tasks. We claim that this modification improves the registration process's accuracy and robustness while maintaining computational efficiency.

CHAPTER 3

3. EVALUATION OF METHODS

3.1. Datasets

In our work we have dealt with two datasets. First the Multispectral (Optical-Thermal) Image Pair dataset that multipoint was applied on where we applied our image translation approach then performed image registration. Second the Vehicle Detection in Aerial Imagery (VEDAI) Dataset where we applied our second approach of visual transformers.

3.1.1. Multispectral ImagePair Dataset

The Multispectral ImagePair Dataset is uniquely composed of multispectral images, specifically, pairs of optical and thermal images. This dataset, meticulously curated, is the result of ten distinct UAV flights carried out at different times throughout the day. The time variance in data collection aimed to introduce diversity in lighting conditions and environmental features, thus making the dataset more robust and suitable for a wide range of applications [1].

The optical part of the image pairs was captured using a UI-5261SE Rev. 4 RGB camera, known for its high resolution and quality, equipped with a 16 mm focal length lens. This choice of camera and lens configuration ensures a wide field of view and high-quality image capture.

For the thermal images, a FLIR TAU2 19 mm camera was utilized, which operates within the 7.5–13.5 μm spectral band. The selection of this specific thermal camera takes advantage of its unique capabilities in capturing minute temperature variations, offering an insightful view of the thermal properties of the scene under inspection.

The images from the dataset present an interesting challenge due to their multimodal nature, making the dataset especially suitable for algorithm development, model training and testing, and comparative analyses across a spectrum of computer vision tasks. Importantly, the data's multispectral quality allows for a more comprehensive understanding of environmental conditions, which single-modal datasets may not fully capture. Future enhancements to this dataset will aim to diversify the environmental conditions and potentially introduce additional imaging modalities.

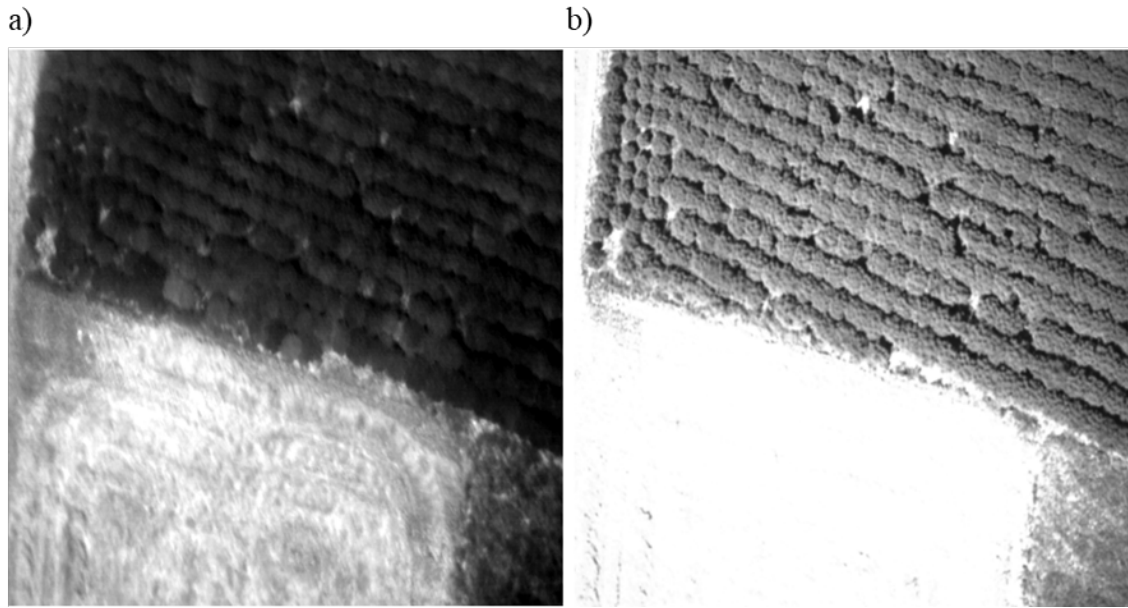


Figure 3.1: a) IR Image, b) Visible Optical Image

The dataset is split across flights into a training and test set. The dataset contains 9000 Image pairs for training and 4391 image pairs for testing. A sample of image pairs shown in **Figure 3.1** where **Figure 3.1a** is the infrared image and **Figure 3.1b** is the visible optical image.

3.1.2. VEDAI Dataset

The VEDAI (Vehicle Detection in Aerial Imagery) dataset serves as a valuable resource for vehicle detection in aerial images [10]. It comprises high-resolution aerial images captured from various sensors and platforms, including optical and infrared imagery. The dataset offers a diverse range of environments, covering urban areas, highways, and rural landscapes, enabling researchers to explore vehicle detection algorithms across different scenarios.

The dataset is typically split into subsets 979 image pair for training and 266 image pair for testing, ensuring unbiased evaluations and facilitating reliable comparisons where the [13].

The VEDAI dataset finds applications in traffic monitoring, urban planning, surveillance, and defense, among others. By leveraging this dataset, we were able to evaluate image registration algorithms, contributing to the advancement of systems relying on aerial imagery.



Figure 3.2: Aligned ImagePair from VEDAI dataset where the left is optical and right is infrared.

In conclusion, the VEDAI dataset is a valuable resource for Aligned Pairs in aerial imagery, offering diverse scenarios and predefined data splits [13].

3.2. Translating images with Pix2Pix

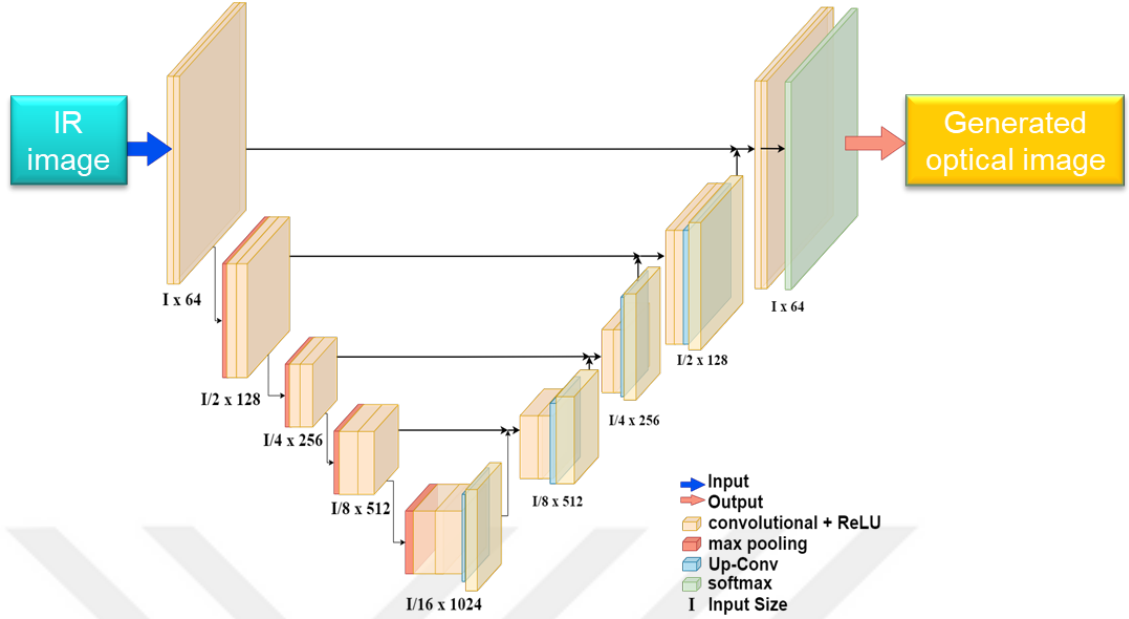


Figure 3.3: U-Net Generator Architecture

The Pix2Pix network, using a U-Net generator and a Pixel GAN [14] discriminator, facilitates feed-forward operations for efficient image-to-image translation. The process initiates with an input image, such as an infrared image, which is fed into the U-Net generator shown in **Figure 3.3**. The input traverses through the encoder section of the U-Net, systematically reducing the spatial dimensions and increasing the depth, thus enhancing feature abstraction. Employing convolutional layers, the encoder efficiently extracts feature maps from the input image. The subsequent decoding phase of U-Net involves a series of up sampling operations applied to these feature maps. The skip connections between the encoder and decoder layers ensure retention and incorporation of spatial information from earlier layers, resulting in accurate image detail localization and reconstruction. This step generates an output image which effectively mirrors the desired translation of the input, say an optical image equivalent of the input infrared image.

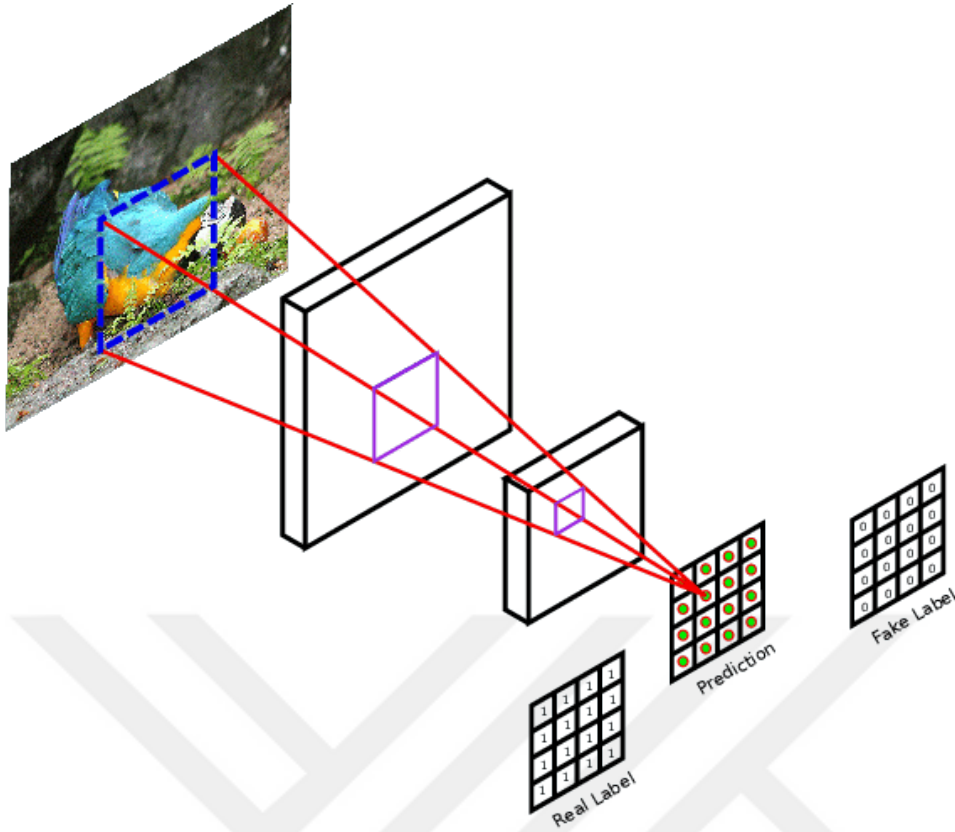


Figure 3.4: Pixel GAN (Discriminator)

The next phase involves evaluation by the Pixel GAN discriminator illustrated on **Figure 3.4**. The only difference between Patch GAN and Pixel GAN is that Pixel GAN classifies each pixel, while Patch GAN classifies a 70x70 Patch. The generated image, coupled with real images from the target domain, is passed to the discriminator for assessment. The discriminator scrutinizes the authenticity of each pixel in the generated image by comparing it with corresponding pixels in the real image. The aim is to classify individual pixels of the generated image as either 'real' (originating from an actual image) or 'fake' (from a synthetic image). Upon this evaluation, the discriminator computes the error or loss, which is subsequently back-propagated through the network. This step is vital in adjusting the parameters of both the generator and the discriminator, enabling the generator to create more realistic images that would challenge the discriminator in differentiating them from real images. This optimization enhances the overall performance of the image-to-image translation task [5]. The entire process is iterated over each epoch during training until the network meets a specific stopping criterion.

3.3. Generating labels for training

Labels were generated for training of the detector and descriptor networks using MultiPoint’s homographic adaptation method was used.

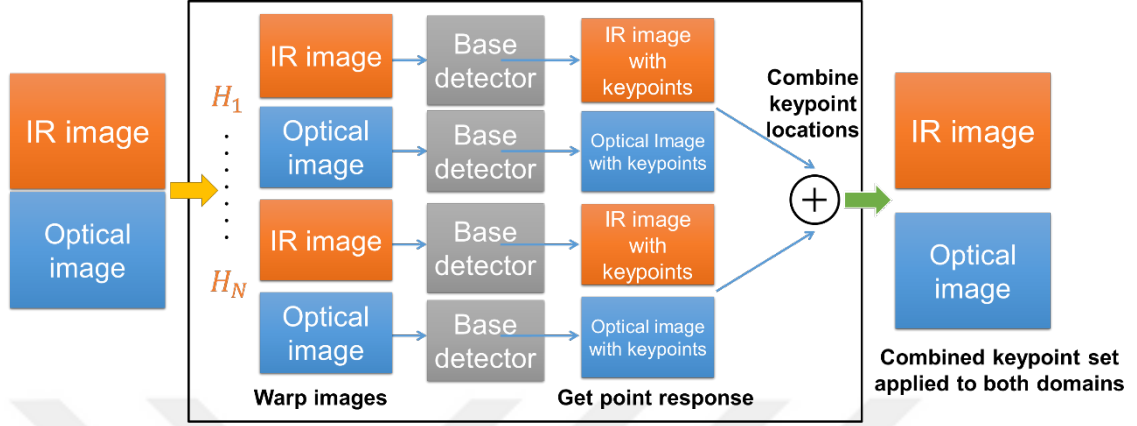


Figure 3.5: Homographic adaptation technique used to improve the consistency of keypoints of multi-modal image matching in MultiPoint, where homographies N were applied to increase the number of keypoints detected across both of the images.

Instead of analyzing individual frames, the method applies homographies to pairs of pre-aligned images, creating a new viewpoint. The interest point detector is then applied to both images, and the intersection of the resulting heatmaps is determined for consistent interest points. The approach utilizes randomly sampled homographies and decomposes transformations for better results. The proposed multi-spectrum-aware scoring function aggregates the heatmaps to generate reliable cross-spectral interest point labels, promising improved performance in various engineering applications.

It uses SURF as a base detector, as it spreads interest points uniformly in the multi-spectral domain [1]. Homographic adaptation as in **Figure 3.5** generates a joint keypoint map for aligned optical-thermal image pairs by matching the corresponding SURF keypoints in the two images, 100 homographies were applied. This approach is more suitable for multispectral image alignment, which is a challenging task for traditional methods like SIFT, SURF, and ORB [3].

3.4. Training losses

3.4.1. Pix2Pix losses

In the Pix2Pix network, a type of conditional Generative Adversarial Network (GAN), training incorporates a dual loss function approach, consisting of adversarial loss and L1 loss (also termed reconstruction or pixel-wise loss) [5].

The adversarial loss follows the typical GAN training regime, presenting a game-theoretic optimization dilemma where the generator aims to maximize the probability of the discriminator incorrectly classifying an image, while the discriminator's goal is to minimize this same probability. This is formalized as:

$$L_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,y}[\log(1 - D(x, G(x, z)))] \quad (3.1)$$

with G representing the optimal generator output, D the discriminator output, x the input image, y the target image, z a noise vector, and E signifying the expectation [5].

In this adversarial setting, the generator G is working to minimize this objective while at the same time, the discriminator D is striving to maximize the same objective.

The L1 loss, on the other hand, is incorporated to ensure pixel-level similarity between the generated and target images. Defined as the mean absolute error between these images, the L1 loss encourages the generator to generate images closely resembling the ground truth on a pixel basis:

$$L1(G) = E_{x,y,z} [||y - G(x, z)||_1] \quad (3.2)$$

This loss function is typically weighted by a scaling factor (lambda, λ) and added to the GAN loss function, with λ regulating the relative importance of the L1 loss compared to the adversarial loss [5].

The combined objective function minimized by the Pix2Pix network [5] is hence:

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (3.3)$$

This dual loss approach ensures that while the adversarial loss guarantees the output's indistinguishability from real images, the L1 loss assures pixel-to-pixel correspondence between the output and target images.

3.4.2. MultiPoint

3.4.2.1. Detector loss

In the process of detecting interest points, each pixel within the output is linked to the probability of that specific pixel in the input representing a "point of interest". Conventionally, the architectural design of networks for dense prediction entails the use of an encoder-decoder pair. In this setup, the spatial resolution is reduced via pooling or strided convolution methods, which is later brought back to full resolution through upconvolution operations, as seen in architectures like SegNet [15]. However, these upsampling layers often result in increased computational load and the potential for unwanted checkerboard artifacts [16]. To mitigate these issues, our design incorporates a distinct decoder within the interest point detection head, aimed at reducing the model's computation.

The detector head of the interest point calculates X , which belongs to a space of dimension $R^{H_c \times W_c \times 65}$, outputting a tensor of size $R^{H \times W}$. The 65 channels each correspond to localized, non-overlapping 4×4 grid regions of pixels, with an additional "non-interest point" dustbin. After executing a channel-wise softmax, the 65th channel is the dustbin dimension, is discarded, followed by a reshape operation from $R^{H_c \times W_c \times 64}$ to $R^{H \times W}$.

$$\mathcal{L}_p(X, Y) = \frac{1}{H_c W_c} \sum_{h=1}^{H_c} \sum_{w=1}^{W_c} l_p(x_{hw}; y_{hw}), \quad (3.4)$$

where

$$l_p(x_{hw}; y) = -\log \left(\frac{\exp(x_{hwy})}{\sum_{k=1}^{65} \exp(x_{hwk})} \right). \quad (3.5)$$

The loss function for the interest point detector, represented as $\mathcal{L}_p$, employs a fully convolutional cross-entropy loss distributed across cells defined as $x_{wh} \in X$. The corresponding ground-truth labels for interest points are collectively referred to as Y , and their individual entries are denoted as y_{wh} .

3.4.2.2. Descriptor loss

The descriptor loss is calculated for all pairs of descriptor cells, where $\mathbf{d}_{hw} \in \mathcal{D}$ from the first image and $\mathbf{d}'_{h'w'}$ belongs to $\mathcal{D}'$ from the second image. The homography-guided match between the cell (h, w) and the cell (h', w') can be expressed as follows:

$$S_{hwh'w'} = \begin{cases} 1, & \text{if } \|\widehat{\mathcal{H}p_{hw}} - p_{h'w'}\| \leq 4, \\ 0, & \text{otherwise} \end{cases}, \quad (3.6)$$

where p_{hw} stands for the position of the center pixel in the (h,w) cell, and $\widehat{\mathcal{H}p_{hw}}$ stands for multiplying the cell location p_{hw} by the homography $\mathcal{H}$ and dividing by the last coordinate, as is typically done when converting between homogeneous and Euclidean coordinates. We use S to represent the entire set of correspondences for a given pair of images. We also include a weighting term λ_d to compensate for the fact that there are more negative correspondences than positive correspondences. We employ a hinge loss with positive and negative margins, m_p and m_n . Overall descriptor loss is defined as follows:

$$\mathcal{L}_d = \frac{1}{(H_c W_c)^2} \sum_{h=1}^{H_c W_c} \sum_{w'=1}^{H_c W_c} l_d(\mathbf{d}_{hw}, \mathbf{d}'_{h'w'}; S_{hwh'w'}), \quad (3.7)$$

where

$$l_d(\mathbf{d}, \mathbf{d}'; S) = \lambda_d * s * \max(0, m_p - \mathbf{d}^T \mathbf{d}') + (1 - s) * \max(0, \mathbf{d}^T \mathbf{d}' - m_n). \quad (3.8)$$

3.4.3. Cosine Similarity loss

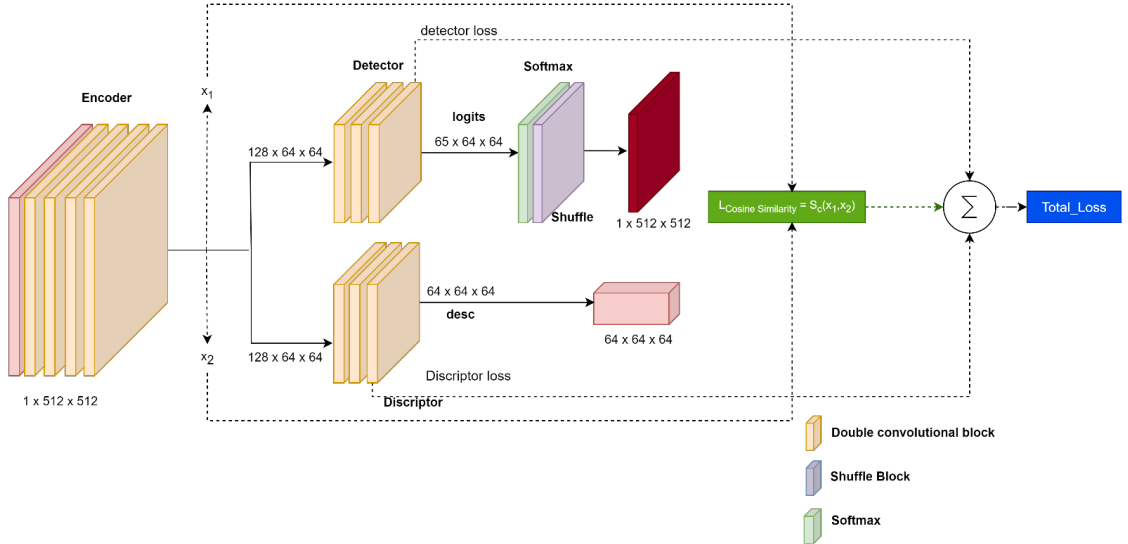


Figure 3.6: MultiPoint Architecture with Cosine Similarity loss added where x_1 and x_2 are the output vectors from thermal and optical encoders respectively.

As shown in **Figure 3.6** MultiPoint has a shared encoder followed by Detector and Descriptor and their mechanism to work is previously explained in **Section 2.1.2**.

$$L_{\cos}(x_1, x_2) = 1 - \frac{x_1 \cdot x_2}{|x_1||x_2|}. \quad (3.9)$$

We contributed to the architecture of MultiPoint and SuperPoint by adding cosine similarity loss in (E 3.9) to the output of the descriptor then feeding network thermal and optical pairs individually through two feedforward operations after that that cosine loss is calculated and added to total loss.

3.5. Evaluation Metrics

3.5.1. Fréchet inception distance

We evaluate our image translation model performance based on the “Fréchet inception distance” (FID) [17].

The "inception score" is a well-performing way to evaluate the performance of GANs [18]. The generated samples are put into an ImageNet-trained inception model.

The enhanced inception score is used to extract vision-relevant characteristics.

The Gaussian distribution is the highest entropy distribution for known covariance and mean. Therefore the outputs of the last max-pooling layers are modeled by multidimensional Gaussian distributions. The difference between two Gaussians of real-world and synthetic images is measured by the Fréchet distance which also known as Wasserstein-2 distance [19].

The final calculation is done by the Wasserstein-2 distance of the two Gaussians through the following equation:

$$d^2 = \|\mu_1 - \mu_2\|^2 + \text{Tr}(C_1 + C_2 - 2 \times \sqrt{C_1 \times C_2}). \quad (3.10)$$

Where μ_1, μ_2 are the means, C_1, C_2 are the covariances of synthetic data and real-world data (i.e. real optical images), respectively.

3.5.2. Repeatability

Repeatability is a key metric in evaluating the performance of keypoint detection algorithms, particularly in the context of image matching and registration tasks. It measures how consistently the same physical point in the real world, appearing in different images, is detected as a keypoint.

Given two images of the same scene captured from different viewpoints, a repeatability score estimates the likelihood that a keypoint detected in the first image will also be detected in the second image. High repeatability implies that an algorithm can robustly detect the same points under different viewing conditions, which is desirable for applications like feature matching, panorama stitching, or 3D reconstruction [20].

To compute repeatability, one needs to establish correspondences between keypoints in the two images. This process often involves applying a known geometric transformation (like a homography in case of planar scenes or wide-baseline stereo pairs) to the keypoints in the first image and then determining how many of these transformed points fall within a certain distance of a keypoint in the second image.

Mathematically repeatability is obtained through the following equation:

$$Rep = \frac{1}{N_1 + N_2} \left(\sum_i Correct(x_i) + \sum_j Correct(x_j) \right), \quad (3.11)$$

where

$$Correct(x_i) = \left(\min_{j \in \{1, \dots, N_2\}} \|x_i - \hat{x}_j\| \right) \leq \epsilon. \quad (3.12)$$

where N_1, N_2 are the number of keypoints, and x_i, x_j are the locations of keypoints in the first and second images, respectively ($\hat{x}_j$ is the homography transformed location of keypoint j). ϵ is the detection threshold.

3.5.3. Nearest neighbor mean average precision

nn-mAP (nearest-neighbor mean average precision) metric computes the descriptor power, by evaluating the correctness of descriptors at multiple descriptor distance thresholds. Nearest neighbor matching is used to match descriptors. nn-mAP is calculated by the area under the precision-recall curve.

3.5.4. Homography estimation

The capability of an algorithm to deduce the homography linking a pair of images is evaluated by contrasting the inferred homography, denoted as $\hat{\mathcal{H}}$ with the actual homography, denoted as $\mathcal{H}$. A direct comparison of the 3×3 $\mathcal{H}$ matrices is not straightforward due to varying scales in the matrix entries. Therefore, the evaluation relies

on determining the effectiveness of the homography in mapping the four corners of one image onto the corresponding positions in the second image.

The corners of the first image are represented as c_1, c_2, c_3, c_4 . When the ground truth $\mathcal{H}$ is applied to these corners, the resulting positions in the second image are c'_1, c'_2, c'_3, c'_4 . Similarly, applying the estimated homography $\hat{\mathcal{H}}$ yields the corresponding positions $\hat{c}'_1, \hat{c}'_2, \hat{c}'_3, \hat{c}'_4$.

A specific threshold ε is utilized to categorize a homography as accurate. This threshold signifies the maximum allowable deviation between the actual and estimated corner positions in the second image. Therefore, the smaller the deviation from the true corner positions, the better the estimated homography.

$$\text{Homography Correctness} = \frac{1}{N} \sum_{i=1}^N \left(\left(\frac{1}{4} \sum_{j=1}^4 \|c'_{ij} - \hat{c}'_{ij}\| \right) \leq \varepsilon \right). \quad (3.13)$$

The scores lie within the range of 0 to 1, with a higher score being more favorable.

3.5.5. Matching accuracy

The Matching score, often referred to as M-score, it is ratio of true positive matches over total number of matches as show in **E(3.14)**, this metric evaluates the combined effectiveness of the interest point detector and descriptor. It determines the proportion of ground truth correspondences recoverable by the entire process relative to the number of features the process proposes in the shared viewpoint region. This metric is symmetrically computed across the image pair and then averaged.

$$\text{Matching score} = \frac{\text{True Postive Matches}}{\text{Total Number of Matches}} \quad (3.14)$$

CHAPTER 4

4. OUR APPROACH

In our study, we set out on an exploratory journey to navigate the domain of image registration, with a focus on multi-modal image registration. Our work evolved through three distinct approaches, each specifically tailored to address the challenges of multi-modal image registration.

4.1. Image Translation With Pix2Pix

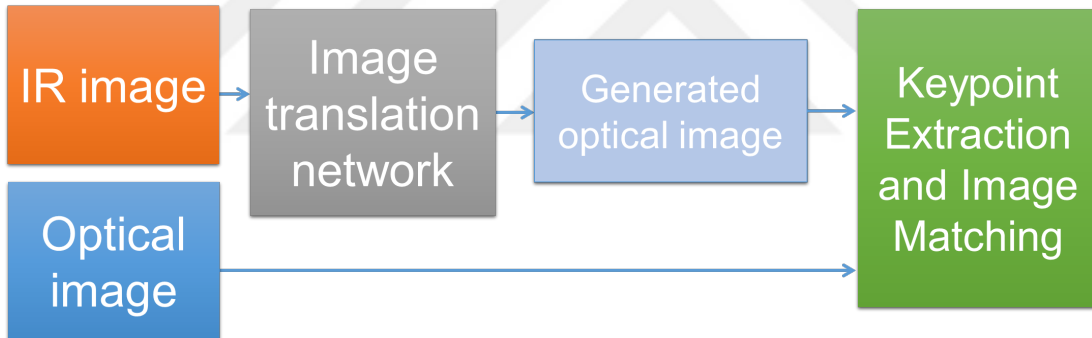


Figure 4.1: Our new architecture where image translation network applied before feature matching.

Our first approach leveraged the capabilities of the pix2pix network, a powerful tool for image-to-image translation. With this network, we post processed the images before feeding them into the MultiPoint encoder network for subsequent image matching tasks. This method capitalizes on the strength of the pix2pix network in transforming an image's appearance while preserving its structural integrity, thereby harmonizing the visual disparity between images of different modalities thereby the infrared and optical images were bring closer in the visual representation. Upon image translation, the processed images were introduced to the MultiPoint encoder, where learning feature representations and establishing robust correspondences between images was easier due to both images being similar in visual representation. This approach required training the pix2pix

network from scratch and employ it as pre-processing step in the image matching workflow as show in **Figure 4.1**.

4.2. SwinV2 Encoder and Cosine Loss

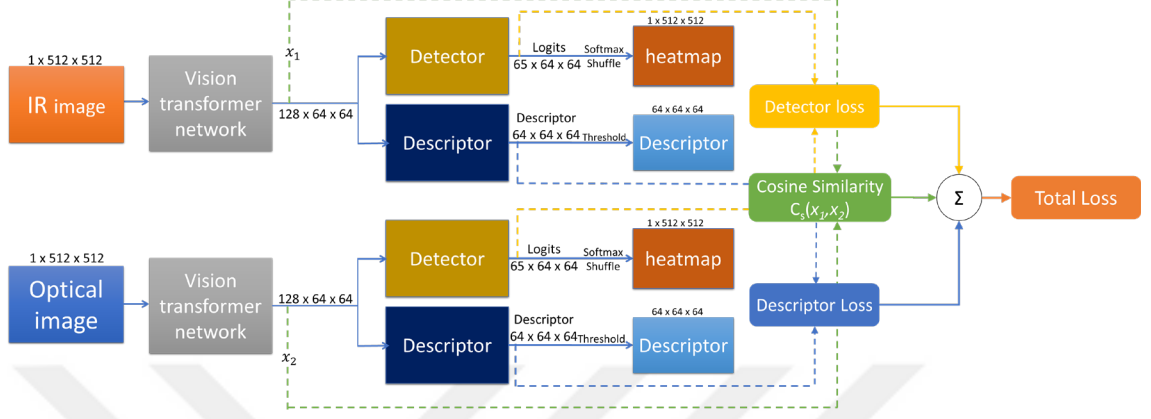


Figure 4.2: Block diagram of the network and its training process.

We claim in this work that increasing the matching accuracy can be achieved by decreasing the perceptual difference between images from different domains. Feature extractors have a hard time dealing with multi-modal images because their keypoint feature representations vary greatly from image to image. Based on this observation, our method consists of two primary components: (1) a dissimilarity reduction mechanism, which minimizes the distance between corresponding feature representations in different modalities by employing a cosine similarity, and (2) a hierarchical alignment module, which further refines the keypoint correspondences by leveraging the multi-scale feature maps generated by the Swin Transformer's encoder as in **Figure 4.2**.

We utilize the encoder part of the Swin Transformer V2 as a substitute for MultiPoint's CNN encoder to extract feature representations from multi-modal images. The encoder of the Swin Transformer V2, known for its ability to capture both local and global contextual information, is well-suited for this purpose. By incorporating the Swin Transformer V2 encoder into the Siamese network framework, our approach effectively aligns keypoints across multi-modal images and achieves improved performance in downstream tasks such as image registration, fusion, and retrieval.

Our second component aims to reduce the perceptual difference between images from different domains to improve the matching accuracy. We address the challenge of multimodal images having substantially different feature representations by employing a

dissimilarity reduction mechanism. Specifically, we minimize the distance between corresponding feature representations in different modalities by using a loss function that employs cosine similarity metric. To accomplish this, we first extract feature maps from the input images using a Swin Transformer V2 encoder. The feature maps are then flattened into vectors and we compute the cosine similarity between the feature vectors in different modalities. Ultimately, this leads to a reduction in the difference in the perceptual information between images from different domains. As we navigated the course of our second component, we were mindful to meticulously document the changes we made and the challenges we encountered.

Our aim was not just to develop a robust model for multi-modal image registration, but also to contribute to the broader knowledge in this field by sharing our experiences and learnings.

CHAPTER 5

5. RESULTS AND DISCUSSION

5.1. Training Details

5.1.1. Image translation using Pix2Pix

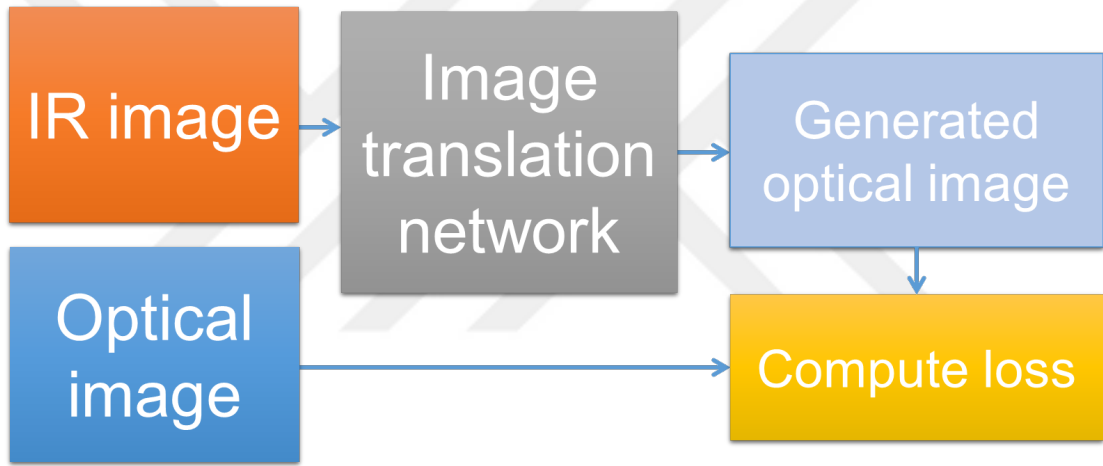


Figure 5.1: Training structure of image translation network.

The image translation network in **Figure 5.1** consists of U-Net model [15] shown in **Figure 3.3**.

The dataset utilized for training contains 9001 image pairs in total. Each input image is resized to a resolution of 512x512 pixels. The models are trained in batches, with each batch consisting of 4 images. Two generator models are evaluated, U-Net 256 and U-Net 128 where 256 and 128 are input dimension of initial layer of U-Net network.

Two discriminator models are also compared: the Basic model and the Pixel model. The Basic model is a Patch GAN with an overlapping patch size of 70x70, comprising approximately 9.76 million parameters. On the other hand, the Pixel model is a Pixel GAN [14], which can be seen as a Patch GAN with a patch size of 1x1, consisting of roughly 9,000 parameters.

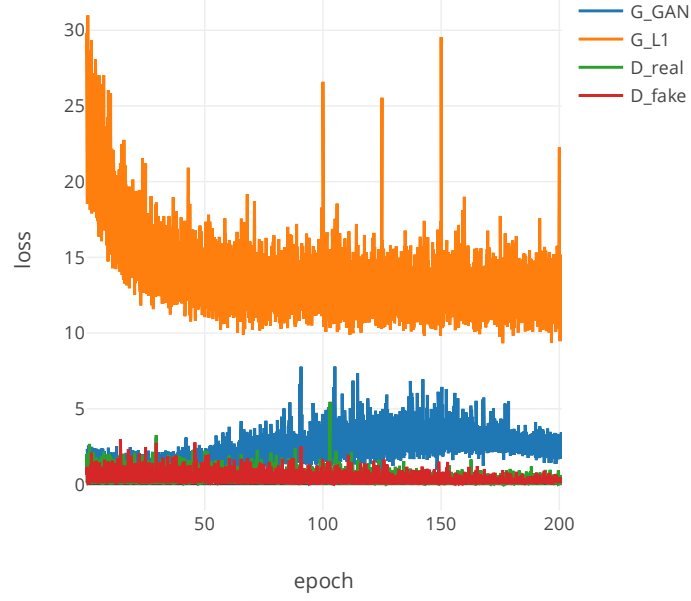


Figure 5.2: Image translation network training loss curve

Each of these models is trained over 200 epochs, using a learning rate of 0.0002. A learning rate decay is implemented starting from epoch 100, where the learning rate progressively decreases to zero. This scheme ensures a smoother convergence to the optimal solution, thus enhancing the learning efficiency and final performance of the models.

5.1.2. Swin V2 Transformer

The Swin Transformer-based Siamese network is employed with ground truth Optical and IR image pairs (See **Figure 4.2**). The Swin Transformer has an original image size of 512×512 , a patch size of 4, 1 input channel, an embedding dimension of 96, depths of [4, 8], and 3, 6, 12, and 24 heads in different layers. The window size is set to 8, with a multi-layer perceptron (MLP) ratio of 4. The model uses Layer Normalization and no axial positional encoding. Total number of pairs utilized for training is 979 from *VEDAI* dataset, with a batch size of 4. For testing, 266 pairs are used.

We train our model from scratch using PyTorch. Each model is trained for 1500 epochs with a batch size of 4. We employ a static learning rate of 0.001 with the Adam Solver. Photometric augmentations e.g. illumination changes, motion blur, and contrast changes are not used in the training. The Swin Transformer model uses patch normalization and default configuration, including a drop out rate of 0.1, original multipoint detector and descriptor network modules are fine-tuned using the loss functions as in [1] and **Section 3.4.2** training descriptor loss graph is shown in **Figure 5.3**.

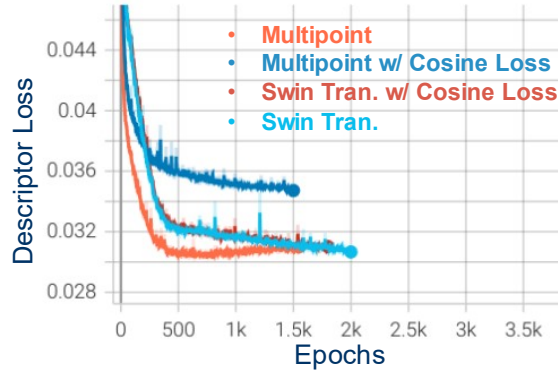


Figure 5.3: Descriptor losses training curves for second approach where it is observed that Swin transformer made Descriptor loss converges more and better than the other models.

5.2. Image Translation Approach Results

To verify the quality of image translation results we used Fréchet inception distance metric explained on **Section 3.5.1** since it measures the similarity of the visual characteristics of the image. The comparison of models based on FID Scores is given in **Table 5.1**. We measure the similarity between generated optical images and the ground truth optical images based on FID score (the lower the better).

Table 5.1: FID score results

Model	Backbone		
	Generator	Discriminator	FID Score
Pix2Pix	U-Net 256	Basic	121.7
		Pixel	107.0
	U-Net 128	Basic	119.2
		Pixel	113.8

We can observe that Pix2Pix with Generator U-Net 256 and Pixel as discriminator was best score among the other tested models. Illustration is shown in **Figure 5.4**.

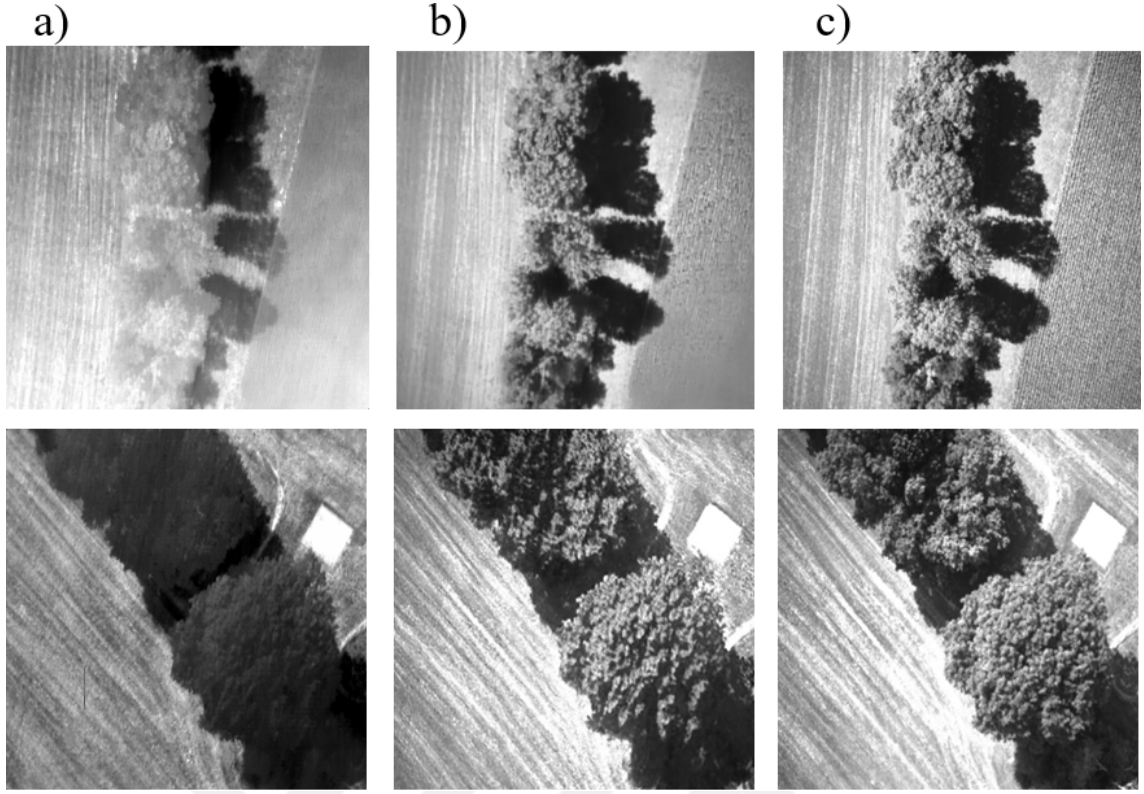


Figure 5.4: Visual results for Pix2Pix with U-Net 256 backbone a) thermal image b) generated optical from thermal c) ground truth optical image.

in **Figure 5.4** there is two samples each row is a different sample, in first column **Figure 5.4a** the IR image, in second column **Figure 5.4b** is the generated optical image and in third column **Figure 5.4c** the ground truth optical image.

Then We trained MultiPoint using original and generated optical images. Adam solver was used for 5000 epochs with a batch size of 16 and a learning rate of 0.001. ImagePair dataset was used to generate optical images from original thermal images using the Pix2Pix network. Homographic adaptation was applied between generated and original optical images. Photometric augmentations such as motion blur, illumination changes, and contrast changes were not used in this training. We used a total of 9000 image pairs during the training and 4391 during testing.

Table 5.2 provides evaluation scores for the original methods and methods using the generated optical images (see **Figure 5.5**). Number of keypoints corresponds to the average number of the detected keypoints in the test dataset.

Table 5.2: Detector and descriptor metrics comparison

Model	Evaluation Metrics				
	Data	# of keypoints	Repeatability	nn-mAP	M-Score
MultiPoint	Original	387	0.303	0.446	0.187
Surf	Original	630	0.248	0.006	0.022
	Generated	331	0.193	0.024	0.046
SIFT	Original	740	0.275	0.060	0.049
	Generated	556	0.234	0.042	0.048
SuperPoint	Original	280	0.190	0.327	0.106
	Generated	158	0.157	0.280	0.113
Retrained Multipoint	Generated	91	0.231	0.524	0.198

Figure 5.5 provides the visual comparison of detected points and matching between the two images for all the tested methods. The generated images in **Figure 5.5** and results in **Table 5.2** are from U-Net 256 generator and *Pixel* as discriminator since it outperformed the other discriminators based on **Table 5.1**

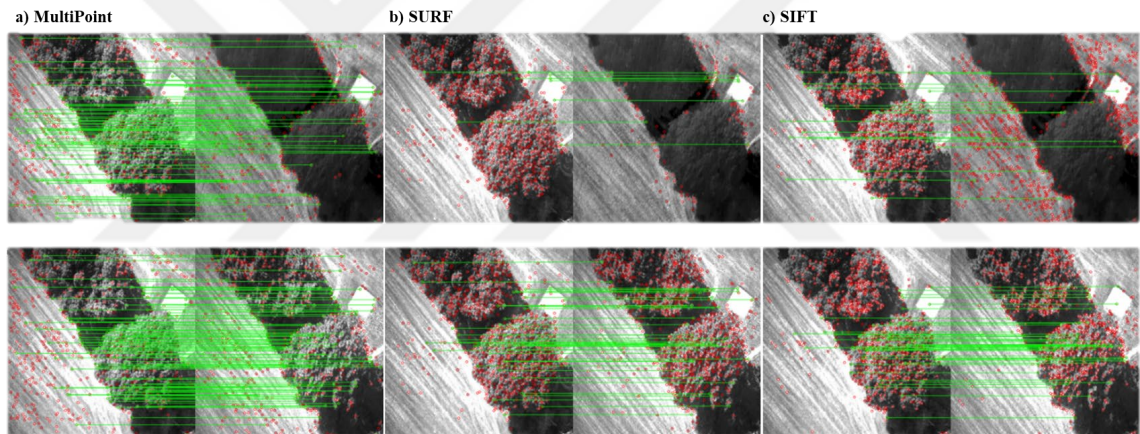


Figure 5.5: Qualitative results on the multi-modal *ImagePair* dataset. Correct matches are highlighted in green. The first row shows results from original methods (left image: optical, right image: IR). The second row shows results obtained using generated optical images in the corresponding methods (left image: optical, right image: generated). Where in a) matching is done using MultiPoint, b) using SURF and in c) using SIFT Matching performance is best when generated image is used in MultiPoint.

Results show that, even though original MultiPoint produces higher repeatability score, the retrained MultiPoint detector combined with the image translation network produces the highest nn-mAP and M-score values. Even though fewer keypoints are detected on

generated optical images, they are more reliable and better matched than their counterparts in IR images.

5.3. Visual Transformer and Cosine Loss Approach Results

We evaluate our keypoint extraction and matching model performance based on four score metrics: Repeatability, nn-mAP, M-score and Homography correctness. Each was explained in **sections 3.5.2-3.5.5** in detail.

Table 5.3 shows evaluation metrics score with 200 epochs of training for the MultiPoint encoder compared to using the Swin V2 transformer with and without cosine loss (see **Figure 5.6**). Avg number of keypoints refers to the average number of detected keypoints in the test dataset between the IR and Optical images. Results show that the Swin V2 transformer as encoder produces the highest repeatability, nn-mAP and M-score values.

Table 5.3: Keypoint detector and descriptor metrics evaluation of Swin transformer against multipoint encoder

Encoder Model	Evaluation Metrics				
	Number of Keypoints	Repeatability	nn-mAP	M-score	Homography Correctness
MultiPoint	344	0.576	0.661	0.436	0.436
MultiPoint w/ Cosine Loss	384	0.548	0.824	0.549	0.571
Swin Tran.	438	0.591	0.834	0.547	0.676
Swin Tran. w/ Cosine Loss	578	0.582	0.807	0.545	0.669

In **Figure 5.5**, we can observe that Swin Transformer has more keypoints and correct matches between the image pairs. In **Table 5.3**, we observe that Swin Transformer model overall outperforms the other models; it performs better than the MultiPoint encoder by 1.5% in terms of repeatability, 17.3% in terms of nn-mAP, 8.4% in terms of matching score and 24% in terms of homography correctness. The use of cosine loss improves the MultiPoint 12% on average in terms of nn-mAP, M-Score and homography correctness, but decreases the repeatability by 2.8%. For the Swin Transformer, the use of cosine loss results in a slight decrease in performance with respect to Swin Transformer without cosine loss. We also observe that the use of cosine loss results in a larger number of keypoints for both encoders.

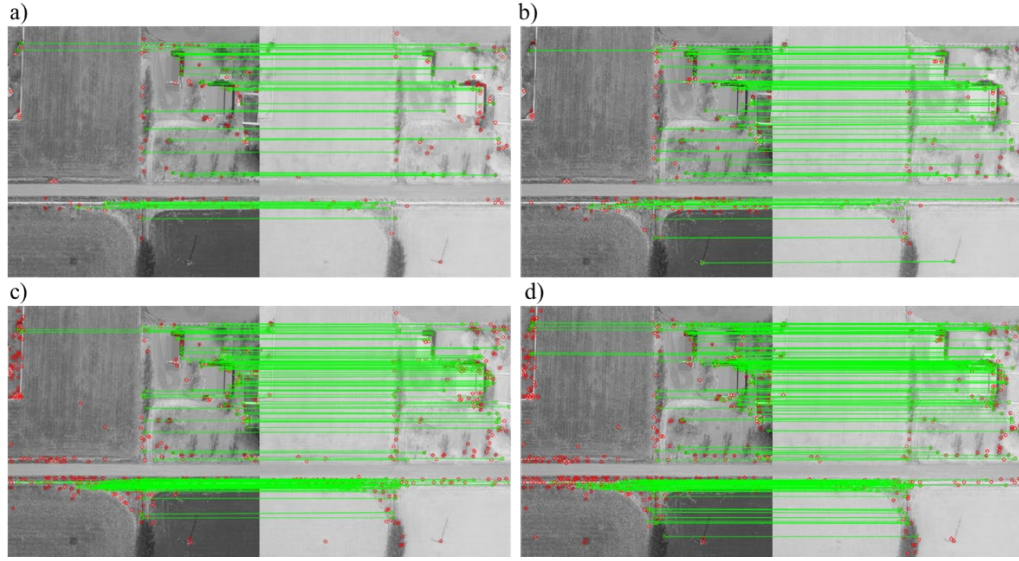


Figure 5.6: Qualitative results on the multi-modal *VEDAI* dataset. Green lines highlight correct matches (left image: Optical, right image: IR). **a,b)** shows results from original MultiPoint encoder w/ and w/o cosine loss. **c,d)** shows results obtained using Swin transformer w/ and w/o cosine loss. Matching performance is best when Swin transformer w/ Cosine loss is used. **b,d)** is results w/o cosine loss.

In **Figure 5.6**, we present an example where MultiPoint fails to estimate the correct homography while the Swin Transformer succeeds.

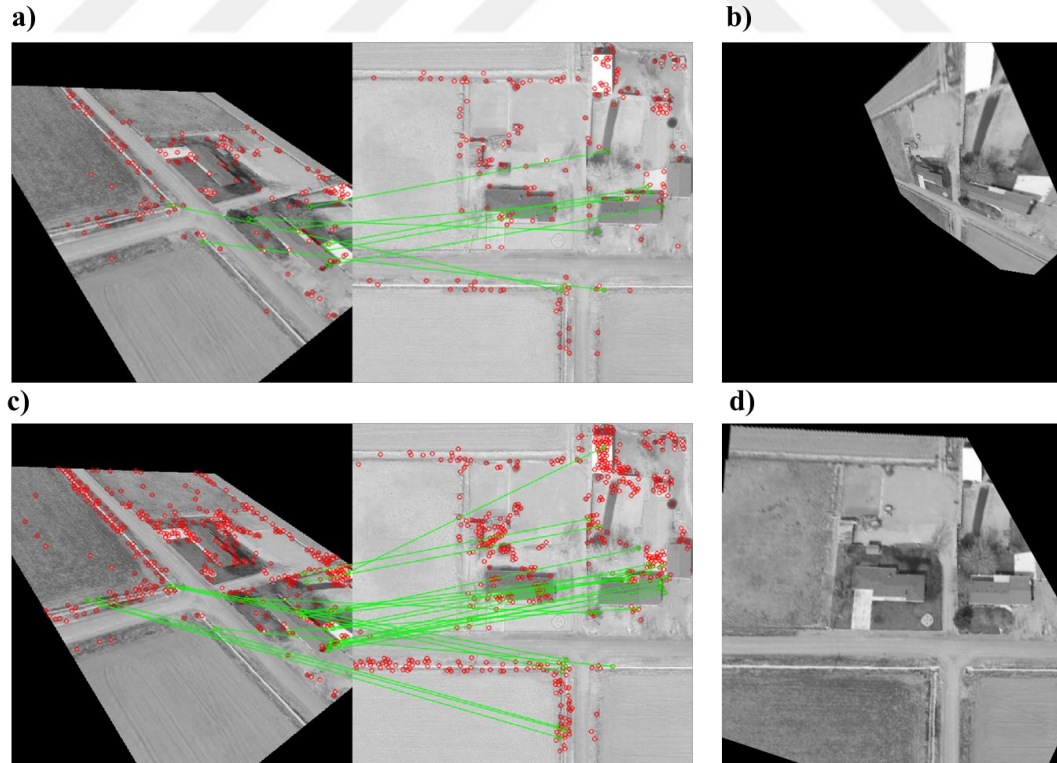


Figure 5.7: Matching with homographic adaptation where **a)** and **c)** are the matching with multipoint and swin transformer respectively and **b)** and **d)** are their homograph estimation.

CHAPTER 6

6. CONCLUSIONS AND FUTURE WORK

In this study, we have put forth two distinct approaches aimed at enhancing the accuracy of keypoint detection and matching in multi-modal image registration.

The first approach utilized the Pix2Pix image-to-image translation network as a pre-processing tool to minimize the visual disparities between infrared (IR) and optical images. By reducing these visual differences, we observed an appreciable improvement in matching accuracy. This method demonstrated that generative adversarial network (GAN) structures can be effectively employed as pre-processing modules for keypoint extraction methods, thereby achieving satisfactory accuracy without compromising run-time.

Conversely, The second approach in our study pivots around the Swin Transformer V2. This approach was developed with the aim of elevating the process of keypoint detection and matching accuracy, crucial for the task of multi-modal image registration. By deploying the Swin Transformer network, we were able to significantly diminish the disparity found in the feature map representations of multi-modal images. Consequently, this helped to augment the number of matching keypoints and improved their quality between IR and optical images. Swin Transformer V2, with its unique architecture and inherent capacity to process visual data in a holistic way, proved invaluable in our project. By exploiting its self-attention mechanisms, we were able to process entire images, accounting for both local and global image features, and effectively bridge the semantic gaps that exist between different imaging modalities.

Another contributing factor to our approach's success was the application of a Siamese network training strategy, particularly the use of a simple cosine loss function. The Siamese network architecture, designed to rank similarity between inputs, paired with the cosine loss function, enabled the optimization of our model to recognize and learn more

robust and relevant features. This innovative combination was instrumental in bolstering the overall matching performance of our system.

Moving forward, we plan to delve deeper into Swin Transformer's potential. With its versatile and adaptable nature, we are optimistic that further improvements and customization of its operational parameters and architectural structure can lead to even more effective solutions for multi-modal image registration. We aim to investigate more advanced feature encoding techniques, different optimization strategies, and possibly augment the network architecture to enhance its effectiveness in our specific task.

In sum, the second approach offered us a transformative view of handling multi-modal image registration problems. The utilization of the Swin Transformer network not only led to immediate improvements in our work but also opened up a promising path for future explorations and enhancements in this field.

As we look forward, future work will focus on extending these two approaches. For the first approach, we aim to investigate the possibility of an end-to-end training of the Pix2Pix network and the feature extractor network to further enhance the matching accuracy of multi-modal images. We plan to assess the improved keypoint detector and feature extractor network in a variety of applications, including SLAM, SfM, remote sensing, and autonomous visual navigation.

For the second approach, we envision focusing on enhancing the functionality of the Swin Transformer for multi-modal image registration. Our intention is to explore additional feature encoding techniques, optimization strategies, and network architectures that could bolster the Transformer's performance.

This dual-pronged approach offers unique perspectives in dealing with the challenges of multi-modal image registration. The insights gathered in this work, we hope, will contribute to ongoing advancements in the field.

BIBLIOGRAPHY

- [1] F. Achermann *et al.*, “MultiPoint: Cross-spectral registration of thermal and optical aerial imagery”.
- [2] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded Up Robust Features,” in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 404–417. doi: 10.1007/11744023_32.
- [3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2564–2571. doi: 10.1109/ICCV.2011.6126544.
- [4] D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperPoint: Self-Supervised Interest Point Detection and Description,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 337–33712. doi: 10.1109/CVPRW.2018.00060.
- [5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.
- [6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” arXiv, May 18, 2015. Accessed: Jul. 14, 2023. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [7] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.
- [8] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Dec. 05, 2017. doi: 10.48550/arXiv.1706.03762.
- [9] Z. Liu *et al.*, “Swin Transformer V2: Scaling Up Capacity and Resolution.” arXiv, Apr. 11, 2022. Accessed: Jul. 19, 2023. [Online]. Available: <http://arxiv.org/abs/2111.09883>
- [10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.
- [11] H. Hu, Z. Zhang, Z. Xie, and S. Lin, “Local Relation Networks for Image Recognition.” arXiv, Apr. 25, 2019. doi: 10.48550/arXiv.1904.11491.
- [12] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-Alone Self-Attention in Vision Models,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: Jul. 22, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/3416a75f4cea9109507cacd8e2f2aefc-Abstract.html
- [13] S. Razakarivony and F. Jurie, “Vehicle detection in aerial imagery : A small target detection benchmark,” *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016, doi: 10.1016/j.jvcir.2015.11.002.
- [14] A. Makhzani and B. Frey, “PixelGAN Autoencoders.” arXiv, Jun. 01, 2017. doi: 10.48550/arXiv.1706.00531.

- [15] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
- [16] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and Checkerboard Artifacts,” *Distill*, vol. 1, no. 10, p. e3, Oct. 2016, doi: 10.23915/distill.00003.
- [17] “Markov Processes over Denumerable Products of Spaces, Describing Large Systems of Automata”.
- [18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs.” arXiv, Jun. 10, 2016. doi: 10.48550/arXiv.1606.03498.
- [19] M. Fréchet, “Sur la distance de deux lois de probabilité,” *C R Acad Sci Paris*, vol. 244, no. 6, pp. 689–692, 1957.
- [20] T. Hinzmann, J. L. Schönberger, M. Pollefeys, and R. Siegwart, “Mapping on the Fly: Real-Time 3D Dense Reconstruction, Digital Surface Map and Incremental Orthomosaic Generation for Unmanned Aerial Vehicles,” in *Field and Service Robotics*, M. Hutter and R. Siegwart, Eds., in Springer Proceedings in Advanced Robotics. Cham: Springer International Publishing, 2018, pp. 383–396. doi: 10.1007/978-3-319-67361-5_25.

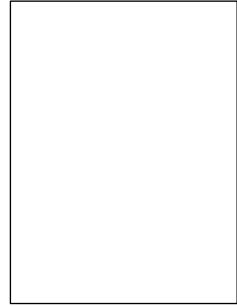
CURRICULUM VITAE

Name Surname : Mohamed Sabbah Mohamed

Mesbah Elsaeidy

Place and Date of Birth : BEHEIRA

E-Mail :



EDUCATION:

B.Sc. : 2021, Istanbul Medipol University, Faculty of Engineering and natural sciences, Department of electrical engineering

PUBLICATIONS, PRESENTATIONS, AND PATENTS ON THE THESIS:

- M. Elsaeidy, M. Emin Erkol, B. Kürşat Güntürk and H. Fehmi Ateş, "Infrared-to-Optical Image Translation for Keypoint-Based Image Registration," *2022 30th Signal Processing and Communications Applications Conference (SIU)*, Safranbolu, Turkey, 2022, pp. 1-4, doi: 10.1109/SIU55565.2022.9864815.
- M. Elsaeidy, İsmail Can YAĞMUR, H. Fehmi Ateş and B. Kürşat Güntürk, "Swin Transformer based Siamese Network for Thermal and Optical Image Registration", *2023 31th Signal Processing and Communications Applications Conference (SIU)*, Istanbul, Turkey, 2023, pp. 1-4, doi: 979-8-3503-4355-7/23/\$31.00 ©2023

MULTI-MODAL IMAGE REGISTRATION

ORIGINALITY REPORT

12%

SIMILARITY INDEX

7%

INTERNET SOURCES

8%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

- | | | |
|-------|---|-----|
| 1 | <p>Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich. "SuperPoint: Self-Supervised Interest Point Detection and Description", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018</p> <p>Publication</p> | 2% |
| <hr/> | | |
| 2 | <p>deepai.org</p> <p>Internet Source</p> | 1% |
| <hr/> | | |
| 3 | <p>mafiadoc.com</p> <p>Internet Source</p> | <1% |
| <hr/> | | |
| 4 | <p>Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021</p> <p>Publication</p> | <1% |
| <hr/> | | |
| 5 | <p>proceedings.mlr.press</p> <p>Internet Source</p> | <1% |
-