SELF-SUPERVISED DEEP LEARNING FOR MULTISPECTRAL IMAGE MATCHING

İSMAİL CAN YAĞMUR

ÖZYEĞİN UNIVERSITY

JUNE, 2025

SELF-SUPERVISED DEEP LEARNING FOR MULTISPECTRAL IMAGE MATCHING

by

İSMAİL CAN YAĞMUR

Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in
Computer Science

Advisor: Prof. Hasan F. Ateş

Graduate School of Science and Engineering
Özyeğin University
İstanbul

June, 2025

SELF-SUPERVISED DEEP LEARNING FOR MULTISPECTRAL IMAGE MATCHING

Approved by:

Prof. Hasan F. Ateş, Advisor
Department of Artificial Intelligence and Data Engineering
Özyeğin University

Prof. Çiğdem Eroğlu Erdem
Department of Electrical and Electronics Engineering
Özyeğin University

Prof. Bahadır Kürşat Güntürk
Department of Electrical and Electronics Engineering
İstanbul Medipol University

Approval Date: May 16, 2025

To my advisors, mentors, colleagues, and family — thank you for your guidance, support, and unwavering belief in me.

DECLARATION OF ORIGINALITY

I hereby declare that I am the sole author of this thesis and that this is the true copy of my thesis, including the final revisions, approved by my thesis committee. All data and information have been obtained, produced, and presented in accordance with the rules of research ethics and principles of academic honesty. As required by these rules, to the best of my knowledge I have acknowledged ideas, thoughts, and any copyrighted material in accordance with the standard referencing rules. I certify that any part of this thesis has not been submitted for a degree or diploma in another educational institution.

İsmail Can Yağmur

ABSTRACT

This thesis presents a self-supervised deep learning framework for multispectral image matching, addressing the core challenges posed by nonlinear radiation distortions (NRDs), viewpoint variations across spectral modalities, and the scarcity of annotated datasets. Existing methods often exhibit strong modality dependence and rely heavily on costly supervision, such as depth maps or calibrated camera poses, thereby limiting their generalizability across diverse spectral domains. The framework introduces an improved self-supervision strategy—Improved Multispectral Homographic Adaptation—that enhances pseudo ground truth keypoint generation in cross-spectral settings while ensuring invariance to viewpoint changes. By incorporating a spectrum-aware windowing rule, this method increases both the repeatability and the density of detected keypoints under spectral differences, thereby improving matching performance. This enhancement ultimately leads to more accurate multispectral image registration and is validated on UAV-acquired visible–thermal datasets. Building on this self-supervision strategy, the XPoint framework is proposed as a modular and fully self-supervised image matching architecture. It integrates a pretrained VMamba encoder for robust, modality-invariant feature extraction, alongside lightweight decoder heads for keypoint detection, feature description, and homography regression. This design enables efficient, label-free learning from aligned image pairs and facilitates rapid adaptation across diverse spectral modalities. The framework is designed to be scalable and easily adaptable, requiring no additional supervision beyond image pair alignment. The approach is evaluated across five public benchmarks spanning Visible-Thermal (VIS-TH), Visible-Near-Infrared (VIS-NIR), Visible-Longwave-Infrared (VIS-LWIR), Visible-Synthetic-Aperture-Radar (VIS-SAR)

datasets. Experimental results demonstrate competitive or superior performance in feature matching and multispectral image registration tasks, while maintaining high computational efficiency. This progression—from spectrum-aware self-supervision to a generalizable matching framework—positions XPoint as a practical solution for real-world multispectral applications, particularly in environments characterized by limited supervision and high spectral variability.

Keywords: Image Matching, Self-Supervised, Deep Learning, Multispectral, Image Registration

ÖZET

Bu tez, çok spektrumlu görüntü eşlemesinde spektral modaliteler arası doğrusal olmayan ışınım bozulmaları, görüş açısı farklılıkları ve etiketli veri eksikliği gibi temel zorlukları çözmeyi hedefleyen özdenetimli bir derin öğrenme çerçevesi sunmaktadır. Mevcut yöntemler genellikle tek bir modaliteye bağımlıdır ve derinlik haritaları veya kalibre edilmiş kamera pozları gibi maliyetli zemin doğrularına büyük ölçüde dayanır; bu durum, yöntemlerin farklı spektral alanlarda genelleştirilebilirliğini sınırlandırmaktadır. Bu bağlamda geliştirilen çerçeve, ”Geliştirilmiş Multispektral Homografik Adaptasyon” adlı özdenetimli öğrenme stratejisidir. Bu strateji, spektral farklılıklar altında anahtar nokta tespiti açısından daha yüksek tekrar edilebilirlik ve yoğunluk sağlayan, spektruma duyarlı bir pencereleme kuralı içermektedir. Bu sayede, farklı spektrumlarda daha güvenilir yapay zemin doğrusu anahtar noktaları üretilebilmekte ve eşleştirme performansı artırılmaktadır. Bu iyileştirme, insansız hava araçları (İHA) ile elde edilen görünür-termal veri kümeleri üzerinde test edilmiştir. Bu özdenetim stratejisi üzerine inşa edilen XPoint, modüler ve tamamen özdenetimli bir görüntü eşleştirme mimarisi olarak tasarlanmıştır. XPoint, sağlam ve modaliteden bağımsız öznitelik çıkarımı için önceden eğitilmiş bir VMamba kodlayıcıyı; anahtar nokta tespiti, öznitelik tanımlaması ve homografi regresyonu içinse kod çözücü başlıkları entegre etmektedir. Bu tasarım, hizalanmış görüntü çiftlerinden etiket gerektirmeyen, hızlı uyum sağlayabilen ve ölçeklenebilir bir öğrenme süreci sunmaktadır. Sistem, görüntü çiftlerinin hizalanması dışında ek bir denetim gerektirmeyen esnek bir yapıya sahiptir. Geliştirilen yöntem, görünür-termal, görünür-NIR, görünür-IR, görünür-LWIR ve görünür-SAR olmak üzere beş farklı açık kaynak veri kümesi üzerinde değerlendirilmiştir. Deneysel sonuçlar, önerilen çerçevenin öznitelik eşleştirme ve multispektral görüntü çakıştırma görevlerinde, yüksek hesaplama verimliliğini korurken, rekabetçi veya üstün performans gösterdiğini ortaya koymaktadır.

Bu ilerleme, spektruma duyarlı özdenetimden genelleştirilebilir bir eşleştirme çerçevesine uzanan yapısıyla XPoint'i, sınırlı denetim ve yüksek spektral değişkenlik içeren gerçek dünya multispektral problemleri için pratik bir çözüm haline getirmektedir.

Anahtar Kelimeler: Görüntü Eşleme, Öz Denetimli Öğrenme, Derin Öğrenme, Çok Spektralli, Görüntü Çakıştırma

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my esteemed supervisor, Prof. Hasan F. Ateş, for his unwavering support, guidance, and mentorship from the very beginning of my undergraduate studies. His wisdom, patience, and encouragement have been invaluable to me throughout this journey. Among many things I learned from him, perhaps the most important was the ability to stay calm under pressure—something I had often struggled with. I always felt that he genuinely valued my ideas and efforts, which I believe is essential for building self-confidence in research.

I would also like to extend my sincere thanks to Prof. Bahadır Kürşat Güntürk for his mentorship and insightful advice, which greatly contributed to my academic development.

I am particularly grateful to Prof. Hasan Fatih Uğurdağ, who first encouraged me to pursue research. His continuous support and belief in my potential have played a pivotal role in shaping my academic path. I will always carry a deep sense of gratitude for his impact on my career.

Lastly, and most importantly, I am profoundly grateful to my family. My father Yaşar, my mother Semiha, and my sister Nisa have always believed in me without hesitation, supported me unconditionally, and encouraged me to pursue my own path. None of my accomplishments would have been possible without their love and support. I am forever thankful.

This work is supported in part by TUBITAK Grant - Project No: 118E891.

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY iv

ABSTRACT v

ÖZET vii

ACKNOWLEDGEMENTS ix

LIST OF TABLES xii

LIST OF FIGURES xiii

LIST OF ACRONYMS AND ABBREVIATIONS xvii

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Contributions 6

1.3 Scope of the Thesis 9

1.4 Outline of the Thesis 9

2 RELATED WORK 11

2.1 Image Matching 11

2.1.1 2-D Geometric Transformation Models 13

2.1.2 Area-Based Methods 17

2.1.3 Feature-Based Methods 18

2.1.4 Direct Homography Regression 32

2.1.5 Summary of Image Matching Methods 34

2.2 Multispectral Image Matching 34

2.3 Homographic Adaptation 38

2.3.1 Multispectral Homographic Adaptation 40

3 METHODOLOGY 42

3.1 Overview 42

3.2 Improved Multispectral Homographic Adaptation 43

3.2.1 Motivation 43
3.2.2 Proposed Approach 45
3.3 XPoint: A Modular Self-Supervised Architecture for Multispectral Matching 49
3.3.1 Motivation 49
3.3.2 XPoint Architecture 54
3.3.3 Model Training Details 63
4 EXPERIMENTAL EVALUATION 68
4.1 Overview 68
4.2 Datasets 69
4.3 Evaluation Protocol 71
4.4 Experiments of Improved Multispectral Homographic Adaptation (IMHA). 73
4.4.1 Self-Supervision and Training 74
4.4.2 Loss Functions 74
4.4.3 Evaluation on MultiPoint's Optical–Thermal Dataset 75
4.4.4 Generalization Results on VEDAI Dataset 76
4.4.5 Ablation Study: Effect of Window Size 76
4.5 Experiments of XPoint 78
4.5.1 Feature Matching Performance 78
4.5.2 Homography Correctness and Image Registration Performance 81
4.5.3 System Runtime and Keypoint Statistics 83
4.5.4 Ablation Study 83
4.5.5 Summary of Experimental Results 90
5 CONCLUSION AND FUTURE WORK 93
5.1 Conclusion 93
5.2 Future Work 93
REFERENCES 95

LIST OF TABLES

Table 2.1: Geometric transformations for planar images. The homography is the transformation studied throughout this thesis. 13

Table 2.2: Representative local feature matching methods. 2024–2025 methods marked with †. 32

Table 2.3: Recent multispectral image–matching methods, grouped by pipeline category. Columns list the publication year, the high-level method type (descriptor learning, detector-based, detector-free, direct regression, or "universal" cross-modal models), and a brief note on each method's main technical contribution. 35

Table 3.1: Comparison of Homographic Adaptation Variants 56

Table 4.1: Summary of the multimodal datasets used for training and evaluation. 71

Table 4.2: Evaluation metrics on the MultiPoint test set with random homographies applied. 75

Table 4.3: Evaluation metrics on the VEDAI test set with random homographies applied. 77

Table 4.4: Comparison of the effects of different window sizes in evaluation metrics. 77

Table 4.5: Ablation Study of XPoint Architectural Configurations on the Optical-Thermal Dataset. 85

LIST OF FIGURES

Figure 1.1: Example of a visible–thermal image pair from the MultiPoint [1] dataset. Both images are in grayscale; the thermal image is warped using a random homography with mirror-padding at the borders. The top row shows the unaligned pair, highlighting geometric and radiometric differences. The bottom row shows the matched result, with green lines indicating sparse keypoint correspondences across modalities. 5

Figure 2.1: Overview of Image Matching Methods: Image pairs can be processed using area-based, feature-based, or direct regression approaches to estimate transformation parameters like the fundamental matrix, essential matrix, or homography matrix, which are then utilized in applications such as structure from motion (SfM), Visual Simultaneous Localization and Mapping (VSLAM), image registration, and image fusion. 12

Figure 2.2: Concept of RANSAC-based mismatch removal. Left: initial correspondences include many outliers. Right: after iterative sampling, RANSAC isolates the inlier set that supports a single geometric model while rejecting the outliers. 15

Figure 2.3: Illustration of an area-based matching algorithm. The reference window from the first image is compared across a search window in the second image to find the most similar region using metrics such as SSD or NCC. 16

Figure 2.4: Schematic overview of feature-based image matching pipelines. Image pairs are first passed through a feature extraction module. The resulting features are processed by either detector-based or detector-free branches to estimate correspondences. A resampling-based outlier removal step is then applied to refine the matches, and the final output is the geometric transformation parameters between the input images. 18

Figure 2.5: Overview of Detector-Based Methods: Features extracted from the encoder are processed through interest point and descriptor decoders, followed by putative match set construction and context aggregation. Outlier removal is then performed to obtain the final set of robust correspondences used to compute the desired transform parameters..... 21

Figure 2.6: Overview of Detector-Free Matching Frameworks: Features extracted from the encoder are processed using either coarse-to-fine matching strategies or pixel-level matching, followed by resampling-based outlier removal to estimate the final transformation parameters. 28

Figure 2.7: Homographic adaptation in SuperPoint [2]. An unlabeled image is warped by random homographies $\{H_i\}$; each warped view passes through the base detector trained on synthetic-data, MagicPoint, the responses are unwarped with H_i^{-1}, averaged, and non-maximum-suppressed to yield a viewpoint-invariant keypoint superset. 38

Figure 2.8: Multispectral homographic adaptation in MultiPoint [1]. A random homography $\mathcal{H}_i$ is applied to each image of an aligned optical–thermal pair. Detector responses are computed independently, multiplied pixel-wise to keep only cross-spectrally consistent peaks, unwarped with $\mathcal{H}_i^{-1}$, and averaged over N_h samples to form a robust interest-point superset valid in both spectra. 40

Figure 3.1: Pipeline of the proposed *Improved Multispectral Homographic Adaptation* for generating robust pseudo-ground truth keypoints across modalities. Given aligned optical–thermal image pairs, random homographies are applied and the warped images are processed by a base keypoint detector (e.g., SURF [3]). A windowing-based acceptance criterion is used to tolerate spatial shifts between modalities. The resulting filtered keypoint maps are unwarped and aggregated to produce separate training labels for each spectrum. 45

Figure 3.2: Illustration of the keypoint acceptance criterion in the windowing strategy. A candidate keypoint from Spectrum X is accepted only if there exists at least one keypoint within a predefined window region in the corresponding location of Spectrum Y. The figure shows two scenarios: in the top row, the candidate keypoint is accepted due to spatial proximity to a matching keypoint in Spectrum Y; in the bottom row, the candidate is rejected due to the absence of nearby keypoints. Here, 1 denotes a keypoint, 0 a non-keypoint, and x represents a 'don't care' region outside the search window. 46

Figure 3.3: Illustration of the windowing computation used for keypoint aggregation. Two keypoint heatmaps from different spectral modalities are compared using a sliding window strategy. For each candidate keypoint (value 1), a local patch is extracted from the other modality's heatmap and summed. If the sum exceeds a threshold (e.g., at least one overlapping keypoint), the keypoint is accepted. Aggregated heatmaps are then thresholded to produce the final keypoint maps. The figure shows padding, window-based comparison, and final acceptance logic applied to two sample cases. 47

Figure 3.4: Overview of the XPoint pipeline. **(A)** Self-supervision stage applies improved multispectral homographic adaptation using RIFT2 and probabilistic windowing to generate robust pseudo-ground truth keypoints. **(B)** Training with joint optimization of keypoint detection, descriptors, and homography. **(C)** Inference stage with matching and robust transformation estimation. 52

Figure 3.5: Random homographies are generated by combining simple transformations such as translation, scaling, rotation, and symmetric perspective distortion, sampled within predefined ranges. 55

Figure 3.6: Overview of the adopted VMamba encoder. The input image of size $H \times W \times 3$ is partitioned into patches and embedded to a feature map of size $\frac{H}{4} \times \frac{W}{4} \times 48$. This representation passes through four stages of VSS blocks, each reducing the spatial resolution by a factor of 2. After patch unembedding, the final encoded output $\mathcal{B}$ has dimensions $\frac{H}{8} \times \frac{W}{8} \times \frac{C}{2}$ and serves as the compact feature representation for downstream tasks. 57

Figure 3.7: Interest Point and Descriptor Decoders. The interest point decoder outputs keypoint heatmaps, while the descriptor decoder generates dense descriptors for the input image. 59

Figure 3.8: Homography Head Architecture. 60

Figure 4.1: Sample image pairs from the five multispectral datasets used in this study. Each row presents an example image pair from a specific dataset, showcasing the diversity of spectral combinations and scene types: agricultural aerial imagery in MultiPoint (VIS-TH), urban aerial scenes in Vehicle Detection in Aerial Imagery (VEDAI) (VIS-NIR), driving scenarios in VIS-IR (VIS-LWIR), architectural scenes in VIS-NIR (VIS-NIR), and satellite imagery in VIS-SAR (VIS-SAR). Visible spectrum images are shown in the left columns, while their corresponding multispectral counterparts are displayed on the right. 70

Figure 4.2: Qualitative results on the MultiPoint dataset. In each matched image pair, the left image is optical and the right is thermal. Green lines show correct matches within a 4-pixel threshold. Both **Window-5** and **MultiPoint** succeed when enough texture is present, while **SuperPoint** struggles under large cross-spectral transformations. All methods degrade under extreme modality differences (bottom row). 76

Figure 4.3: Repeatability (left) and Matching Score (right) for detector-based (RIFT2, MultiPoint, ReDFeat), detector-free (ASpanFormer, DKM, LoFTR, RoMa, XoFTR) and proposed XPoint variants across five multispectral datasets: MP (VIS-TH), VEDAI (VIS-NIR), VIS-NIR (VIS-NIR), VIS-IR (VIS-LWIR) and VIS-SAR (VIS-SAR). XPoint_001 exhibits the highest repeatability and matching consistency on all datasets except VIS-NIR, where detector-free methods marginally outperform. 79

Figure 4.4: Homography correctness versus warp threshold (ϵ) for detector-based (RIFT2, MultiPoint, ReDFeat), detector-free (ASpanFormer, DKM, LoFTR, RoMa, XoFTR), and proposed XPoint variants across five benchmarks: OPT-TH (VIS-TH), VEDAI (VIS-NIR), VIS-NIR, VIS-IR (VIS-LWIR), and VIS-SAR. 82

Figure 4.5: Comparison of average runtime (seconds) and keypoint counts across datasets for various methods. 84

Figure 4.6: Visual comparison of keypoint detection results for optical–thermal image pairs using different base detectors (SURF vs. RIFT2) and homographic adaptation strategies. Rows correspond to different aggregation techniques (Gaussian, weighted windowing, and proposed unweighted windowing), and columns indicate the base detector used. Nkp denotes the number of detected keypoints in each image. 88

Figure 4.7: Visual comparison between Cross-Entropy and Weighted Cross-Entropy loss strategies. 90

Figure 4.8: Qualitative results on three datasets—OPT-TH, VEDAI, and VIS-NIR—using four methods: XPoint (ours), ReDFeat, RoMa, and XoFTR. Top row per method shows MAGSAC-filtered keypoint matches; bottom row shows warped image alignment. 91

Figure 4.9: Qualitative results on two datasets—VIS-IR and VIS-SAR—using the same four methods. Match quality and registration accuracy are compared using MAGSAC-filtered keypoints and resulting image warps. 92

LIST OF ACRONYMS AND ABBREVIATIONS

CNN Convolutional Neural Network

DHR Direct Homography Regression

DHR Direct Homography Regression

GPU Graphics Processing Unit

IR Infrared

IMHA Improved Multispectral Homographic Adaptation

LWIR Long-Wave Infrared

MS-COCO Microsoft Common Objects in Context

MIM Multispectral Image Matching

MMIM Multimodal Image Matching

NRD Nonlinear Radiation Distortion

NIR Near-Infrared

NN Nearest Neighbor

NCC Normalized Cross-Correlation

NN mAP Nearest Neighbor mean Average Precision

NMS Non-Maximum Suppression

RANSAC Random Sample Consensus

SAR Synthetic Aperture Radar

SWIR Short-Wave Infrared

SSD Sum of Squared Differences

SAD Sum of Absolute Differences

TIR Thermal Infrared

TH Thermal

VIS Visible

VIS-NIR Visible-Near-Infrared

VIS-IR Visible-Infrared

VIS-LWIR Visible-Longwave-Infrared

VIS-SAR Visible-Synthetic-Aperture-Radar

VIS-TH Visible-Thermal

VMamba Vision Mamba

VSLAM Visual Simultaneous Localization and Mapping

VEDAI Vehicle Detection in Aerial Imagery

CT Computed Tomography

MRI Magnetic Resonance Imaging

PET Positron Emission Tomography

1. INTRODUCTION

1.1 Motivation

Image matching establishes correspondences between points, regions, or structures in two or more images by identifying visual features that represent the same underlying physical elements across different views, times, or sensing conditions. Its objective is to spatially align images despite differences in viewpoint, illumination, scale, modality, or other acquisition parameters. Image matching serves as a fundamental component in a wide range of applications, including 3D reconstruction [4, 5], VSLAM [6, 7], structure-from-motion (SfM) [8, 9], aerial and satellite image analysis [10, 11], image mosaicking and fusion [12, 13], image retrieval and information integration [14, 15], and optical flow estimation [16, 17]. Image matching has become increasingly vital for enabling machines to perceive, interpret, and make decisions based on their environment, thanks to the latest developments in deep learning. Hence, developing accurate, fast, and robust matching techniques remains a critical research challenge to support autonomous systems and real-world visual understanding tasks.

Multimodal Image Matching (MMIM) refers to the process of establishing dense or sparse correspondences between image pairs acquired under different sensing modalities, acquisition conditions, or temporal settings. The primary objective of MMIM is to integrate complementary information from diverse sources—such as images captured by different sensors [18, 1], taken at different times [19, 20, 21], under varying weather or seasonal conditions [22, 23, 24, 25], or across heterogeneous data types such as image–text pairs [26, 27, 28]. MMIM plays a critical role in enhancing scene understanding and enabling robust perception in complex environments. Within this broader scope, Multispectral Image Matching (MIM) constitutes a specialized subtask focused on aligning images captured across different spectral bands—such as VIS, IR, NIR, SWIR, or

LWIR. Unlike generic multimodal scenarios, MIM assumes that the data remains image-to-image and structurally consistent, while differing primarily in spectral response. The key challenge in MIM lies in handling NRD and spectrum-dependent feature shifts, which significantly hinder matching task. As such, MIM demands specialized techniques that are robust to spectral variability while preserving spatial and geometric consistency across modalities, thereby enabling a more comprehensive and robust understanding of the observed environment [29]. In remote sensing, multispectral matching facilitates the fusion of optical, SAR, infrared, and LiDAR imagery [30, 31], playing a crucial role in environmental monitoring, disaster management, and land use analysis. Medical imaging enables the alignment of Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans [32, 33], supporting accurate diagnosis, treatment planning, and surgical guidance. Autonomous navigation systems rely on multispectral image matching to fuse visual, thermal, and LiDAR data under dynamic conditions [34], enhancing the robustness and safety of self-driving vehicles and UAVs. In security and surveillance, aligning visible and infrared imagery [35] improves object detection and human identification in low-light scenarios. The ability to adaptively match multispectral images strengthens decision-making processes and significantly improves the operational effectiveness of these technologies in real-world applications.

Multispectral image matching presents significantly greater challenges than traditional single-modal matching due to the enormous diversity in sensing mechanisms, environmental conditions, and data characteristics. Images captured by different modalities—such as visible, near-infrared, thermal infrared (TIR), or synthetic aperture radar (SAR)—exhibit substantial differences in texture, contrast, spatial resolution, and radiometric properties, often caused by distinct underlying physical phenomena: reflection in the visible spectrum versus thermal radiation in TIR imagery, for instance. Moreover,

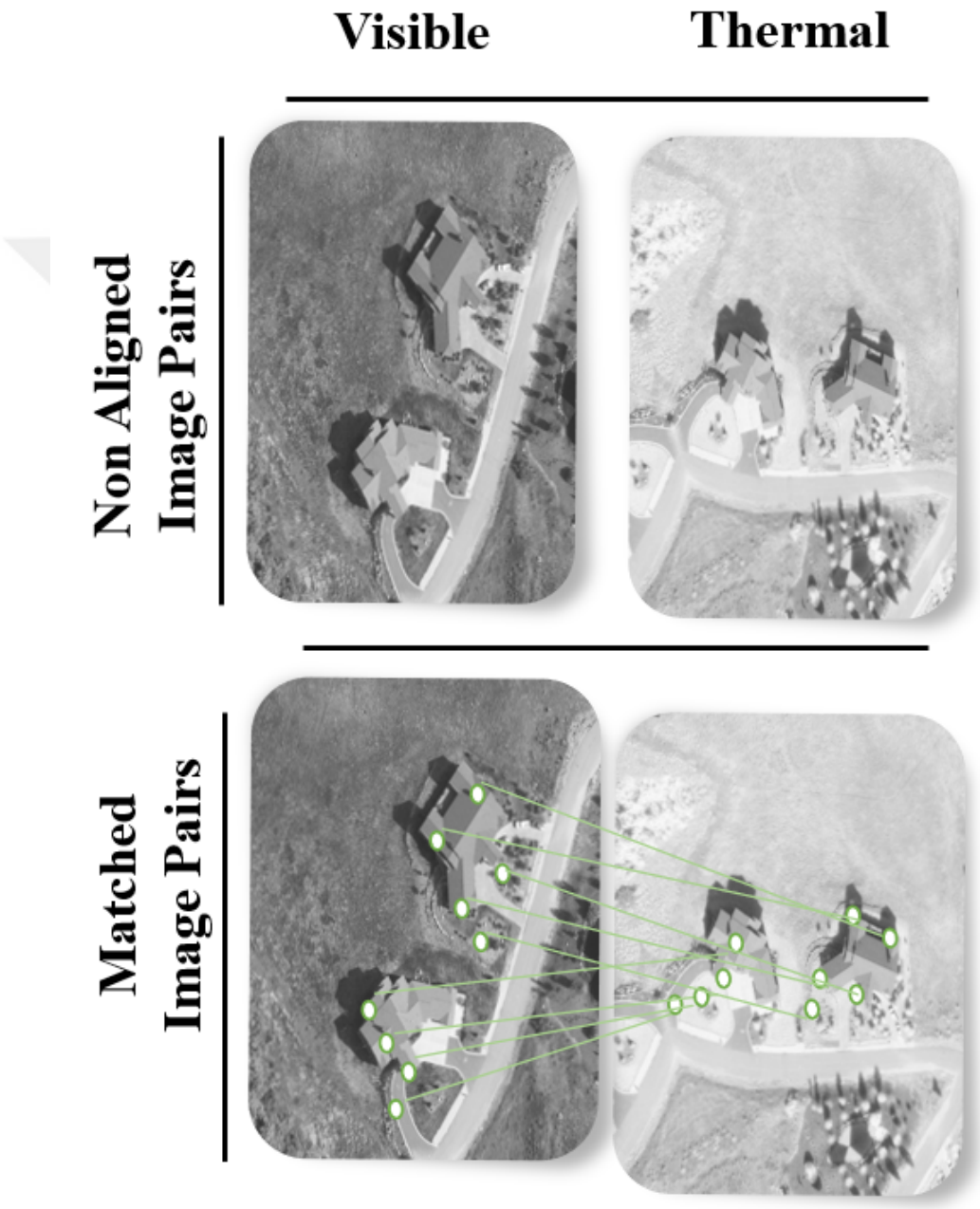
these differences are further exacerbated by viewpoint changes, scale variations, and low-texture regions, complicating the reliable extraction of correspondences across modalities.

Early solutions primarily relied on handcrafted feature detectors such as SIFT [36] and SURF [3]. These methods exhibit robustness to scale, rotation, and moderate illumination changes but struggle under extreme modality differences and require careful parameter tuning. They are also computationally expensive and fail to maintain performance across severe appearance variations, particularly in thermal-visible matching scenarios. With the advent of deep learning, learning-based methods such as SuperPoint [2] and LoFTR [37] have achieved remarkable improvements in single-spectrum (visible-visible) matching. However, their generalization to multispectral settings remains limited, primarily due to their dependence on appearance consistency and the scarcity of large-scale, diverse multispectral datasets. Visible-thermal matching, in particular, poses unique challenges due to the non-linear and non-invertible relationship between modalities, low resolution, narrow field of view in TIR sensors, and substantial differences in texture representations. In response, several specialized multispectral image matching frameworks have been developed, including handcrafted approaches [38, 39] and learning-based models [1, 18]. While these approaches show promise within their target domains, they typically focus on a specific modality pair (e.g., visible–thermal or visible–infrared) and fail to generalize across a broader range of spectral combinations. Furthermore, recent attempts to enhance generalization via large foundational models [40] have encountered challenges such as computational inefficiency, overfitting, and performance degradation in extreme cross-modal scenarios. Given these limitations, there is a clear need for a unified, modular, and self-supervised image matching framework that can adapt across multiple spectral domains without reliance on dense supervision (e.g., depth maps, camera poses) or task-specific retraining. Such a framework should utilize flexible architectural components and robust self-supervised training strategies to bridge

the domain gaps inherent in multispectral image matching tasks. Figure 1.1 illustrates the challenges of multispectral alignment and the motivation for robust matching. The top row shows an unaligned visible–thermal image pair with clear geometric and radiometric differences, while the bottom row displays sparse keypoint correspondences. These accurate matches can then be used to estimate a homography and warp the thermal image to align with the visible one, enabling cross-spectral registration.

To address the challenges inherent in multispectral image matching, this thesis introduces a two-stage approach that progressively advances from a specialized solution to a general, scalable framework. The initial focus is on the specific case of visible–thermal image matching, a particularly challenging setting characterized by extreme spectral discrepancies. Building on the MultiPoint [1] pipeline, the proposed approach includes an improved homographic adaptation strategy that reformulates the pseudo-ground truth generation process. Instead of relying on pixel-wise multiplication to combine keypoints detected in each modality—a method that can inadvertently suppress salient features under non-linear spectral differences—a spectrum-aware windowing rule is introduced. This rule accepts a keypoint detected in one spectrum if a corresponding response falls within a small spatial window in the other spectrum. This straightforward yet effective modification generates separate but overlapping keypoint supersets for each modality, resulting in richer and more robust pseudo ground-truth heatmaps. Experimental evaluations on two UAV-acquired visible–thermal datasets demonstrate that this approach significantly improves repeatability, matching score, and homography estimation accuracy, all without requiring manual labels [18]. These results validate that self-supervised training remains effective even under extreme cross-spectral conditions and that careful design of keypoint acceptance criteria can bridge much of the gap between visible-only and cross-spectral deployment.

Figure 1.1: *Example of a visible–thermal image pair from the MultiPoint [1] dataset. Both images are in grayscale; the thermal image is warped using a random homography with mirror-padding at the borders. The top row shows the unaligned pair, highlighting geometric and radiometric differences. The bottom row shows the matched result, with green lines indicating sparse keypoint correspondences across modalities.*



While the improved keypoint strategy addresses the visible–thermal case, practical systems often operate across a wide variety of spectral modalities, necessitating more flexible and scalable solutions. Therefore, in the second stage, we introduce *XPoint* [41], a fully self-supervised and modular architecture designed for general multispectral image matching. XPoint incorporates the improved multispectral homographic adaptation for keypoint generation but replaces the original CNN backbone with a pretrained VMamba encoder, using its 2D-Selective-Scan capabilities for robust feature extraction. The encoder outputs are processed by three lightweight decoder heads: two specialized for keypoint detection and descriptor extraction, and a third for homography regression, which enforces geometric consistency during self-supervised training. Crucially, XPoint requires only aligned image pairs without depth maps, camera poses, or external annotations, enabling rapid adaptation to new domains. Extensive experiments across five diverse datasets—covering VIS-TH, VIS-NIR, VIS-LWIR, VIS-SAR, matching tasks—demonstrate that XPoint achieves state-of-the-art performance in both feature matching and homography estimation, while maintaining near real-time efficiency. Together, the two studies chart a coherent progression: we first solve the visible–thermal matching challenge through an enhanced pseudo ground truth keypoint generation method, then extend these insights into a general-purpose framework capable of robust and scalable multispectral image matching without reliance on costly supervision.

1.2 Contributions

This thesis introduces two self-supervised learning frameworks to address the challenges of feature matching and geometric alignment in multispectral image registration. The methods are designed to reduce dependence on labeled data and improve adaptability across spectral domains. The main contributions of this thesis are as follows:

1. **Improved Homographic Adaptation for Cross-Spectral Keypoint Generation**:

We develop a pseudo ground truth keypoint generation method for visible–thermal image matching by reformulating the homographic adaptation stage used in MultiPoint. Our method introduces a spectrum-aware window-based keypoint acceptance rule that yields higher repeatability and more informative pseudo-labels without relying on pixel-wise map multiplication. This improvement enables self-supervised training with superior performance in matching accuracy and homography estimation on UAV-collected visible–thermal datasets.

2. **XPoint: A Modular, Self-Supervised Framework for General multispectral Registration**: We propose *XPoint*, a spectrum-agnostic matching framework based on a pretrained VMamba encoder and modular decoder heads for keypoint detection, description, and homography regression. XPoint is designed for rapid adaptation to new spectral domains and achieves strong performance in visible–NIR, visible–IR, visible–LWIR, and visible–SAR matching tasks. The method requires only aligned image pairs and operates without supervision, external annotations, or depth/camera metadata, establishing new baselines on five public benchmarks.

3. **End-to-End Label-Free Training Paradigm**: Both proposed frameworks use self-supervision derived from homographic consistency and aligned image pairs. This eliminates the need for expensive manual annotations or sensor metadata, offering a scalable training strategy suitable for real-world multispectral deployments with minimal data curation.

4. **Demonstration of Practical Relevance and Scalability**: The proposed methods span from a task-specific solution (visible–thermal registration) to a general-purpose, modular framework (XPoint), demonstrating a clear and reproducible trajectory from domain-specific insights to scalable architectures. The models are

validated on challenging real-world datasets and support applications in UAV navigation, cross-spectral mapping, and autonomous perception systems.

Related Publications :

- **Published and Presented**: The work on improved homographic adaptation for visible–thermal image matching was published in the proceedings of the *SPIE Conference on Image and Signal Processing for Remote Sensing XXIX*.
  - I. C. Yagmur, H. F. Ates, "Improved Homographic Adaptation for Keypoint Generation in Cross-spectral Registration of Thermal and Optical Imagery," *Proc. of SPIE*, Vol. 12733, 2023.
- **Accepted (To be Presented)**: The CS-REG-NET framework was accepted for presentation at the *33rd IEEE Signal Processing and Communications Applications Conference (SIU 2025)*. The paper introduces a self-supervised visual-state-space-based architecture for robust image matching of thermal and optical imagery.
  - I. C. Yagmur, H. F. Ates, S. U. Ulu, A. Naghavi, B. K. Gunturk, "CS-REG-NET: A Self-Supervised Visual-State-Space Based Architecture for Cross-Spectral Registration of Thermal and Optical Imagery," *Proc. of SIU 2025* (Accepted).
- **Under Preparation**: The XPoint framework is currently under preparation. The paper presents a general-purpose, self-supervised architecture for cross-spectral image registration, with extensive experiments validating its robustness across five datasets.
  - I. C. Yagmur, H. F. Ates, B. K. Gunturk, "XPoint: A Self-Supervised Visual-State-Space based Architecture for Multispectral Image Registration," arXiv, 2024.

1.3 Scope of the Thesis

This thesis investigates self-supervised methods for robust multispectral image matching, with a focus on handling the spectral diversity encountered in real-world sensing scenarios. The primary aim is to improve keypoint detection and geometric alignment across modalities such as *Visible* (VIS), *Thermal* (TH), *Near-Infrared* (NIR), *Long-Wave Infrared* (LWIR), and *Synthetic Aperture Radar* (SAR). To this end, two main contributions are proposed: an enhanced homographic adaptation technique for visible–thermal pairs, and XPoint, a modular self-supervised framework for general multispectral matching.

The research includes a spectrum-aware keypoint generation strategy and a scalable encoder-decoder architecture that predicts interest points, descriptors, and homographies without manual labels. Evaluation is conducted on multiple public datasets, benchmarking against both handcrafted and deep learning-based methods in terms of repeatability, matching accuracy, and transformation estimation. This work excludes topics such as non-rigid registration, medical image alignment, and dense supervision (e.g., depth maps or camera poses). Instead, it adopts a self-supervised training paradigm using only aligned multispectral image pairs, addressing the need for generalizable matching under limited supervision and high spectral variability.

1.4 Outline of the Thesis

This thesis is organized into five chapters, each addressing a distinct aspect of the proposed self-supervised multispectral image matching framework:

- **Chapter 1 – Introduction:** Presents the motivation for tackling the problem of multimodal image matching, outlines the main contributions of the thesis, defines its scope, and provides a summary of the thesis structure.

- **Chapter 2 – Related Work:** Reviews foundational and recent work on image matching, covering 2D geometric transformation models, area-based and feature-based methods, and direct homography regression techniques. The chapter then focuses on prior studies in multispectral image matching and surveys different homographic adaptation strategies, identifying the limitations of pixel-level supervision in cross-spectral contexts.

- **Chapter 3 – Methodology:** Introduces the proposed self-supervised pipeline. First, it details the Improved Multispectral Homographic Adaptation (IMHA) technique, which enhances keypoint label generation across spectral modalities using a spatial windowing strategy. Then, it presents XPoint—a modular and scalable architecture for multispectral matching—outlining its motivation, network design, and training procedures.

- **Chapter 4 – Experimental Evaluation:** Describes the experimental setup and datasets used for training and evaluation. It presents extensive quantitative and qualitative evaluations for both IMHA and XPoint across multiple metrics. The chapter includes evaluations on cross-spectral datasets (MultiPoint and VEDAI), analyses the impact of key architectural choices, compares performance with state-of-the-art methods, and discusses runtime and keypoint statistics.

- **Chapter 5 – Conclusion and Future Work:** Summarizes the main findings of the thesis and proposes future research directions, such as incorporating meta-learning for adaptive matching, extending the framework to other modality pairs, and exploring its use in downstream tasks like semantic segmentation or 3D reconstruction.

2. RELATED WORK

This chapter surveys the literature on image matching, with an emphasis on multispectral settings. It begins by framing image matching as the task of finding spatial correspondences between two or more images so that downstream applications, such as structure-from-motion, VSLAM, and image fusion, can recover the underlying scene geometry. When scenes are approximately planar, these correspondences enable the estimation of a single projective transformation, or homography, that aligns the views. The review then organises existing image matching methods into three broad families: (i) area-based techniques that compare raw intensity patches, (ii) feature-based pipelines that extract features and match them, and (iii) direct regression networks that predict the transformation parameters without forming explicit correspondences. After introducing these generic approaches, the discussion turns to the specific challenges of multispectral imagery, including non-linear radiometric differences, sensor-dependent noise, and the scarcity of annotated data, and it summarises recent self-supervised and modality-agnostic solutions designed to address these issues. The chapter concludes with homographic adaptation and its multispectral extensions, which generate spectrum-consistent pseudo labels for training detector-based methods and form a key component of the methodology developed in this thesis.

2.1 Image Matching

Image matching is a fundamental task in computer vision that involves establishing spatial correspondences between two or more images of the same scene. These correspondences enable the estimation of geometric transformation parameters, such as homographies, fundamental matrices, or essential matrices, which in turn support a wide range of downstream applications. These include 3D reconstruction [4], VSLAM [6], structure-from-motion (SfM) [8], aerial image analysis [10], and multimodal image fusion [13].

Figure 2.1: *Overview of Image Matching Methods: Image pairs can be processed using area-based, feature-based, or direct regression approaches to estimate transformation parameters like the fundamental matrix, essential matrix, or homography matrix, which are then utilized in applications such as structure from motion (SfM), VSLAM, image registration, and image fusion.*

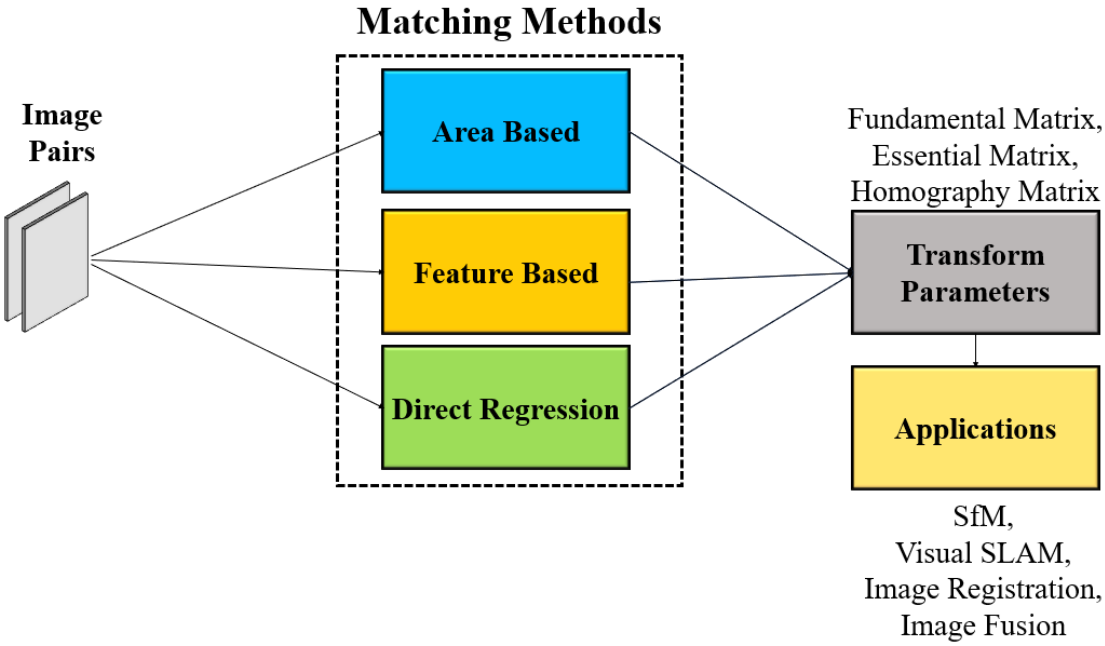


Figure 2.1 illustrates this process as a multi-stage pipeline that begins with paired input images and proceeds through different matching strategies to ultimately estimate transformation parameters and enable high-level vision tasks. At the core of this pipeline are three principal categories of image matching methods: *area-based*, *feature-based*, and *direct regression-based* approaches. Each category represents a different philosophy for extracting and comparing information across images. Area-based methods operate directly on pixel intensities, comparing fixed-size windows to find the most similar regions across views. Feature-based methods identify and match features, and are generally more robust to variations in viewpoint, scale, and modality. Direct regression methods bypass explicit matching altogether by using deep neural networks to directly estimate transformation parameters from paired images. While each family of methods has unique strengths, they all

ultimately aim to recover accurate spatial mappings between images in the form of transformation parameters. These parameters describe the underlying geometric relationship between views and are explored in detail in the following subsection.

Table 2.1: *Geometric transformations for planar images. The homography is the transformation studied throughout this thesis.*

Type (DoF)	**Matrix Form**	**Invariant Properties**
Homography (8)	$\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$	Collinearity, cross-ratio of four collinear points
Affine (6)	$\begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix}$	Parallelism, ratios of lengths on parallel lines
Similarity (4)	$\begin{bmatrix} sR & t \\ 0 & 1 \end{bmatrix}$	Angles and length ratios
Euclidean / Rigid (SE(2), 3)	$\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$	Lengths, angles, areas

2.1.1 2-D Geometric Transformation Models

In image matching, a geometric transformation specifies how coordinates in one image map to another. For planar scenes—or images captured by rotating a camera about its optical centre—the most expressive yet tractable model is the homography. Table 2.1 lists the homography and its three constrained variants (affine, similarity, and Euclidean), each variant reducing the degrees of freedom by imposing additional geometric constraints. Throughout this thesis we rely on the full homography model, because multispectral pairs often exhibit viewpoint-induced distortion and mild lens asymmetries that simpler models cannot absorb. The terms projective and perspective are used interchangeably with homography throughout the remainder of this thesis.

Homography : A homography is a 3×3 matrix $\mathcal{H}$ acting on homogeneous image coordinates. Although $\mathcal{H}$ contains nine scalars, it has only eight independent parameters because it is defined up to a non-zero scale factor:

$$\lambda \begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} & \mathcal{H}_{13} \\ \mathcal{H}_{21} & \mathcal{H}_{22} & \mathcal{H}_{23} \\ \mathcal{H}_{31} & \mathcal{H}_{32} & \mathcal{H}_{33} \end{pmatrix}}_{\mathcal{H}} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix}, \tag{2.1}$$

where (u, v) and (u', v') are corresponding pixel coordinates, and λ is removed by normalising with the third (homogeneous) component.

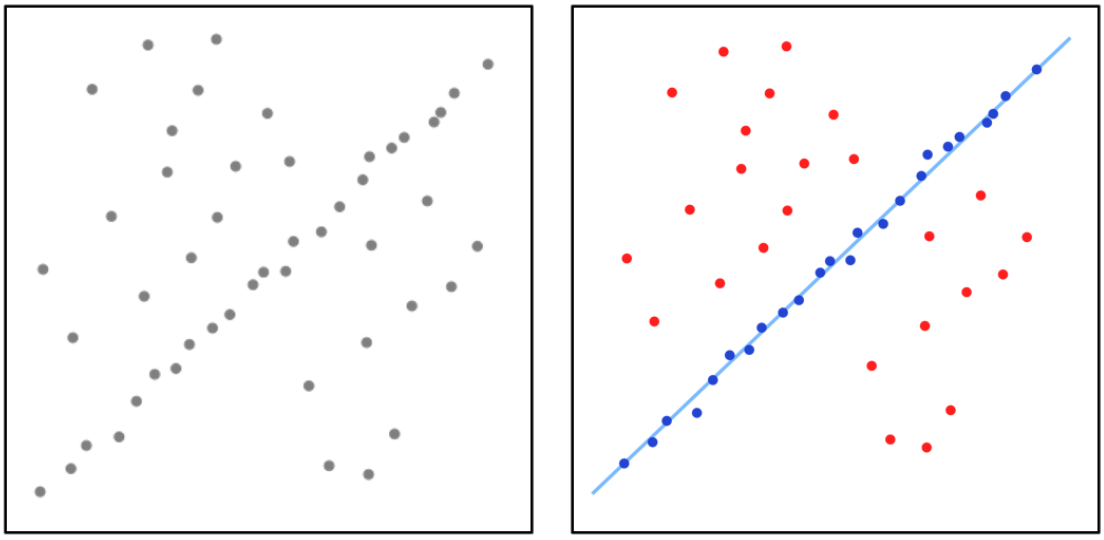
The 2×2 block in the upper left of $\mathcal{H}$ encodes rotation and, when the singular values differ, anisotropic scaling. The third column $(\mathcal{H}_{13}, \mathcal{H}_{23})^\top$ introduces in-plane translation, while the third row $(\mathcal{H}_{31}, \mathcal{H}_{32})$ contributes terms that bend parallel lines toward vanishing points. With these elements, a homography preserves straight lines and the cross-ratio of four collinear points, providing enough flexibility to model planar scenes without resorting to a full 3-D formulation.

Beyond the planar assumption : If a scene contains significant depth variation or involves wide-baseline camera motion, a homography is no longer sufficient. The relationship between views then follows epipolar geometry and is described by a 3×3 fundamental matrix in uncalibrated settings or an essential matrix when camera intrinsics are known. These 3-D entities are noted here only to mark the boundary of the planar approach adopted in the remainder of this work.

Outline of matching strategies : The remainder of this chapter surveys three families of techniques for estimating homographies or related parameters. Area-based methods align raw intensity or spectral patches by maximising a similarity score. Feature-based

pipelines employ repeatable keypoints and descriptors, obtained either from explicit detectors or from dense learned features. Direct-regression frameworks predict transformation parameters end-to-end, bypassing explicit correspondence extraction. Each family ultimately yields a tentative correspondence set, which must be converted into a geometry-consistent homography by a robust estimation procedure.

Figure 2.2: *Concept of RANSAC-based mismatch removal. Left: initial correspondences include many outliers. Right: after iterative sampling, RANSAC isolates the inlier set that supports a single geometric model while rejecting the outliers.*



(a) A dataset with many outliers for which a line has to be fitted.

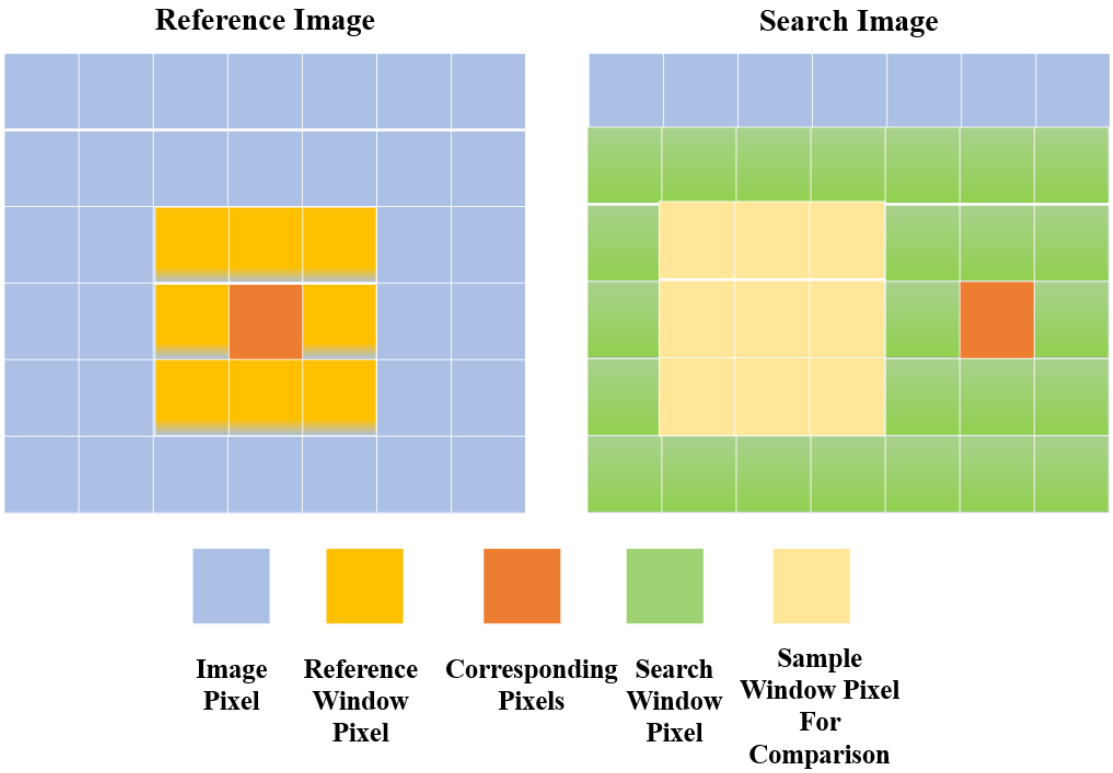
(b) Fitted line with RANSAC; outliers have no influence on the result.

Transformation parameter estimation and mismatch removal : Given tentative matches, outputs of image matching methods, erroneous correspondences are discarded before a reliable homography is fitted. The de-facto standard is Random Sample Consensus (RANSAC) [42], which repeatedly samples minimal subsets, estimates trial homographies, and retains the model with the largest consensus of geometrically consistent matches. Figure 2.2 illustrates the concept: inliers survive the consensus test, whereas outliers are rejected.

Plain RANSAC may become slow or unstable when the outlier ratio is high or when the inlier noise scale is difficult to set. To address these issues, the experimental section adopts MAGSAC++ [43], which replaces the fixed inlier threshold with an adaptive, data-driven bound and introduces an automatic termination rule. MAGSAC++ produces more accurate homography estimates for multispectral pairs without manual parameter tuning. Direct-regression frameworks incorporate robustness internally and therefore avoid a separate RANSAC stage. Additional outlier-rejection strategies—non-parametric model-based methods and relaxed-constraint approaches—are reviewed by Huang et al. [44] but are not applied here.

Next section 2.1.2 now examines area-based image-matching methods in detail, following the order outlined above.

Figure 2.3: *Illustration of an area-based matching algorithm. The reference window from the first image is compared across a search window in the second image to find the most similar region using metrics such as SSD or NCC.*



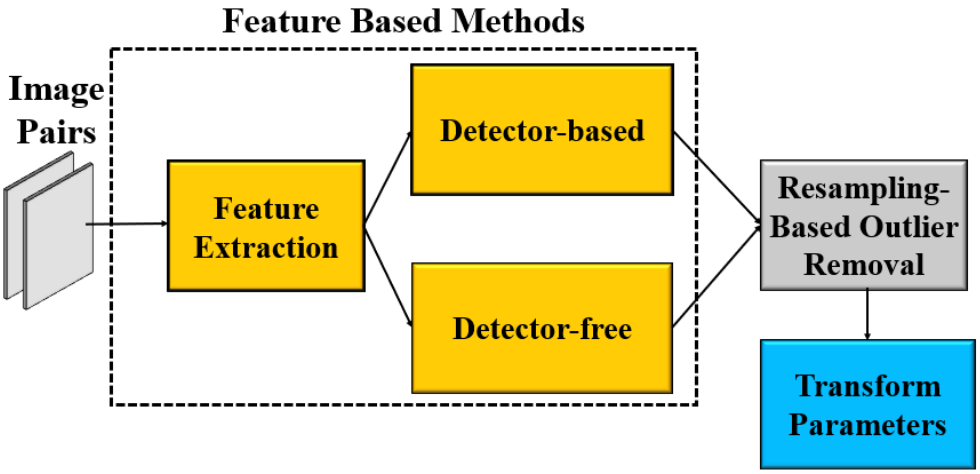
2.1.2 Area-Based Methods

Area-based image matching techniques establish correspondences by directly comparing pixel intensities between two images, without relying on keypoint detection or descriptor extraction. For each pixel in the first image (referred to as the reference image), a fixed-size reference window is defined and compared against multiple candidate regions within a larger search window in the second image (the search image). The objective is to identify the location within the search window that yields the highest similarity with the reference window, according to a predefined metric such as Normalized Cross-Correlation (NCC), Sum of Absolute Differences (SAD), or Sum of Squared Differences (SSD) [45, 46]. As illustrated in Figure 2.3, the reference window (highlighted in orange and yellow) is systematically compared with overlapping windows across the search region, and the location with the highest similarity score is selected as the match. This exhaustive sliding-window process is simple to implement and can be effective in well-textured scenes where radiometric conditions between the images are consistent.

However, the reliance on raw intensity values makes area-based methods highly sensitive to radiometric and geometric distortions caused by changes in illumination, scale, viewpoint, or sensor modality [47]. Their performance tends to degrade in low-texture or repetitive-pattern regions, where many patches may appear similar, increasing the risk of incorrect matches. Larger search windows improve robustness to misalignment but also increase computational cost and the likelihood of false positives. To address these issues, pre-processing steps such as histogram equalization, radiometric normalization, or adaptive window sizing are commonly applied to improve contrast and reduce sensitivity to intensity variations [48, 49]. The final result of this process is a set of pixel-level correspondences, where each pixel (or selected pixel) in the reference image is matched to a location in the search image that yields the highest similarity. These matched coordinate pairs can be used to estimate geometric transformation parameters as discussed in 2.1.1.

Once estimated, these transformations can then be used for downstream applications.

Figure 2.4: *Schematic overview of feature-based image matching pipelines. Image pairs are first passed through a feature extraction module. The resulting features are processed by either detector-based or detector-free branches to estimate correspondences. A resampling-based outlier removal step is then applied to refine the matches, and the final output is the geometric transformation parameters between the input images.*



2.1.3 Feature-Based Methods

Feature-based image matching identifies visual correspondences by detecting and associating distinctive structures that are repeatable across images, despite variations in viewpoint, illumination, or scale. These structures, commonly referred to as interest points or keypoints, form the foundation of a matching pipeline that typically consists of four main stages: feature detection, feature description, feature matching, and geometric transformation estimation. Classical methods use handcrafted detectors and descriptors, while contemporary approaches have embraced deep learning to learn both stages directly from data. Modern pipelines are broadly divided into two categories: detector-based and detector-free methods, as summarized in Figure 2.4 [50, 51].

Detector-based pipelines extract a sparse set of keypoints and compute local descriptors in their neighborhoods. These descriptors are matched across images using nearest-neighbor search or learning based methods [52, 53], often followed by outlier filtering [42, 43]. These methods are interpretable, modular, and compatible with sparse supervision, making them suitable for low-data or self-supervised regimes. However, their effectiveness depends on the presence of repeatable structures, and their performance often degrades in low-texture regions, under occlusions, or in the presence of repetitive patterns. Detector-free methods, on the other hand, remove the keypoint detection stage entirely and instead estimate dense correspondence maps directly from learned image features. These models typically demonstrate better robustness and higher accuracy in complex visual conditions but come with increased computational demands and dependence on dense ground-truth supervision, which is difficult to obtain in many real-world or multispectral settings. Both approaches ultimately rely on a resampling-based outlier removal stage to reject spurious matches and estimate geometric transformation parameters such as the homography or fundamental matrix. The following sections examine the structure, strengths, and limitations of detector-based and detector-free paradigms in greater detail.

2.1.3.1 Detector-based Methods

Detector-based pipelines follow the classical three-stage recipe shown in Figure 2.5. The design of each stage has been refined over two decades of research, and the most widely adopted variants are summarised below.

Keypoint detection and descriptor extraction : The pipeline starts by selecting salient, repeatable points and describing their local appearance. Classical detectors such as SIFT [36], Harris [54], and SURF [3] rely on handcrafted scale–space extrema or corner measures, whereas modern approaches (e.g. SuperPoint [2], R2D2 [55], KeyNet [56])

learn both the detector and the descriptor in a single CNN. Regardless of implementation, the goal is to provide a set of distinctive feature vectors that are robust to modest changes in viewpoint, scale, illumination, and—in the multispectral case—moderate radiometric differences. These features feed directly into the next stage.

Putative match set construction : Given two images, all descriptors from the first image are compared with those from the second to produce tentative correspondences. The default method is the Nearest Neighbor (NN) matcher, which assigns to each descriptor its closest partner in Euclidean or cosine space. Although fast, pure NN is susceptible to local optima and produces a high outlier rate in repetitive or textured regions. Lowe's distance-ratio test, often referred to as the mutual nearest-neighbour rule, mitigates this problem by retaining a correspondence only if the distance to the best match is significantly smaller than the distance to the second-best match [36]. Unlike plain NN, the ratio rule adapts automatically to local descriptor distribution and removes many ambiguous matches without hand-tuning a global distance threshold. Once this filtered set is assembled, the pipeline proceeds to robustness filtering.

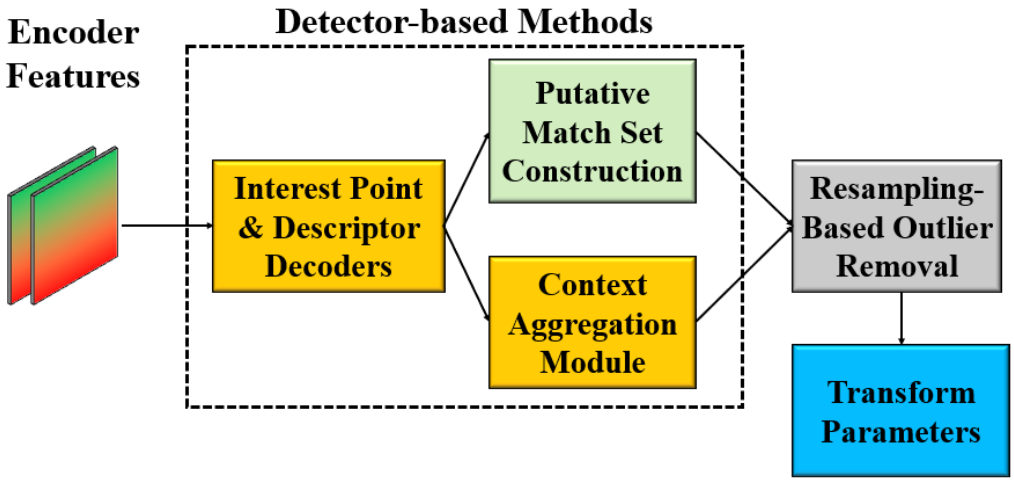
Outlier removal and transformation estimation : After the ratio test, a robust estimator is still required to separate inliers from residual outliers and recover the homography. The standard choice is RANSAC, which samples minimal four-point subsets, fits trial models, and keeps the homography supported by the largest consensus of matches (see Section 2.1.1). This thesis employs the more recent MAGSAC++ variant [43], which replaces the fixed inlier threshold with an adaptive noise bound and stops automatically when further sampling is unlikely to improve the result; the modification yields higher accuracy in high-outlier scenarios and eliminates manual threshold tuning, an advantage for multispectral imagery. Alternative strategies include non-parametric model-based filters and relaxed geometric-constraint frameworks; a comprehensive review can be found

in the survey by Huang et al. [44]. Regardless of the specific filter, the output of this stage is a clean inlier set and the final homography estimate.

Detector-based pipelines remain attractive because every component is interpretable and can be swapped or upgraded independently. Their modularity also allows easy integration of spectrum-specific preprocessing, making them a strong baseline for the multispectral registration experiments that follow.

Figure 2.5 illustrates this workflow, highlighting the key components involved in this pipeline, from feature extraction through descriptor matching to final transformation estimation.

Figure 2.5: *Overview of Detector-Based Methods: Features extracted from the encoder are processed through interest point and descriptor decoders, followed by putative match set construction and context aggregation. Outlier removal is then performed to obtain the final set of robust correspondences used to compute the desired transform parameters.*



With the advent of deep learning, the traditional detector-based workflow has been significantly enhanced through data-driven approaches that often surpass handcrafted methods in performance and generalization. One influential early example is LIFT [57], which uses convolutional neural networks to jointly learn keypoint detection, orientation

estimation, and descriptor extraction in an end-to-end fashion. This approach provides enhanced robustness to variations in illumination, scale, and viewpoint, addressing many limitations of traditional methods.

Modern detector-based pipelines can be broadly categorized into four primary families:

- Detect-then-Describe approaches first identify keypoints and then extract descriptors. Their modular design allows stage-wise optimisation and supports real-time operation when training data are limited; however, performance degrades in low-texture or repetitive regions because reliable keypoint detection becomes difficult.
- Joint Detection and Description models integrate both tasks within a single network. The shared representation improves repeatability and discriminability under significant viewpoint and illumination changes, yet the tight coupling reduces interpretability and can complicate domain adaptation.
- Describe-then-Detect strategies compute dense features before selecting keypoints. This ordering is advantageous in low-texture or repetitive scenes because keypoint selection is guided by rich feature maps, but the dense extraction step increases computational cost and may introduce redundant detections.
- Graph-based Matching methods operate on a pre-computed keypoint set and optimise their spatial relationships in a graph framework. The global reasoning enforces geometric consistency and improves robustness under wide baselines or semantic ambiguity, although it depends on a reliable initial keypoint set and entails additional memory and runtime overhead.

This diverse set of approaches reflects ongoing efforts to enhance descriptor quality, robustness, and generalization, even under challenging real-world conditions with limited or noisy annotations.

Detect-then-Describe Methods : The detect-then-describe paradigm has long served as a foundational framework in local feature matching. It proceeds in a sequential manner: initially detecting salient and repeatable keypoints in the input images, followed by the extraction of local descriptors from patches centered at these keypoints, and concluding with a matching step based on similarity in descriptor space. Classic handcrafted techniques such as Harris [54], SIFT [36], and ORB [58] exemplify this approach, offering robustness across moderate changes in scale, rotation, and illumination.

To enhance correspondence reliability, post-processing strategies like GMS [59] utilize grid-based motion constraints to eliminate false matches, while neural models such as OANet [60] incorporate spatial order and context via attention mechanisms to refine pairwise feature assignments. The advent of deep learning has further enabled end-to-end learning-based alternatives like LIFT [57], which integrates keypoint detection, orientation prediction, and descriptor learning within a single CNN-based pipeline. These models typically employ metric learning losses that encourage compact intra-class and dispersed inter-class descriptor embeddings.

Despite the efficiency and modularity of the detect-then-describe design, its reliance on sparse keypoints can hinder performance in low-texture or highly repetitive scenes. Nevertheless, the approach remains a preferred choice in scenarios requiring interpretability, computational efficiency, and compatibility with legacy systems.

Fully Supervised Methods : Fully supervised learning approaches have redefined local feature matching by replacing handcrafted descriptors with data-driven alternatives trained on annotated datasets. Utilizing convolutional neural networks (CNNs), these methods directly learn discriminative, invariant descriptors robust to scale, viewpoint, and illumination changes.

Pioneering works like L2-Net [61] introduced intermediate supervision and progressive sampling to enhance descriptor quality, while HardNet [62] emphasized metric-based training objectives to simplify learning. Geodesc [63] marked a significant step by incorporating geometric constraints derived from Structure-from-Motion (SfM), aligning descriptor training with 3D reconstruction tasks. SOSNet [64] further refined this line by enforcing second-order similarity constraints, improving the relational consistency of learned features.

Keypoint detection has also evolved with hybrid designs such as Key.Net [56], which combines handcrafted filters with learned CNN features, and ALIKED [65], which introduces differentiable detection for subpixel localization.

To address robustness across modalities and scenes, ContextDesc [66] augments local features with contextual cues, while RIFT [39] and SemLA [67] extend descriptor learning into the frequency and semantic domains, enhancing performance in multispectral and cross-domain scenarios.

Weakly Supervised and Other Methods : Weakly supervised and semi-supervised approaches offer scalable alternatives to fully supervised training by utilizing geometric cues—such as relative camera poses and epipolar constraints—rather than densely annotated correspondence labels. AffNet [68] introduced affine shape learning using a hard-negative constant loss to enhance region-level matchability. GLAM-Points [69] applied reinforcement learning to reward keypoints that contribute to improved image alignment. CAPS [70] utilized camera pose supervision and epipolar constraints within a coarse-to-fine architecture to train dense descriptors. These approaches reduce reliance on extensive manual annotation and exhibit strong generalization across varied environments.

Joint Detection and Description: While sparse feature matching performs reliably under moderate variations, its robustness significantly degrades under extreme conditions such as drastic lighting changes, seasonal differences, or scenes with limited texture. These limitations are often due to the decoupled design of keypoint detectors and local descriptors, where isolated optimization can lead to information loss across stages. Joint Detection and Description (JDD) frameworks address this issue by integrating both tasks into a single, end-to-end trainable architecture, enabling shared feature representations and more coherent optimization. Among the earliest and most influential approaches is SuperPoint [2], which employs a self-supervised pipeline using homographic adaptation to generate pseudo ground-truth keypoints and descriptors. D2Net [71] extends this idea by jointly optimizing both tasks through a unified network, identifying keypoints as local maxima across depth-augmented feature maps. R2D2 [55] complements dense descriptors with a reliability score for keypoints, balancing repeatability and discriminability. Further architectural innovations include RF-Net [72], which introduces receptive field-aware detection, and Reinforced SP [73], which uses reinforcement learning to handle the non-differentiability of keypoint selection. HDD-Net [74] focuses on decoupled yet interactive learning of detection and description modules, while RoRD [75] enhances viewpoint invariance through ensemble-based descriptor correspondence. Recent methods also explore semantic awareness and multi-scale representations. ASLFeat [76] incorporates multi-level feature peaks and deformable convolutions for precise and invariant representations. SFD2 [77] further filters unreliable regions using high-level scene semantics for globally stable feature extraction. Together, these joint models significantly improve robustness and accuracy in challenging visual conditions by allowing feature detection and description to co-evolve within a unified optimization framework.

Describe-then-Detect : In contrast to the conventional detect-then-describe paradigm, the Describe-then-Detect pipeline first computes dense feature descriptors across the image and subsequently identifies keypoints from these descriptors. This reversal allows the model to exploit rich intermediate features to guide keypoint selection. D2D [78] introduced this formulation by generating dense descriptors from convolutional features and selecting keypoints based on absolute and relative saliency within local depth maps. PoSFeat [79] addressed the weakly supervised regime by decoupling the description and detection stages. Using pose supervision, it trained descriptors independently, followed by a keypoint detector informed by descriptor saliency. ReDFeat [80] proposed a mutual weighting scheme to balance detection and description in multispectral learning. SCFeat [81] further improved robustness by introducing shared coupling bridges and cross-normalization layers to facilitate parallel yet consistent optimization of both branches.

Graph-Based Matching : Traditional feature matching pipelines typically rely on nearest-neighbor search in descriptor space, often followed by geometric outlier rejection. More recent methods formulate matching as a structured problem using attention-based graph neural networks (GNNs), where each keypoint acts as a node and relationships are modeled using cross-attention mechanisms. SuperGlue [52] pioneered the use of Transformer-style self- and cross-attention layers in a Graph Neural Network to process sets of keypoints and descriptors jointly, producing a soft assignment matrix through the Sinkhorn algorithm. While highly effective, its quadratic complexity motivated subsequent efficiency-focused designs. SGMNet [82] reduced graph complexity using a seeding mechanism and sparse connections. ClusterGNN [83] partitioned nodes into clusters to suppress redundancy and localize attention. GlueStick [84] extended this framework by integrating both point and line correspondences, while LightGlue [53]

adapted network depth and attention width dynamically per image pair to optimize runtime. DenseGAP [85] used anchor points to propagate context sparsely across the feature map, enhancing robustness in dense scenarios.

2.1.3.2 Detector-free Methods

Despite their effectiveness, detector-based systems can struggle in environments with low structure, modality variations, or wide-baseline viewpoint differences, where repeatable keypoint detection is challenging. To overcome these limitations, **detector-free methods** eliminate the explicit keypoint detection stage entirely, instead focusing on dense correspondences directly computed from learned image features.

Detector-free methods have emerged as a powerful alternative, utilizing dense feature extraction to achieve higher robustness and more complete spatial coverage. These methods are particularly effective in scenarios involving low-texture surfaces, severe occlusions, or modality shifts, where conventional detector-based pipelines tend to fail.

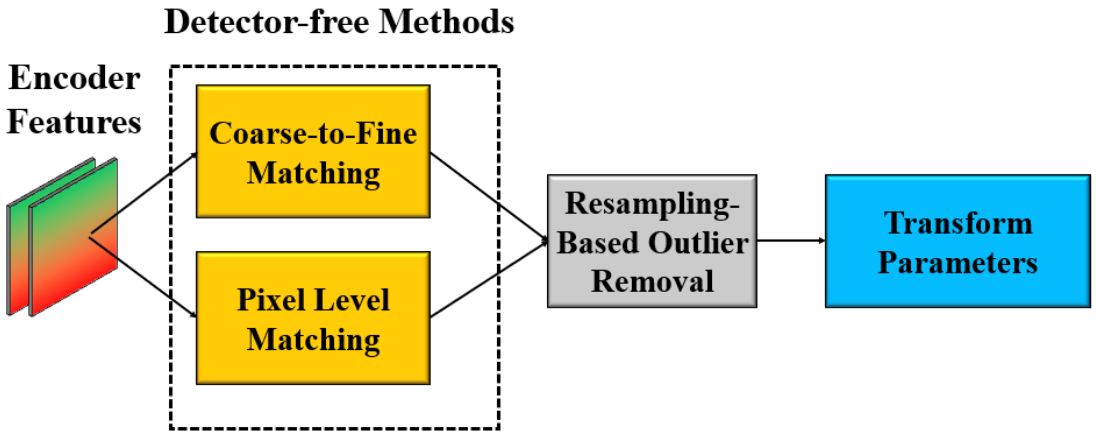
Recent advances in detector-free local feature matching have given rise to two main categories, as illustrated in Figure 2.6:

- **Coarse-to-Fine Matching:** These methods operate in a hierarchical manner, extracting features at multiple scales to balance computational efficiency with accuracy. Initial matches are established at lower resolutions, followed by progressive refinement at finer scales. This approach captures both global context and local detail, making it well-suited for wide-baseline matching.

- **Pixel-Level Matching:** These models compute dense correspondences directly at the original image resolution, providing higher accuracy in regions with large displacements or uniform textures. However, this approach often incurs higher computational costs due to the dense nature of the matching process.

Figure 2.6 provides a high-level overview of these detector-free matching frameworks, capturing the general architectural stages shared across these approaches, including feature extraction, dense matching, and final outlier removal.

Figure 2.6: *Overview of Detector-Free Matching Frameworks: Features extracted from the encoder are processed using either coarse-to-fine matching strategies or pixel-level matching, followed by resampling-based outlier removal to estimate the final transformation parameters.*



Coarse-to-Fine Matching Methods : Coarse-to-fine methods constitute the most prominent class of detector-free matching architectures. These models begin by computing sparse, patch-level correspondences at a lower resolution (coarse stage), which are then refined to subpixel accuracy at higher resolutions (fine stage). This hierarchical design improves robustness in large displacement and low-texture scenarios while maintaining computational efficiency.

DRCNet [86] first introduced this paradigm by generating a 4D correlation volume and refining matches through a learnable neighborhood consensus module. LoFTR [37], a seminal work in this category, incorporated Transformer-based self- and cross-attention to compute dense correspondences with global context, significantly improving performance in textureless regions. SE2-LoFTR [87] extended LoFTR by employing a CNN-based

backbone with enhanced rotation and translation sensitivity.

QuadTree Attention [88] improved scalability by computing attention hierarchically across token pyramids, reducing Transformer complexity from quadratic to linear. MatchFormer [89] adopted a multi-scale attention mechanism that interleaved self- and cross-attention across hierarchical feature levels for robust matching. TopicFM [90] introduced a semantic latent topic model, treating image patches as probabilistic topic distributions to enhance correspondence in semantically consistent regions.

AdaMatcher [91] proposed block-level many-to-one allocation, co-visible region estimation, and refinement modules for robust interaction across overlapping regions. DeepMatcher [92] incorporated SlimFormer, a slimmed Transformer with vector-based attention to enable long-range context aggregation with reduced complexity. OA-Matcher [93] further emphasized overlapping-region awareness, combining region prediction with a weighted label loss to focus matching on co-visible areas.

Other variants, such as EfficientLoFTR [94], DiffRegV2 [95], ShapeFormer [96], XoFTR [97], and JamMa [98], propose improvements in runtime, registration accuracy, or modality adaptation. While coarse-to-fine models effectively balance accuracy and efficiency, they remain sensitive to failure in the fine refinement stage, which can lead to misaligned matches in highly detailed regions.

Pixel-Level Matching Methods : Pixel-level matching methods differ from coarse-to-fine approaches in that they estimate dense correspondences directly at the original image resolution, thereby enabling subpixel-accurate predictions without an intermediate coarse stage. These models aim to compute a dense warp field by estimating matchable pixel pairs across the entire image domain, making them particularly effective in capturing fine geometric details and large displacements.

DGC-Net [99] introduced one of the earliest CNN-based dense correspondence

models, extending optical flow techniques to large transformations. It produced subpixel-accurate predictions using a global correlation volume. Building on this foundation, GLU-Net [100] proposed a unified architecture that integrates both global and local correlation layers to accommodate varying displacement magnitudes. Its adaptive-resolution design allows for flexible input sizes, forming the basis for several subsequent improvements.

Notably, PDC-Net [101] and its successor PDC-Net+ [102] enhanced GLU-Net by incorporating probabilistic modeling to jointly estimate correspondence reliability and flow uncertainty. These models produce dense confidence maps that indicate the trustworthiness of predicted matches, improving robustness in ambiguous regions.

Recent state-of-the-art methods include DKM [103], which introduces a kernelized global matcher and Gaussian embedding decoder for accurate global matching, and RoMA [40], which reframes the correspondence estimation problem through a Markov chain lens. RoMA achieves significant gains in matching accuracy, outperforming many coarse-to-fine models in benchmarks. However, the increased precision of pixel-level matchers comes at the cost of higher computational demands, making them less suitable for real-time or resource-constrained applications.

Other Architectures : Beyond the dominant coarse-to-fine and pixel-level strategies, several detector-free frameworks adopt alternative architectural designs to address unique challenges in feature matching.

DFM [104] employs a pre-trained VGG network as a fixed feature extractor and performs dense matching directly from the final convolutional activations. By utilizing semantic-level features, it bypasses explicit detector modules entirely.

COTR [105] introduces a recursive transformer-based matching pipeline. It progressively refines correspondences by querying image regions around predicted matches

at multiple scales, enabling high-precision alignment through contextual amplification.

ASpanFormer [106] integrates adaptive sampling within a transformer-based matcher. At each attention stage, it estimates uncertainty and dynamically generates local sampling grids around high-confidence regions, ensuring that fine-grained correspondences are derived from regions with high correlation.

MIVI [107] targets multispectral correspondence, particularly between infrared and visible spectra. Its multi-stage design captures long-range dependencies and fine-grained local structure, making it especially effective under severe modality gaps and low-texture conditions.

These frameworks highlight the growing trend of incorporating task-specific priors or architectural innovations to overcome the inherent limitations of traditional detector-free pipelines.

2.1.3.3 Summary of Feature-Based Methods

Feature-based methods constitute a fundamental class of image matching techniques that rely on establishing correspondences through local visual features. Detector-based pipelines remain widely used due to their interpretability, modularity, and compatibility with sparse training supervision. They span handcrafted and deep learning variants, including recent advances in joint detection and description, weak supervision, and graph-based optimization. However, their dependence on repeatable keypoints limits performance in textureless or severely distorted scenes. In contrast, detector-free approaches eliminate the need for explicit keypoint detection, enabling dense correspondence estimation directly from learned features. These methods offer improved robustness and matching completeness, particularly in multispectral or low-structure settings, albeit at the cost of increased computational demand and supervision complexity. Taken together,

the field has witnessed a shift from sparse, modular frameworks to integrated, dense architectures—driven by advances in CNNs, transformers, and large-scale self-supervised learning—with each paradigm presenting distinct trade-offs between accuracy, efficiency, and applicability. Table 2.2 provides an overview of landmark methods in this evolution.

Table 2.2: *Representative local feature matching methods. 2024–2025 methods marked with* †.

Method	Year	Type	Key contributions
SIFT [36]	2004	Det.	Classic handcrafted desc.; scale/rotation invariant
HardNet [62]	2017	Det.	Metric-learned, highly discriminative descriptor
SuperPoint [2]	2018	Det.	Self-supervised joint detector–descriptor
SuperGlue [52]	2020	Det.	Transformer matcher with OT attention
LoFTR [37]	2021	D-free	Dense transformer; no keypoints
DKM [103]	2023	D-free	Global kernelized matcher, Gaussian decoder
RoMa [40]	2024†	D-free	Markov-chain robust dense matching
LightGlue [53]	2024†	Det.	Dynamic sparse matching with learned confidence
XoFTR [97]	2024†	D-free	Cross-spectral transformer, sub-pixel refine
XPoint [41]	2024†	Det.	Self-sup. multispectral keypoints + homography
JamMa [98]	2025†	D-free	Modality-aware dense matcher with DA refinement

Despite continual progress, all feature-based pipelines still depend on an explicit correspondence set and a subsequent robust estimator. Recent learning-based research therefore explores a complementary strategy: predicting the geometric transformation itself—most often the homography—directly from paired images, without forming discrete matches. This family of direct homography regression methods is introduced next.

2.1.4 Direct Homography Regression

Classical matching pipelines detect keypoints, compute descriptors, match them across views, and fit a homography with RANSAC. Each step can fail: detectors miss repeatable points in low-texture scenes, descriptors drift under spectral shifts, and robust sampling adds computational overhead. Direct Homography Regression (DHR) removes these intermediates by feeding an image pair—or two aligned patches—into a network

that outputs the eight independent homography parameters (or an unnormalised 3×3 matrix later scaled). The simplification reduces latency and code complexity, but at the cost of interpretability and a risk of over-fitting to the training distribution.

Early analytic solutions such as the Lucas–Kanade aligner minimise a photometric error iteratively; they require good initialisation and assume brightness constancy, making them fragile under illumination change. HomographyNet [108] demonstrated that a shallow CNN trained on synthetically warped patches can predict the homography in a single forward pass, sparking modern DHR work. Hierarchical encoders [109] learn from labelled patch pairs, expanding the receptive field while keeping inference fast. Unsupervised or self-supervised models such as SSR-Net [110] replace labels with reconstruction losses that warp one image by the predicted homography and minimise pixel-wise error; hybrid frameworks [111] start with synthetic supervision and then refine using unsupervised objectives.

Recent architectures follow several trends. Hierarchical feature pyramids [109] enable coarse-to-fine refinement; attention modules concentrate computation on salient regions [112]; and recurrent or iterative decoders, implemented with GRU or gradient-based updates [113, 114, 115, 116], progressively polish the estimate. Spatial-pyramid pooling, as used in SSR-Net, improves robustness to scale change.

DHR performs well where keypoint repeatability is low—texture-poor, repetitive, or cross-spectral scenes—and its unsupervised variants remain attractive when ground-truth correspondences are scarce. High accuracy, however, often depends on dense supervisory signals or strong geometric priors, and compute cost increases with attention blocks or recurrence though lightweight versions mitigate this. Extending a monolithic regressor beyond planar scenes and diagnosing its failure modes also remain open challenges. Despite these limitations, DHR offers a label-efficient alternative for homography

estimation in learning-based VSLAM, multispectral image matching, and embedded vision systems.

2.1.5 Summary of Image Matching Methods

Current image-matching research can be grouped into three ideas that range from simple pixel rules to learned, context-aware models. Area-based methods slide a window over the image and search for the best intensity match. They are straightforward and need no features, but fail when lighting, scale, or spectrum changes. Feature-based pipelines are still the main choice in many vision systems. Detector-based versions are easy to interpret and can be tuned stage by stage, yet they struggle when a scene lacks strong corners or edges. Detector-free networks skip keypoints and predict dense matches from learned features; coarse-to-fine designs balance speed and accuracy, while pixel-level models reach sub-pixel precision at higher cost. Direct homography regression removes matching altogether. A network maps a pair of images directly to the eight homography values. This approach works well in texture-poor or repetitive areas and can train without ground-truth matches, but it often overfits to the training domain and offers limited insight into its errors. No single family is best in every situation. Modern systems often mix ideas, for example by pairing learned features with a classical RANSAC fit or adding a direct-regression module for difficult regions.

Multispectral registration adds extra challenges: large brightness shifts, sensor-specific distortions, and scarce labelled data. The next section examines how the methods above are adapted, or newly designed, to handle these multispectral problems.

2.2 Multispectral Image Matching

Extending image matching techniques to the multispectral domain introduces a set of unique challenges that arise due to heterogeneous sensing modalities. These include non-linear radiometric distortions, spectrum-dependent noise characteristics, and the lack

Table 2.3: *Recent multispectral image–matching methods, grouped by pipeline category. Columns list the publication year, the high-level method type (descriptor learning, detector-based, detector-free, direct regression, or "universal" cross-modal models), and a brief note on each method's main technical contribution.*

Method	Year	Type	Remarks
CoMIR [117]	2020	Desc. learn.	Contrastive embeddings for mono-modal registration
MultiPoint [1]	2022	Detector	Homographic adaptation across spectra
RIFT2 [118]	2023	Detector	Fast rotation-invariant descriptor, NRD robust
RoMa [40]	2023	Det-free	Dense matcher on foundation features
XPoint [41]	2024	Detector	Self-supervised keypoints+ homography
XoFTR [97]	2024	Det-free	Cross-modal transformer, sub-pixel refine
POS-GIFT [12]	2024	Regression	Geo-intensity invariant homography
MINIMA [119]	2025	Universal	Synthetic data scaling for any modality
MatchAnything [120]	2025	Universal	Large-scale pretrain; > 8 unseen modalities

of well-defined ground truth keypoints across domains. Conventional methods that rely on pixel-level similarity or handcrafted features often fail in these scenarios due to substantial differences in spectral appearance. This section adopts the taxonomy previously introduced—area-based, feature-based, and direct regression-based approaches—and reviews methods specifically adapted or designed for the task of multispectral image registration. Each category is discussed in terms of its adaptability to spectral variation, supervision strategy, and matching performance across domains.

Area-based multispectral matching methods extend classical similarity measures by incorporating descriptors and metrics that are more robust to radiometric inconsistencies. Traditional approaches rely on Normalized Cross-Correlation (NCC) or Mutual Information (MI), which are commonly used in medical and remote-sensing image registration. Hybrid similarity measures such as EMPCA-MI [121], RSNCC [122], and HOPC[123] improve alignment under extreme modality gaps by using statistical structure or phase congruency. PCA-based hybrids have shown improved robustness in retinal

imaging, while morphological-gradient methods [124] and multi-scale phase-congruency descriptors [123] offer structure-aware metrics to suppress modality-specific variance. These methods, however, are typically computationally expensive and sensitive to initialization errors.

Feature-based techniques dominate multispectral registration due to their flexibility and generalization capabilities. Hand-crafted descriptors such as SIFT [36] and SURF [3] are widely used but often degrade under cross-spectral variation. Spectral-specific variants like SAR-SIFT [125], GA-SURF[126], and C-SIFT [124] address modality-induced distortions through structural priors and rotation-aware design. RIFT2 [118] achieves invariance to non-linear intensity changes, establishing itself as a strong baseline in this domain. Other hand-crafted descriptors—such as HOPC [123], HoDM[127], HGEO [128], and HPCG [129]—combine structural and textural cues that remain stable across modalities. Co-occurrence-aware approaches [130] and phase-consistent gradients[129] further improve robustness in noisy or low-texture scenes, though most hand-crafted solutions are limited by heuristic design and lack adaptability to new modalities.

Learning-based methods have introduced significant advances by utilizing deep architectures and data-driven optimization. Self-supervised pipelines such as MultiPoint [1] and XPoint [41] exploit homographic adaptation and cross-spectral augmentation to overcome the lack of cross-modal ground truth. XPoint integrates a VMamba encoder with a modular decoder for keypoint detection, description, and geometric regression. Contrastive-learning frameworks like CoMIR [117] and SDNet [131] construct shared embedding spaces for multispectral pairs using contrastive or self-distillation losses. RFNet [72] combines coarse-to-fine alignment with image-fusion feedback, achieving mutually reinforcing optimization. Transformer-based models such as XoFTR [97] and RoMa [40] learn dense cross-modal correspondences using masked-image modeling or

hybrid pyramids, while GRiD [132] combines detector-free correspondence prediction with guided refinement and phase-congruency anchors.

Within learning-based frameworks, two key design choices emerge: detector-based and detector-free. Detector-based methods like XPoint [41] and Multi-Point [1] retain explicit interest-point detection while incorporating modality-invariant training strategies such as spectrum-consistent pseudo-labeling. Detector-free approaches, including GRiD [132], RoMa [40], and XoFTR [97], bypass keypoint detection entirely by inferring dense correspondences via transformers, deformable attention, or hierarchical feature aggregation. These models improve spatial coverage and robustness but often incur higher computational complexity.

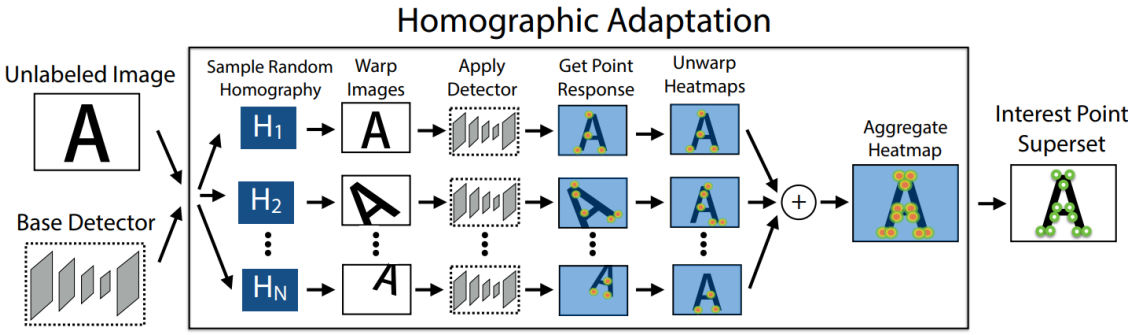
Direct homography regression has also emerged as a viable alternative, particularly in low-texture or high-distortion settings where feature-based methods struggle. Models such as SSR-Net [110], XPoint [41], and POS-GIFT [12] learn to estimate geometric transformations directly from paired inputs using photometric or geometric loss functions. POS-GIFT further introduces a modality-guided sampling strategy and position, orientation and scale invariant constraints to improve robustness under severe nonlinear distortions. GAN-based translation methods [133] attempt to normalize modality gaps prior to registration, but often at the cost of interpretability and data efficiency.

Universal matchers trained across multiple modality pairs have recently gained momentum. MINIMA [119] introduces synthetic modality scaling and large-scale training on a generative multispectral dataset, enabling strong cross-domain generalization. MatchAnything [120] adopts large-scale pretraining using synthetic supervision to create a single network capable of generalizing to over eight unseen cross-modal registration tasks. These methods mark a paradigm shift from modality-specific to modality-agnostic architectures, with growing emphasis on zero-shot performance.

Collectively, these methods reflect a transition from traditional handcrafted

pipelines to adaptive, self-supervised, and modality-agnostic architectures, as summarized in Table 2.3. This growing body of research demonstrates a clear shift toward learning-based solutions capable of generalizing across a wide range of spectral domains. The following chapter builds on this foundation to introduce a novel self-supervised registration framework designed for robust and scalable multispectral image matching.

Figure 2.7: *Homographic adaptation in SuperPoint [2]. An unlabeled image is warped by random homographies $\{H_i\}$; each warped view passes through the base detector trained on synthetic-data, MagicPoint, the responses are unwarped with H_i^{-1}, averaged, and non-maximum-suppressed to yield a viewpoint-invariant keypoint superset.*



2.3 Homographic Adaptation

Homographic adaptation is a self-supervised technique that transforms a viewpoint-variant detector into a viewpoint-invariant one by aggregating its predictions across synthetically warped versions of the same image [2]. It forms the backbone of the SuperPoint pipeline, enabling joint learning of keypoint detection and description without manual annotations (see Fig. 2.7).

The process begins with MagicPoint, a corner detector trained on approximately 30 k synthetic glyph images. While effective on clean inputs, MagicPoint struggles with real-world images due to its limited exposure to natural photometric and geometric variations. Homographic adaptation bridges this gap by generating pseudo ground truth keypoints from real scenes, allowing the detector to be fine-tuned in a self-supervised manner

using datasets like Microsoft Common Objects in Context (MS-COCO).

Algorithm 1 outlines the procedure, where each step aligns with the blocks in Fig. 2.7. The method initializes an empty heatmap $\hat{F}$, then iteratively samples random planar homographies. Each warped image $I_i = \mathcal{H}_i(I)$ is processed by the base detector f, yielding a response map S_i, which is then unwarped using $\mathcal{H}_i^{-1}$. These are aggregated into $\hat{F}$, which is finally averaged to produce the output heatmap that represents pseudo keypoint probabilities.

Algorithm 1 Homographic Adaptation (SuperPoint)

Require: unlabeled image I; base detector f; number of homographies N_h
Ensure: pseudo keypoint heatmap $\hat{F}$
$\hat{F} \leftarrow \mathbf{0}$ {initialize aggregated heatmap}
for $i = 1$ **to** N_h **do**
  sample planar homography $\mathcal{H}_i$
  $I_i \leftarrow \mathcal{H}_i(I)$ {warp image}
  $S_i \leftarrow f(I_i)$ {run MagicPoint}
  $\tilde{S}_i \leftarrow \mathcal{H}_i^{-1}(S_i)$ {unwarp response}
  $\hat{F} \leftarrow \hat{F} + \tilde{S}_i$
end for
$\hat{F} \leftarrow \hat{F}/N_h$ {pixel-wise average}
return $\hat{F}$

The aggregation step can be compactly written as:

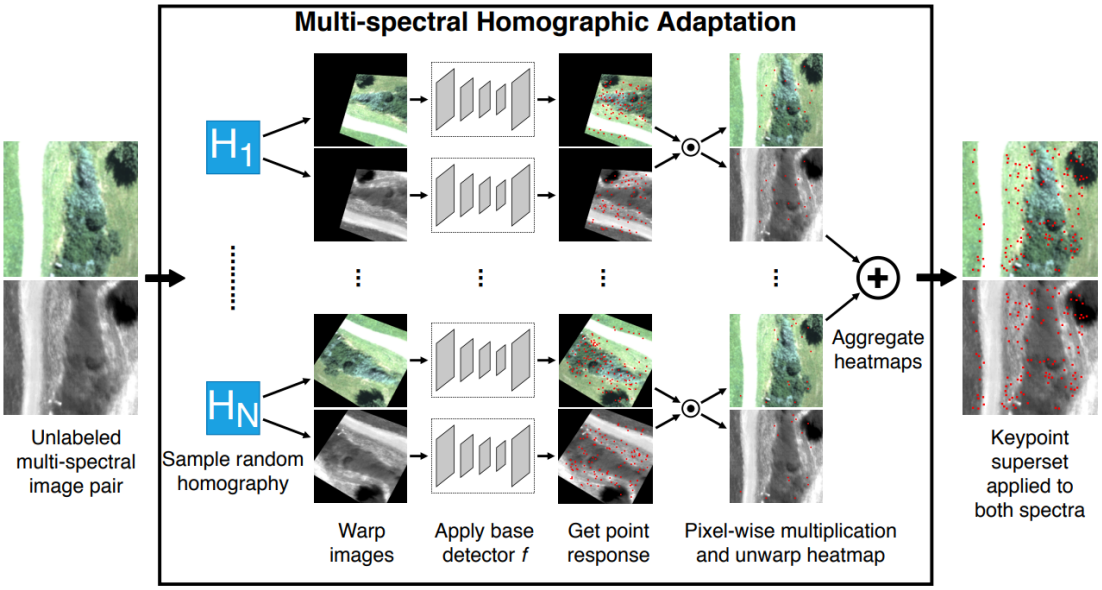
$$\hat{F}(I; f) = \frac{1}{N_h} \sum_{i=1}^{N_h} \mathcal{H}_i^{-1} \, f\big(\mathcal{H}_i(I)\big), \tag{2.2}$$

where $\hat{F}(I; f)$ denotes the viewpoint-invariant heatmap of keypoint activations. For an ideal homography-covariant detector f, the equality $\mathcal{H}(f(I)) = f(\mathcal{H}(I))$ would hold. Equation (2.2) empirically approximates this behavior by averaging over multiple randomly sampled homographies.

Applying homographic adaptation to large unlabeled datasets such as MS-COCO yields dense pseudo-labels for keypoints, enabling self-supervised fine-tuning of SuperPoint. This process allows the network to learn more robust and transferable features

without relying on manual annotations, making it a scalable solution for diverse image matching tasks. Although highly effective, the technique assumes planar scene geometry and may introduce border artifacts due to strong warps or image cropping. Nonetheless, it remains a foundational component in self-supervised training pipelines for keypoint detection.

Figure 2.8: *Multispectral homographic adaptation in MultiPoint [1]. A random homography $\mathcal{H}_i$ is applied to each image of an aligned optical–thermal pair. Detector responses are computed independently, multiplied pixel-wise to keep only cross-spectrally consistent peaks, unwarped with $\mathcal{H}_i^{-1}$, and averaged over N_h samples to form a robust interest-point superset valid in both spectra.*



2.3.1 Multispectral Homographic Adaptation

MultiPoint [1] generalizes SuperPoint's homographic adaptation to aligned multispectral pairs, most commonly visible–thermal imagery. The key modifications are: (1) a spectrum-consistency filter, and (2) a final Gaussian smoothing step. After warping both images with the same homography, the two detector heatmaps are multiplied element-wise so that only locations active in *both* spectra are retained. To further stabilize the aggregated heatmap and suppress noise, a Gaussian smoothing filter with kernel size

$K_{\mathrm{gb}} = 3$ is applied after averaging. Figure 2.8 and Algorithm 2 present the complete procedure.

Algorithm 2 Multispectral Homographic Adaptation (MultiPoint)

Require: aligned optical image I_o, thermal image I_t, base detector f, number of homographies N_h
Ensure: smoothed pseudo keypoint heatmap $\hat{F}_G$
1: $\hat{F} \leftarrow \mathbf{0}$ {initialize aggregated heatmap}
2: **for** $i = 1$ **to** N_h **do**
3: sample planar homography $\mathcal{H}_i$ {same limits as single-spectrum HA}
4: $I_{o,i} \leftarrow \mathcal{H}_i(I_o), \quad I_{t,i} \leftarrow \mathcal{H}_i(I_t)$
5: $S_{o,i} \leftarrow f(I_{o,i}), \quad S_{t,i} \leftarrow f(I_{t,i})$
6: $\tilde{S}_i \leftarrow \mathcal{H}_i^{-1}\big(S_{o,i} \odot S_{t,i}\big)$ {element-wise $\odot$ enforces cross-spectral agreement}
7: $\hat{F} \leftarrow \hat{F} + \tilde{S}_i$
8: **end for**
9: $\hat{F} \leftarrow \hat{F}/N_h$ {pixel-wise average}
10: $\hat{F}_G \leftarrow \text{Gaussian}(\hat{F}, K_{\mathrm{gb}} = 3)$ {apply 2D Gaussian smoothing with kernel size 3}
11: **return** $\hat{F}_G$

The aggregation step in line 7 can be written compactly as:

$$\hat{F}(I_o, I_t; f) = \frac{1}{N_h} \sum_{i=1}^{N_h} \mathcal{H}_i^{-1}\big(f(\mathcal{H}_i(I_o)) \odot f(\mathcal{H}_i(I_t))\big), \tag{2.3}$$

where $\odot$ denotes element-wise multiplication.

Compared with the single-spectrum average in Eq. (2.2), the multiplicative fusion retains only interest points that appear in *both* modalities, thereby increasing cross-spectral repeatability while discarding modality-specific responses.

Although this may suppress some valid detections, the resulting smoothed heatmap $\hat{F}_G$ is highly reliable for training. The next section introduces a window-based operator that relaxes the strict pixel-wise agreement in Eq. (2.3), yielding denser and more complete cross-spectral keypoint sets.

3. METHODOLOGY

3.1 Overview

The problem of multispectral image matching presents significant challenges due to the diversity of sensor modalities, illumination conditions, and spectral pixel intensity differences. Existing approaches often struggle to generalize beyond the modality they are trained on, particularly in cross-spectral scenarios such as the visible–thermal case.

To address these challenges, this thesis proposes a two-stage development:

- First, we introduce an **Improved Homographic Adaptation** method tailored for the visible–thermal domain. By reformulating the pseudo ground-truth keypoint generation process with a spectrum-aware windowing mechanism, we significantly improve the repeatability and matchability of interest points across modalities. This enables robust training of detector-based image matching frameworks without relying on manual labels or expensive annotations such as depth maps or camera poses.
- Second, we generalize the insights gained from this domain-specific solution into a fully self-supervised, modular architecture called **XPoint**. XPoint employs a Visual-State-Space [134] based backbone for feature extraction and integrates specialized heads for keypoint detection, descriptor generation, and homography estimation. It operates solely on aligned image pairs, requiring no additional supervision such as depth, pose, or dense correspondences, and is designed to scale across diverse spectral combinations, including VIS-TH, VIS-NIR, VIS-LWIR, and VIS-SAR.

This two-stage approach first addresses a critical cross-spectral matching case to establish strong keypoint supervision, and then extends the same principle into a scalable, spectrum-agnostic framework for general multispectral image matching.

The following section introduces our improvements to homographic adaptation, which serve as the cornerstone for this unified methodology.

3.2 Improved Multispectral Homographic Adaptation

3.2.1 Motivation

Image matching techniques have traditionally followed two paradigms: pixel-based and feature-based approaches. While pixel-based methods can achieve high accuracy [135], their computational burden makes them unsuitable for real-time applications. In contrast, feature-based methods such as SIFT [36] and SURF [3] offer computational efficiency and robustness for multi-view image matching. However, their performance degrades significantly in multispectral scenarios due to the nonlinear intensity differences across spectral modalities.

Recent advances in data-driven methods have shown strong performance under challenging conditions involving varying viewpoints and illumination. One such method is SuperPoint [2], which employs a technique called *Homographic Adaptation* to learn viewpoint-invariant keypoints in a self-supervised manner. This technique simulates different viewpoints by applying random homographies to unlabeled images and aggregating the resulting keypoint responses, enabling the model to generalize to new views.

Building on this idea, MultiPoint [1] extended SuperPoint's architecture to tackle multispectral image alignment, particularly in the visible–thermal setting. They generalized Homographic Adaptation to operate on aligned multispectral image pairs and introduced a multiplicative fusion step, where the detector responses from each spectrum are multiplied element-wise. This fusion retains only keypoints that are active in both spectra, ensuring high cross-spectral consistency. Additionally, they released a UAV-captured dataset of thermal-optical image pairs under various conditions.

As noted in Section 2.3.1, this strict pixel-wise multiplication strategy enforces

hard spatial alignment and filters out keypoints that are not perfectly matched between modalities. Although Gaussian smoothing is applied to reduce small localization errors, the method still discards many salient features that are present in only one spectrum or slightly misaligned due to sensor-specific distortions. This leads to a sparse supervision signal and limits the effectiveness of the generated keypoint set.

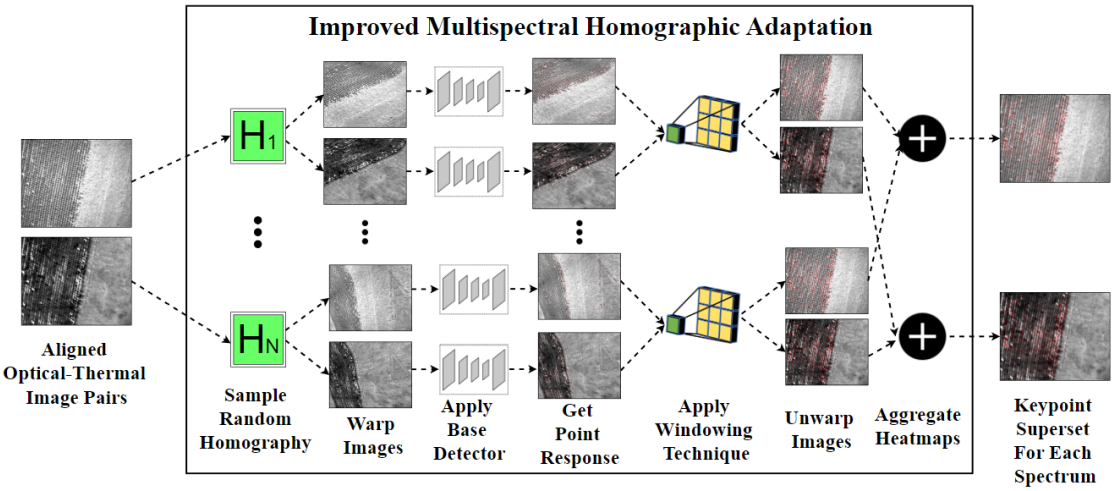
To address this limitation, we propose *Improved Multispectral Homographic Adaptation* (IMHA), a novel windowing-based decision mechanism for cross-spectral Homographic Adaptation. Rather than requiring exact pixel-level agreement, our method accepts a keypoint from one modality if it lies within a defined spatial window of a corresponding keypoint in the other modality. This spatial tolerance accounts for small cross-spectral misalignments and sensor-induced noise, leading to a denser, more representative keypoint superset. Moreover, it allows finer control over localization error during training and improves the robustness of the learned detector.

Contributions. Our contributions can be summarized as:

* We propose the **Windowing Technique**, which determines cross-spectral keypoint consistency based on spatial proximity rather than strict pixel-wise alignment, improving pseudo-label generation in the training of MultiPoint.
* Unlike the original approach that fuses keypoints into a single binary map, our method generates separate keypoint supersets for each modality, preserving modality-specific features while ensuring geometric consistency.
* The improved keypoints lead to stronger supervision, enabling the retrained MultiPoint model to outperform the original baseline and other tested algorithms on two benchmark datasets for optical–thermal image registration.

Building on these observations, we now present our proposed windowing-based improvement tailored specifically for multispectral settings.

Figure 3.1: *Pipeline of the proposed Improved Multispectral Homographic Adaptation for generating robust pseudo-ground truth keypoints across modalities. Given aligned optical–thermal image pairs, random homographies are applied and the warped images are processed by a base keypoint detector (e.g., SURF [3]). A windowing-based acceptance criterion is used to tolerate spatial shifts between modalities. The resulting filtered keypoint maps are unwarped and aggregated to produce separate training labels for each spectrum.*

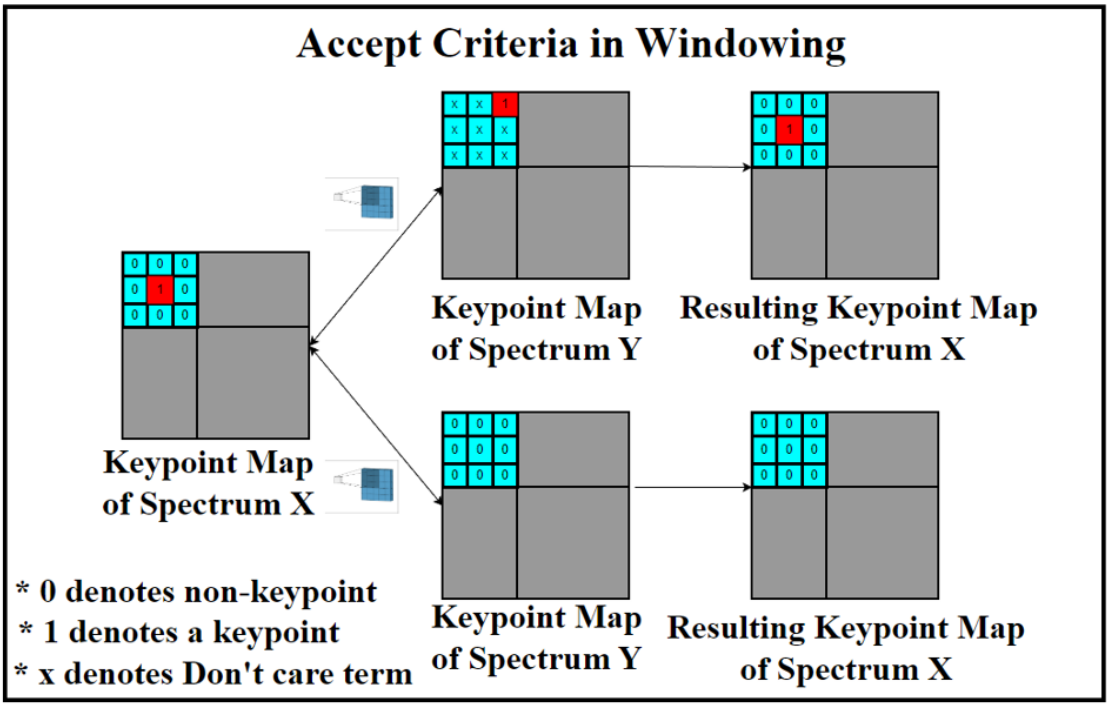


3.2.2 Proposed Approach

We propose an improved version of Multispectral Homographic Adaptation that incorporates a spatially tolerant acceptance criterion through a windowing strategy. The key idea is to accept keypoints from each modality if they fall within a predefined spatial window centered at a corresponding location in the other spectrum. This compensates for misalignments caused by cross-spectral intensity variations, resolution mismatches, and sensor-specific distortions. Relaxing the cross-spectrum supervision to allow small spatial deviations yields denser and more robust pseudo-labels, ultimately improving the model's ability to generalize across modalities. A detailed definition of this acceptance rule is given analytically in Equation (3.1) and implemented in Algorithm 3.

In the Improved Homographic Adaptation depicted in Figure 3.1, random homographies are applied to multispectral image pairs before passing them through a base

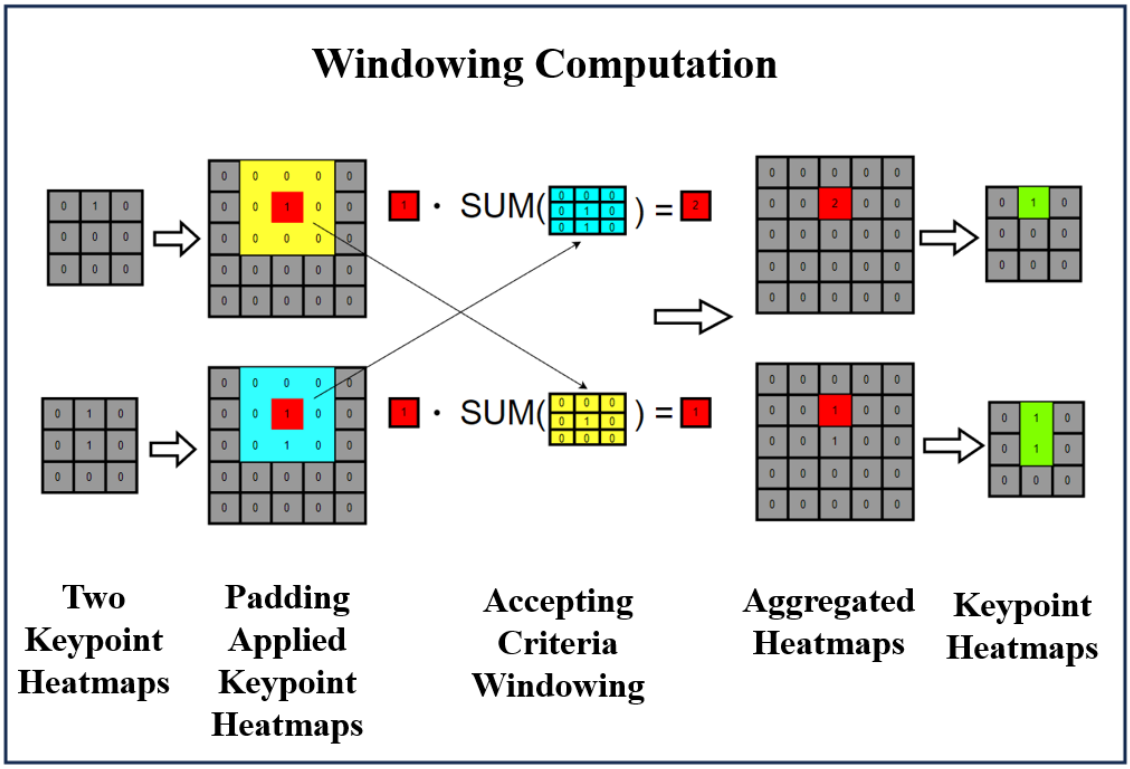
Figure 3.2: *Illustration of the keypoint acceptance criterion in the windowing strategy. A candidate keypoint from Spectrum X is accepted only if there exists at least one keypoint within a predefined window region in the corresponding location of Spectrum Y. The figure shows two scenarios: in the top row, the candidate keypoint is accepted due to spatial proximity to a matching keypoint in Spectrum Y; in the bottom row, the candidate is rejected due to the absence of nearby keypoints. Here, 1 denotes a keypoint, 0 a non-keypoint, and x represents a 'don't care' region outside the search window.*



keypoint detector. The resulting heatmaps are filtered using a window-based consistency check: a keypoint in the optical image is accepted if there exists a corresponding thermal keypoint within a local window centered on the same location—and vice versa (see Figure 3.2). Padding is applied to both heatmaps, and for each candidate keypoint, a sliding window is used to compute a local sum over the other modality. This enables spatially tolerant filtering, addressing modality-specific localization noise (see Figure 3.3).

Let H_{opt} and H_{th} denote binary keypoint heatmaps for the optical and thermal images, respectively. Their zero-padded versions of size $(n + w - 1) \times (m + w - 1)$ are denoted by $\hat{H}_{opt}$ and $\hat{H}_{th}$. Let $W_{\hat{H}_{th}}$ represent the set of all $w \times w$ local windows extracted

Figure 3.3: *Illustration of the windowing computation used for keypoint aggregation. Two keypoint heatmaps from different spectral modalities are compared using a sliding window strategy. For each candidate keypoint (value 1), a local patch is extracted from the other modality's heatmap and summed. If the sum exceeds a threshold (e.g., at least one overlapping keypoint), the keypoint is accepted. Aggregated heatmaps are then thresholded to produce the final keypoint maps. The figure shows padding, window-based comparison, and final acceptance logic applied to two sample cases.*



from $\hat{H}_{th}$, centered on each pixel of the original (unpadded) grid. The same procedure is applied symmetrically for computing B_{th}. The filtering operation is defined as follows:

$$
\begin{aligned}
F_{opt}[k,l] &= \sum_{j}^{w}\sum_{i}^{w} W_{\hat{H}_{th}}[k,l,i,j] \\
H_{opt\text{-}filtered} &= H_{opt} \odot F_{opt} \\
B_{opt} &= H_{opt\text{-}filtered} \geq 1
\end{aligned}
\tag{3.1}
$$

Here, F_{opt} counts the number of keypoints from the thermal modality within a local window centered at each optical keypoint, and B_{opt} is the final binary acceptance map, marking points that meet the window-based acceptance condition.

Algorithm 3 Improved Multispectral Homographic Adaptation (IMHA)

Require: aligned optical image I_o, thermal image I_t, base detector f, number of homographies N_h, window size w
Ensure: binary keypoint maps B_o, B_t for both modalities

$\hat{F}_o \leftarrow \mathbf{0}, \quad \hat{F}_t \leftarrow \mathbf{0}$ {initialize aggregate heatmaps}
for $i = 1$ to N_h **do**
  sample homography $\mathcal{H}_i$
  $I_{o,i} \leftarrow \mathcal{H}_i(I_o), \quad I_{t,i} \leftarrow \mathcal{H}_i(I_t)$ {warp both modalities}
  $S_{o,i} \leftarrow f(I_{o,i}), \quad S_{t,i} \leftarrow f(I_{t,i})$ {apply base detector}
  $(B_{o,i}, B_{t,i}) \leftarrow \texttt{WindowingFilter}(S_{o,i}, S_{t,i}, w)$ {cross-modal window filtering}
  $\tilde{B}_{o,i} \leftarrow \mathcal{H}_i^{-1}(B_{o,i}), \quad \tilde{B}_{t,i} \leftarrow \mathcal{H}_i^{-1}(B_{t,i})$ {unwarp to original coordinates}
  $\hat{F}_o \leftarrow \hat{F}_o + \tilde{B}_{o,i}, \quad \hat{F}_t \leftarrow \hat{F}_t + \tilde{B}_{t,i}$
end for
$\hat{F}_o \leftarrow \hat{F}_o / N_h, \quad \hat{F}_t \leftarrow \hat{F}_t / N_h$ {normalize across homographies}
$B_o \leftarrow \hat{F}_o \geq 1, \quad B_t \leftarrow \hat{F}_t \geq 1$ {final binary keypoint maps}
return B_o, B_t

We benchmark our proposed IMHA method against two established variants of homographic adaptation. The first, SuperPoint [2], achieves viewpoint invariance by averaging detector responses over randomly sampled planar homographies (Section 2.3, Algorithm 1). The second, MultiPoint [1], extends this idea to multispectral pairs via *multiplicative fusion*, retaining only those keypoints that are *exactly* co-localized in both modalities (Algorithm 2, Eq. (2.3)). While highly precise, this strict pixel-wise product yields an overly sparse label set.

IMHA addresses this limitation through a spatially tolerant **windowing rule**: a candidate keypoint is accepted if at least one corresponding keypoint from the opposite modality appears within a $w \times w$ local neighborhood. This behavior is illustrated in Figures 3.2 and 3.3, formalized in Equation (3.1), and implemented in Algorithm 3.

Details of Ground Truth Data Extraction: SURF [3] is selected as the base detector for generating pseudo ground truth keypoints, based on its effectiveness reported in the MultiPoint [1] framework. The detector is configured with Hessian threshold parameters set to $C_1 = 1500$ and $C_2 = 300$, which provide stable keypoint detection across spectral modalities. For homographic adaptation, we apply $N_h = 10$ random planar homographies per image pair. These homographies are generated by composing basic geometric transformations, including translation ($\pm 5\%$), scale ($\pm 20\%$), in-plane rotation ($\pm 90°$), and symmetric perspective distortion ($\pm 20\%$) in width and height (see Figure 3.5). This results in a diverse set of realistic image deformations. To avoid artificial boundary artifacts during warping, border reflection is applied in both training and testing. The proposed windowing-based acceptance strategy uses a window size of 5×5, which was empirically found to offer the best balance between keypoint density and matching reliability, as shown in Table 4.4.

By allowing limited cross-spectral displacement of matching keypoints, IMHA produces *denser* and more representative pseudo-labels, improving both keypoint repeatability and downstream matching performance (see Tables 4.2 and 4.4).

In summary, IMHA mitigates the sparsity issue of MultiPoint while preserving cross-spectral consistency. This richer supervisory signal forms the basis of our fully self-supervised **XPoint** architecture (Section 3.3), enabling scalable training across diverse spectral combinations.

3.3 XPoint: A Modular Self-Supervised Architecture for Multispectral Matching

3.3.1 Motivation

State-of-the-art image matching methods include hand-crafted and learning-based techniques. Hand-crafted methods like SIFT [36] and SURF [3] detect keypoints robust to scale, rotation, and illumination changes but require tuning, are computationally

heavy, and struggle with extreme viewpoint and texture variations in multispectral scenarios. The advent of deep learning has led to substantial improvements in addressing these challenges, particularly within the visible spectrum, through learning-based approaches [2, 52, 37, 106, 84]. Despite these advancements, many learning-based methods encounter difficulties in multispectral settings due to non-linear intensity variations across spectral pairs. To address this, various hand-crafted [12, 136, 137] and learning-based [1, 138, 80] multispectral image matching frameworks have emerged. While these methods show promise on specific benchmarks, they typically target single-modality scenarios and often struggle to generalize across diverse multispectral tasks. For instance, models like RoMa [40] that utilize foundational models to enhance generalization still experience issues with overfitting and remain computationally expensive.

This work introduces XPoint, a modular architecture that enables adaptive training and fine-tuning on new datasets through self-supervision, achieving high performance in multispectral image matching where labeled data is limited. XPoint's modularity allows rapid adaptability to novel tasks by using aligned image pairs, providing a scalable, self-supervised solution for diverse datasets. Existing methods such as detector-free frameworks [37, 106, 40, 97] often rely on labeled data (e.g., camera poses, depth maps) unavailable in multispectral contexts, limiting their generalization. XPoint overcomes this by adopting SuperPoint's self-supervised approach [2] to generate pseudo-ground truth keypoints from aligned pairs, bypassing costly labeling and enhancing cross-modal generalization. Aligned image pairs are considerably easier to obtain than depth maps or precise camera poses, particularly in multispectral settings where additional annotations are scarce or difficult to collect. This self-supervised strategy reduces data collection requirements and enhances generalizability across various modalities. In contrast to similar work, such as ReDFeat [80], a self-supervised geometry-informed network that struggles with large viewpoint changes, XPoint effectively manages both viewpoint and spectrum

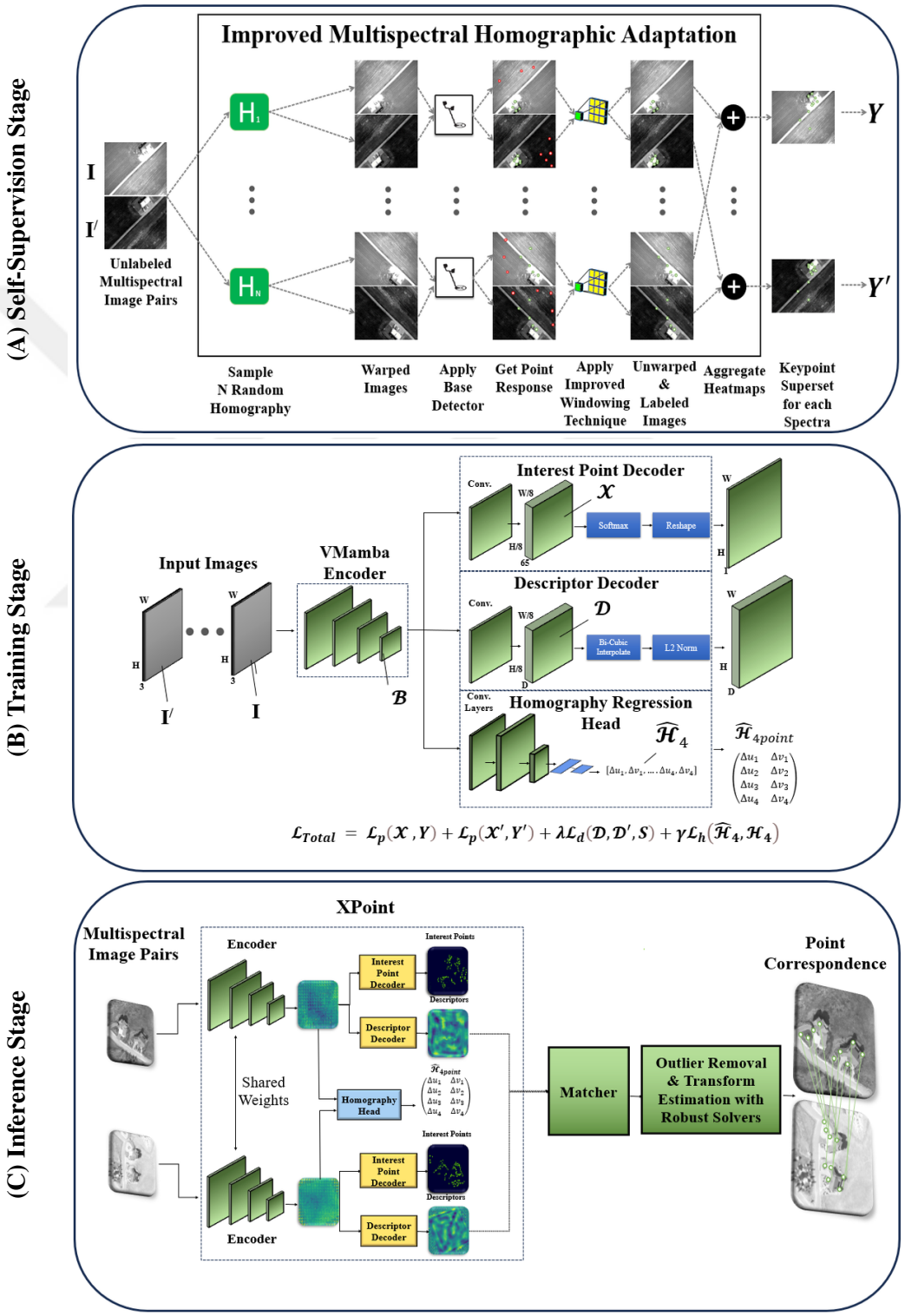
variations.

To handle viewpoint variation and enable self-supervision, SuperPoint [2] uses Homographic Adaptation to create pseudo-ground truth keypoints by simulating different viewpoints of the same scene, allowing a fully convolutional network to recognize scenes from multiple angles. Building on this, MultiPoint [1] adapts Homographic Adaptation for multispectral alignment, combining keypoints from both spectral domains into a unified superset. A key challenge is forming a common feature set across multispectral pairs, which MultiPoint addresses by merging keypoint maps through pixel-wise multiplication and Gaussian smoothing, though some keypoints may be missed. As discussed in the previous section, the windowing technique [18] refines this approach by allowing cross-spectral keypoints within a defined window around corresponding points in other spectra, allowing small localization errors and enhancing keypoint superset generation in MultiPoint.

This work adopts a self-supervised approach, enhancing pseudo-ground truth keypoint generation with an improved multispectral homographic adaptation stage and introducing a modular architecture with a pre-trained encoder and three joint heads: two dedicated to keypoint detection and description, and a task-specific head—homography regression. The main contributions are summarized as follows (see Fig. 3.4 for the complete framework):

Improved Multispectral Homographic Adaptation: We enhance our previously proposed windowing technique in two key ways: (1) by using the RIFT2 keypoint extractor [118] to generate modality-robust response maps; and (2) by accumulating accepted keypoint probabilities across multiple homography-warped views rather than applying binary acceptance. This probabilistic aggregation improves viewpoint consistency and yields pseudo-ground truth keypoints that are robust to both spectral misalignment and geometric variations. In contrast to the earlier binary formulation (see Algorithm 3), this

Figure 3.4: *Overview of the XPoint pipeline.* ***(A)*** *Self-supervision stage applies improved multispectral homographic adaptation using RIFT2 and probabilistic windowing to generate robust pseudo-ground truth keypoints.* ***(B)*** *Training with joint optimization of keypoint detection, descriptors, and homography.* ***(C)*** *Inference stage with matching and robust transformation estimation.*



variant treats accepted keypoints as soft probabilities that accumulate over warped views. Final keypoint masks are generated by thresholding the aggregated responses. This formulation adds temporal consistency to our spatially tolerant windowing mechanism, improving the reliability of pseudo-labels under viewpoint changes.

Pretrained VMamba Encoders: We incorporate 2D-Selective-Scan (SS2D) mechanisms in VMamba encoders, which provide enhanced semantic awareness by extracting features more effectively than CNNs and more efficiently than visual transformers.

Homography Regression Head: We include a dedicated task-specific head in the decoder to refine keypoint detection and description by guiding homography estimation. This constraint improves matching accuracy, enhances metric performance, and enables flexible model adaptation.

Revised Detector Loss: To address imbalances in challenging datasets such as VIS-SAR and VIS-NIR [80], we employ a weighted cross-entropy loss, boosting confidence and performance under significant spectral differences.

Modular XPoint Framework: Our adaptable XPoint framework achieves state-of-the-art results across four multispectral settings (VIS-TH ,VIS-NIR, VIS-LWIR, VIS-SAR) on five datasets [1, 139, 80]. Using a self-supervised approach with aligned image pairs, we reduce data requirements, enhancing generalization across spectrum shifts and viewpoint changes.

We next detail the overall architecture of XPoint, highlighting its three main stages—multispectral homographic adaptation for pseudo ground truth keypoint generation, training with modular heads, and inference for flexible matching—and explain how each component contributes to robust, cross-spectral correspondence estimation.

3.3.2 XPoint Architecture

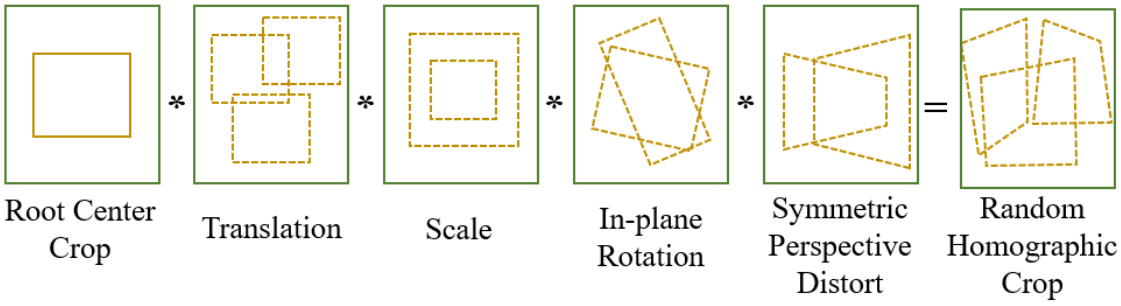
We introduce XPoint, a self-supervised framework developed to address multispectral image-matching challenges through a modular architecture. The framework operates in three stages: multispectral homographic adaptation, training, and inference (see Fig. 3.4). First, pseudo-ground truth keypoints are generated using the proposed multispectral homographic adaptation (see Fig. 3.4(A)). This process applies spectrum-aware keypoint acceptance criteria and utilizes a base detector like RIFT2 [118] to create pseudo-ground truth keypoint supersets that are invariant to both viewpoint and spectrum variations across different modalities.

In the training stage (see Fig. 3.4(B)), the network processes multispectral image pairs to generate keypoint heatmaps, descriptors, and homography parameters. The core VMamba encoder [134], pre-trained on segmentation tasks, enhances the network's capability to align multispectral features in a shared space, improving modality alignment. Compared to traditional CNNs or visual transformers like SwinV2 [140], this architecture optimally balances efficiency and accuracy. Additionally, a lightweight homography regression head guides keypoint detection and description, ensuring geometric consistency and boosting matching accuracy. This modular head is adaptable, supporting flexible configuration. In datasets like VIS-SAR, with significant nonlinear radiation distortions (NRD), a weighted cross-entropy loss manages false positives, enabling the model to prioritize learning keypoints that strengthen feature matching even under challenging conditions.

In the inference stage (see Fig. 3.4(C)), the detected keypoints can be matched using traditional methods like Mutual Nearest Neighbor with RANSAC or MAGSAC, or deep learning-based matchers such as LightGlue [53]. This modular design offers flexibility, enabling users to adjust or replace components to suit specific tasks or datasets. Overall, XPoint's three-stage, modular architecture effectively addresses the complexities

of multispectral image matching, providing adaptability for applications ranging from feature matching to homography estimation across diverse spectral modalities.

Figure 3.5: *Random homographies are generated by combining simple transformations such as translation, scaling, rotation, and symmetric perspective distortion, sampled within predefined ranges.*



3.3.2.1 Improved Windowing Technique in IMHA

Stage A of the XPoint pipeline (see Fig. 3.4(A)) employs an enhanced multispectral homographic adaptation strategy to generate pseudo-ground truth keypoints from aligned multispectral image pairs. This stage builds upon our previously introduced *windowing technique*, formalized in Algorithm 3, which relaxes the strict pixel-wise consistency constraint used in MultiPoint [1] by accepting keypoints that lie within a defined spatial window across spectra.

Rather than repeating the details of Algorithm 3, we focus here on the two novel enhancements introduced in XPoint:

- **Spectrally robust keypoint detector.** Instead of conventional detectors such as SURF, we use RIFT2 [118], a modality-aware keypoint extractor designed to handle extreme spectral differences. This improves the initial heatmap quality prior to applying the windowing criterion.
- **Probabilistic keypoint accumulation.** In contrast to the binary decision rule in our original formulation—where accepted keypoints are assigned a fixed value of

1—we now accumulate the confidence of each accepted keypoint across multiple homography-warped views. This soft aggregation is followed by thresholding to generate the final binary keypoint mask. The result is a label set that reflects both spectral alignment and temporal (viewpoint) consistency.

These two refinements allow the homographic adaptation stage to produce more semantically stable and viewpoint-invariant pseudo-labels. Compared to the original binary formulation, this strategy significantly enhances the reliability of self-supervised keypoint generation, especially under large viewpoint shifts or spectral misalignments.

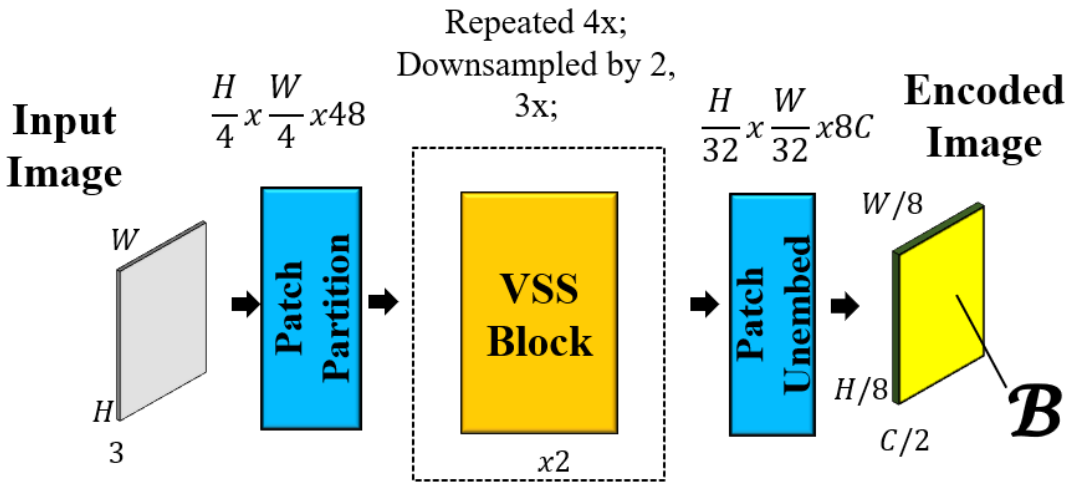
Comparison of Homographic Adaptation Variants. Table 3.1 summarizes the key differences across the four variants of homographic adaptation used throughout this thesis.

Table 3.1: *Comparison of Homographic Adaptation Variants*

Method	**Input Modalities**	**Keypoint Fusion**
SuperPoint [2]	Single-spectrum	Average of warped detections
MultiPoint [1]	Dual-spectrum	Pixel-wise multiplication
IMHA (ours)	Dual-spectrum	Window-based filtering
XPoint (enhanced)	Dual-spectrum	Probabilistic windowing

We adopt the strategy outlined by the authors of SuperPoint [2] for selecting appropriate homographies in Homographic Adaptation. Since not all 3x3 matrices represent plausible camera transformations, we decompose the homography into interpretable transformation classes, including translation, scaling, in-plane rotation, and symmetric perspective distortion. Each transformation is sampled within predefined ranges using a truncated normal distribution. These transformations are then combined with an initial root center crop to reduce bordering artifacts, as illustrated in Figure 3.5.

Figure 3.6: *Overview of the adopted VMamba encoder. The input image of size $H \times W \times 3$ is partitioned into patches and embedded to a feature map of size $\frac{H}{4} \times \frac{W}{4} \times 48$. This representation passes through four stages of VSS blocks, each reducing the spatial resolution by a factor of 2. After patch unembedding, the final encoded output $\mathcal{B}$ has dimensions $\frac{H}{8} \times \frac{W}{8} \times \frac{C}{2}$ and serves as the compact feature representation for downstream tasks.*



3.3.2.2 VMamba as an Encoder

To enhance feature extraction for multispectral image matching, we integrate the VMamba encoder, a state-of-the-art network optimized for computational efficiency and high accuracy with multispectral data. Our modular architecture allows the encoder to be easily replaced or adapted as needed. For this work, we select the VMamba-T[s1l8] [134] variant, pretrained on ADE20K for semantic segmentation, enabling it to capture semantically meaningful features across modalities and reducing the domain gap between spectral bands. This choice balances accuracy and computational efficiency, offering a robust alternative to traditional CNNs and visual transformers. A key feature of VMamba is its 2D-Selective-Scan (SS2D) mechanism, which focuses computational resources on the most relevant image regions, optimizing feature extraction and reducing the overhead associated with high-resolution images. This makes VMamba particularly suited for

aligning and matching high-dimensional multispectral image pairs with enhanced precision and efficiency.

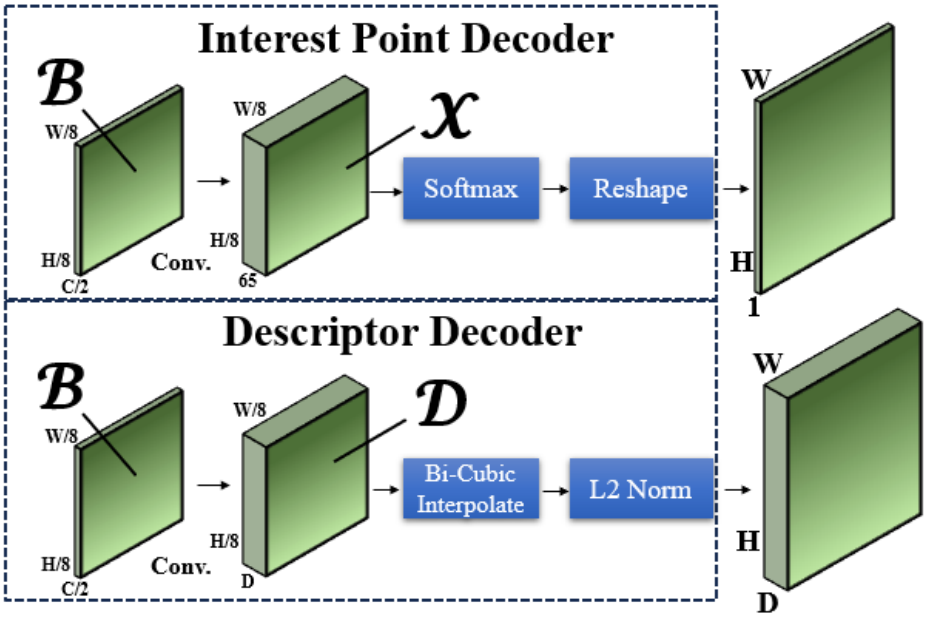
The VMamba encoder processes input images through a four-stage design that progressively refines feature representations across scales (see Fig. 3.6). The encoder receives an image of dimensions $H \times W \times 3$, representing RGB or multispectral inputs. First, the input image is divided into patches, creating an initial tensor of $\frac{H}{4} \times \frac{W}{4} \times 48$, reducing spatial resolution but preserving essential local information. Successive stages, referred to as Stage 1 through 4, apply downsampling and Visual State Space (VSS) blocks to refine features at progressively smaller scales and higher channel depths, ultimately yielding a compact multi-scale feature map of size $\frac{H}{8} \times \frac{W}{8} \times \frac{C}{2}$, suitable for downstream tasks.

The modular design of our architecture permits encoder substitution to meet task-specific needs. While VMamba is used here for its blend of semantic feature extraction and efficiency, alternative backbones, like Swin Transformers or CNNs, can be integrated as required. VMamba's selection is driven by its semantic awareness, efficiency with SS2D, and balanced performance, offering a strong alternative between CNNs and visual transformers. The VMamba encoder output, denoted $\mathcal{B} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{C}{2}}$, provides a dense, multi-scale feature representation that captures spatial and semantic information, facilitating keypoint detection and matching tasks.

3.3.2.3 Interest Point and Descriptor Decoders

Interest Point Decoder. The interest point decoder takes the intermediate tensor $\mathcal{B} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{C}{2}}$ from the encoder and applies a convolutional layer to compute the tensor $\mathcal{X} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 65}$. The 65 channels in this tensor correspond to different non-overlapping 8×8 pixel blocks, with each of the first 64 channels representing potential interest point locations in one of the blocks, and the 65th channel, "dustbin" cell, indicating blocks where no keypoints are present. After computing $\mathcal{X}$, a softmax function is applied along

Figure 3.7: *Interest Point and Descriptor Decoders. The interest point decoder outputs keypoint heatmaps, while the descriptor decoder generates dense descriptors for the input image.*

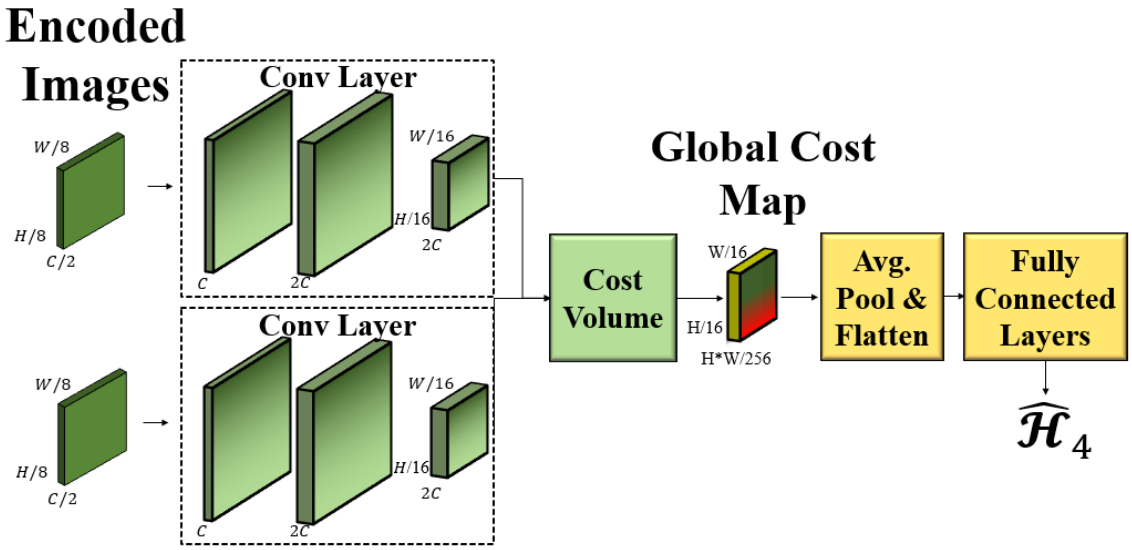


the channel dimension, transforming the raw values into probabilities. This process ensures that the network treats keypoint detection as a classification task, where each pixel block is classified into one of 65 classes. The first 64 channels represent the presence of keypoints in their corresponding locations within 8×8 pixel blocks, while the 65th channel indicates the absence of keypoints. The resulting tensor is reshaped to match the original image size $H \times W$ by selecting the most probable keypoint class/location per block and discarding the "dustbin" channel. The final output is a heatmap of keypoint locations across the entire image.

Descriptor Decoder. The descriptor decoder similarly takes the intermediate tensor $\mathcal{B}$ from the encoder and applies a convolutional layer to compute the tensor $\mathcal{D} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times D}$, where D is the fixed length of the descriptor. This tensor provides a semi-dense grid of descriptors, with each descriptor associated with a specific 8×8 pixel region in the original image. To produce dense descriptors, the output of the descriptor

decoder undergoes bicubic interpolation, which upscales the semi-dense grid to match the original image size $H \times W$. Following the interpolation, L2 normalization is applied to each descriptor, ensuring that all descriptors lie on the unit hypersphere in $\mathbb{R}^D$. This normalization step is essential for stabilizing the matching process, as it ensures that descriptor comparisons are based solely on the angular similarity between vectors. The use of semi-dense descriptors strikes a balance between memory efficiency and performance. Fully dense descriptor extraction can be computationally expensive and memory-intensive, particularly for high-resolution images. By focusing on one descriptor per 8×8 pixel block, the decoder reduces the computational burden while maintaining sufficient spatial resolution for effective matching.

Figure 3.8: *Homography Head Architecture.*



3.3.2.4 Homography Regression Head

In addition to the interest point and descriptor heads, we introduce a task-specific homography regression head to our design. Rather than directly estimating the final homography, this head imposes a homography constraint during training. This constraint enables multitask learning, allowing for the joint training of keypoint detection, descriptor generation, and homography constraint application. By sharing representations and

parameters across these tasks, the model learns more robust and informative features, improving performance in keypoint extraction and description while maintaining geometric consistency between images. Additionally, multitask learning [141] significantly reduces overfitting in hard parameter-sharing models like ours (see Fig. 3.4). The homography regression head builds on methodologies from DIHE [108] and SRHEN [142], following DIHE's supervision and training approach and incorporating an attention-based cost volume inspired by SRHEN. Further details on these aspects are elaborated in the training methodology section.

Our homography regression head (see Fig. 3.8) processes two input image streams through a neural network. The first stage, Layer1, applies convolutional layers with 3×3 kernels, batch normalization, and ReLU activation, followed by max pooling to produce feature maps. These feature maps are used to compute a cost volume via batch matrix multiplication, capturing feature similarities between the images. This cost volume aids in homography estimation and is condensed with adaptive average pooling before being passed through fully connected layers with dropout for regularization. The output is a flattened version of the 4-point parameterization matrix, $\mathcal{H}_4$, from which the homography matrix $\mathcal{H}$ is computed (see Section 3.3.3.1).

Given feature maps x_1 and x_2 from two input images, the cost volume CV is computed by taking the dot product of feature vectors along the depth dimension:

$$CV(i,j) = \sum_k \frac{x_1(i,k) \cdot x_2(j,k)}{\|x_1(i,k)\| \|x_2(j,k)\|}, \tag{3.2}$$

where i and j denote spatial positions within the feature maps, and k traverses the feature channels. This formula assesses feature similarity at various spatial locations, similar to attention scores, aiding in accurate homography estimation between the two images.

To summarize, this chapter introduced a two-stage methodology for robust multispectral image matching: an improved homographic adaptation technique leveraging a novel windowing mechanism for cross-spectral keypoint generation, and the proposed

XPoint architecture—a modular, self-supervised framework that generalizes across a wide range of spectral domains. The synergy between spectrum-aware keypoint supervision and a flexible encoder-decoder design enables accurate correspondence estimation under severe spectral and geometric distortions.

The next chapter presents an extensive experimental evaluation of these contributions, analyzing their effectiveness across multiple datasets, spectral modalities, and benchmark metrics to validate the robustness and generalizability of the proposed methods.

3.3.2.5 Inference Pipeline of XPoint

At inference time, the trained XPoint model processes a pair of multispectral images—such as a visible–thermal pair—to generate interest points and descriptors for both views. Each image is passed through a shared-weight encoder to produce feature maps, which are then decoded into a keypoint probability map (logits) and dense descriptors. The logits are converted into probability heatmaps using a softmax operation, followed by reshaping to recover the spatial resolution. Keypoints are extracted by applying confidence thresholding and Non-Maximum Suppression (NMS) to the probability maps. For each detected keypoint, a corresponding descriptor is gathered from the decoded descriptor map, which is first upsampled via bicubic interpolation and normalized using an L2 norm.

Keypoint correspondences are computed using the **mutual nearest neighbor** strategy: a descriptor in one image is paired with its closest neighbor in the other, and the match is retained only if this relationship is reciprocal. These candidate matches are then refined by **MAGSAC++** [43], which robustly identifies inliers and estimates a homography transformation between the input images. While the XPoint architecture includes an auxiliary **homography regression head** (denoted $\hat{\mathcal{H}}_4$) that predicts control

point offsets, this component is typically *omitted at inference time*. Empirically, the descriptor-based matching followed by robust estimation yields superior accuracy and generalization across spectral domains. The complete inference procedure is summarized in Algorithm 4.

Algorithm 4 XPoint Inference Pipeline

Require: multispectral image pair (I, I'), trained XPoint model, confidence threshold τ
Ensure: estimated homography $\mathcal{H}$

$\mathcal{B} \leftarrow \text{Encoder}(I), \quad \mathcal{B}' \leftarrow \text{Encoder}(I')$ {shared-weight feature extraction}
$\mathcal{X}_{logits}, \mathcal{D} \leftarrow \text{Decoders}(\mathcal{B}), \quad \mathcal{X}'_{logits}, \mathcal{D}' \leftarrow \text{Decoders}(\mathcal{B}')$
$\mathcal{X} \leftarrow \texttt{Softmax}(\mathcal{X}_{logits}), \quad \mathcal{X}' \leftarrow \texttt{Softmax}(\mathcal{X}'_{logits})$
$\mathcal{X} \leftarrow \texttt{Reshape}(\mathcal{X}), \quad \mathcal{X}' \leftarrow \texttt{Reshape}(\mathcal{X}')$
$K \leftarrow \texttt{NMS_and_Threshold}(\mathcal{X}, \tau), \quad K' \leftarrow \texttt{NMS_and_Threshold}(\mathcal{X}', \tau)$
$\mathcal{D}_{interp} \leftarrow \texttt{BiCubic}(\mathcal{D}), \quad \mathcal{D}'_{interp} \leftarrow \texttt{BiCubic}(\mathcal{D}')$
$\mathcal{D}_{norm} \leftarrow \texttt{L2Norm}(\mathcal{D}_{interp}), \quad \mathcal{D}'_{norm} \leftarrow \texttt{L2Norm}(\mathcal{D}'_{interp})$
$\mathcal{D}_K \leftarrow \texttt{GatherDescriptors}(K, \mathcal{D}_{norm})$,
$\mathcal{D}'_{K'} \leftarrow \texttt{GatherDescriptors}(K', \mathcal{D}'_{norm})$
$M \leftarrow \texttt{MutualNN}(\mathcal{D}_K, \mathcal{D}'_{K'})$ {mutual nearest neighbor matching}
$M^* \leftarrow \texttt{MAGSAC++}(M)$ {robust inlier selection}
$\mathcal{H} \leftarrow \texttt{EstimateHomography}(M^*)$
return $\mathcal{H}$

3.3.3 Model Training Details

3.3.3.1 Ground Truth Data Extraction

Keypoint selection : RIFT2 [118] is selected as the base detector due to its significant improvements over traditional methods like SIFT and SURF, particularly in handling nonlinear radiation distortions (NRD) common in multispectral image matching. Unlike SIFT[36], which is sensitive to these distortions, and SURF[3], which may struggle with rotation and scale variances, RIFT2 enhances the robustness of the original RIFT [39] by incorporating a more efficient rotation-invariance technique. Training datasets are labeled using RIFT2 combined with the proposed homographic adaptation technique. The window size for the proposed windowing technique is set to 5, and 100 random homographies are applied in the multispectral homographic adaptation. Random homographies

are generated by decomposing them into simpler transformations: translation ($\pm\%5$), scale ($\pm\%20$), rotation ($\pm 90°$), and distortion ($\pm\%20$ in width and height), resulting in complex transformations. Border reflection is used during training and testing to avoid artificial boundaries in warped images.

4-Point Homography Parameterization : In image alignment, homographies are typically represented by a 3x3 matrix, $\mathcal{H}$, transforming pixel coordinates from $[u, v]$ in one image to $[u', v']$ in another (see Equation 2.1). This standard approach, which includes rotational, translational, and perspective elements, can complicate the optimization process.

To simplify, an alternate parameterization focused on corner locations, known as the 4-point parameterization [108], is adopted. This approach reduces the representation to corner displacements $(\Delta u_1, \Delta v_1, \ldots, \Delta u_4, \Delta v_4)$ as $\mathcal{H}_{4point}$, expressed as:

$$\mathcal{H}_{4point} = \begin{pmatrix} \Delta u_1 & \Delta v_1 \\ \Delta u_2 & \Delta v_2 \\ \Delta u_3 & \Delta v_3 \\ \Delta u_4 & \Delta v_4 \end{pmatrix} \tag{3.3}$$

where $\Delta u_i = u_i' - u_i$ and $\Delta v_i = v_i' - v_i$ represent horizontal and vertical displacements for each corner point. This 4-point method simplifies the standard 3×3 homography matrix by focusing on corner displacements, making homography more intuitive for alignment. Conversion from $\mathcal{H}_{4point}$ to $\mathcal{H}_{matrix}$ can be achieved using the normalized Direct Linear Transform (DLT), which is a well-established algorithm for estimating projective transformations [143].

3.3.3.2 Self-Supervision and Training

Self-Supervision. During training, image pairs linked by a known homography and labeled with keypoint locations from RIFT2 [118] are utilized. Pre-aligned image

pairs from the same or different spectra are randomly selected, with one image warped by a random known homography, in order to provide balanced exposure to multispectral and same-spectrum pairs during training. The model is trained with PyTorch, using the Adam optimizer, for 3000 epochs on the Multispectral (optical-thermal) Image Pair Dataset, with a batch size of 32, automatic mixed precision, and a learning rate of 0.001. For other datasets, the base model is fine-tuned for 100 epochs. Training incorporates translation, rotation, scaling, and perspective changes for homography sampling to enhance model robustness. Additionally, 256×256 patches are randomly cropped from full-resolution images (512×640) as data augmentation.

Encoder and Decoder Training. The encoder, fine-tuned from a pretrained VMamba state, and the decoder, trained from scratch, are simultaneously trained to optimize the encoder's advanced features for multispectral matching.

Homography Regression Head Training. Ground truth corner displacements are derived by applying the ground truth homography transformation to the first image's corners, as described in Equation 2.1. Using the encoded image pairs as input, the homography regression head computes prediction of these eight displacement values. The homography head is optimized with Euclidean (L2) loss between ground truth and predicted displacements during training.

3.3.3.3 Loss Functions

The overall loss function, $\mathcal{L}$, is a weighted sum of three components: $\mathcal{L}_p$ for the interest point detector, $\mathcal{L}_d$ for the descriptor, and $\mathcal{L}_h$ for homography regression. Given pairs of multispectral images related by randomly generated homographies, we obtain pseudo-ground truth interest point locations and correspondences via homography $\mathcal{H}$. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_p(\mathcal{X}, Y) + \mathcal{L}_p(\mathcal{X}', Y') + \lambda \mathcal{L}_d(\mathcal{D}, \mathcal{D}', S) + \gamma \mathcal{L}_h(\hat{\mathcal{H}}_4, \mathcal{H}_4) \tag{3.4}$$

where $\mathcal{X}$ and $\mathcal{X}'$ represent the keypoint tensors in the original and transformed images, respectively, while $\mathcal{D}$ and $\mathcal{D}'$ are the descriptor tensors for these images (see Section 3.3.2.3). Y and Y' denote the ground truth interest point locations in the original and transformed images, respectively, and S denotes the entire set of correspondences for a pair of images. $\mathcal{H}_4$ is the ground truth 4-point homography parameterization, with $\hat{\mathcal{H}}_4$ as the predicted homography.

Interest Point Detector Loss : The loss function for the interest point detector, $\mathcal{L}_p$, is defined as a single-label, multi-class cross-entropy loss with adjustable weights to address class imbalance, especially in challenging datasets like VIS-SAR where keypoints may be sparse or absent in one spectrum. The 65th channel, known as the dustbin class, handles cases without keypoints, and its weight w_{65} is a hyperparameter optimized for challenging spectral conditions. This adjustable weight allows fine-tuning of the model's sensitivity to false positives, encouraging accurate keypoint predictions even in regions without ground truth markers.

The interest point detection loss $\mathcal{L}_p$ is computed over cells $\mathbf{x}_{hw}$ within the predicted tensors $\mathcal{X}$ and $\mathcal{X}'$, each sized $\frac{H}{8} \times \frac{W}{8} \times 65$, for two input images with ground truth tensors Y and Y':

$$\mathcal{L}_p(\mathcal{X}, Y) = \frac{1}{H_c W_c} \sum_{h=1}^{H_c} \sum_{w=1}^{W_c} l_p\left(\mathbf{x}_{hw}, y_{hw}\right) \tag{3.5}$$

where

$$l_p\left(\mathbf{x}_{hw}, y_{hw}\right) = -\sum_{c=1}^{65} w_c \, y_{hwc} \log\left(\frac{\exp(x_{hwc})}{\sum_{i=1}^{65} \exp(x_{hwi})}\right). \tag{3.6}$$

Here, $\mathbf{x}_{hw}$ represents predicted logits for cell (h, w), and y_{hw} provides the ground truth one-hot encoded locations. The weight w_c modulates the contribution of each class, particularly w_{65} for the dustbin class, typically set at 0.025 in the VIS-SAR setting.

Keypoint Cell Computation. The ground truth tensor Y is derived by dividing the input image into non-overlapping 8×8 pixel regions. Each block is represented by a one-hot encoded vector across 65 channels: the first 64 for keypoints, and the 65th for the

dustbin cell, indicating no keypoints. If multiple keypoints appear within a block, one is randomly selected.

Descriptor Loss : The descriptor loss compares pairs of descriptor cells, with $\mathbf{d}_{hw}$ in $\mathcal{D}$ from the first image and $\mathbf{d}'_{h'w'}$ in $\mathcal{D}'$ from the second spectrum. Correspondence between a cell at (h, w) in the first spectrum and its matching cell at (h', w') in the second spectrum is defined by the homography as follows:

$$s_{hwh'w'} = \begin{cases} 1, & \text{if } \left\| \widehat{\mathcal{H}\mathbf{p}_{hw}} - \mathbf{p}_{h'w'} \right\| \leq 4 \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

where $\mathbf{p}_{hw}$ is the central pixel in cell (h, w), and $\widehat{\mathcal{H}\mathbf{p}_{hw}}$ is the transformed position under homography $\mathcal{H}$. A weighting factor λ_d balances the descriptor loss, with hinge loss margins m_p and m_n applied to positive and negative matches, respectively:

$$\begin{aligned} &\mathcal{L}_d\left(\mathcal{D}, \mathcal{D}', S\right) = \\ &\frac{1}{\left(H_c W_c\right)^2} \sum_{\substack{h=1 \\ w=1}}^{H_c, W_c} \sum_{\substack{h'=1 \\ w'=1}}^{H_c, W_c} l_d\left(\mathbf{d}_{hw}, \mathbf{d}'_{h'w'}; s_{hwh'w'}\right), \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} l_d\left(\mathbf{d}, \mathbf{d}'; s\right) = \lambda_d * s * \max\left(0, m_p - \mathbf{d}^T \mathbf{d}'\right) \\ +(1-s) * \max\left(0, \mathbf{d}^T \mathbf{d}' - m_n\right). \end{aligned} \tag{3.9}$$

Homography Regression Head Loss : The homography regression head loss, $\mathcal{L}_h$, uses Euclidean (L2) loss to compare the ground truth 4-point parametrization $\mathcal{H}_4$ of the homography matrix $\mathcal{H}$ and the predicted 4-point parametrization $\hat{\mathcal{H}_4}$. The loss is given by:

$$\mathcal{L}_h(\hat{\mathcal{H}_4}, \mathcal{H}_4) = \sum_{i=1}^{4} ||\mathcal{H}_4^i - \hat{\mathcal{H}_4^i}||_2^2 \tag{3.10}$$

where $\mathcal{H}_4^i$ and $\hat{\mathcal{H}_4^i}$ are the i^{th} corner points in the ground truth and predicted 4-point parametrization, respectively.

4. EXPERIMENTAL EVALUATION

4.1 Overview

This chapter provides a comprehensive evaluation of the proposed contributions: the improved homographic adaptation strategy with windowing for visible–thermal keypoint generation, and XPoint, a modular self-supervised framework designed for general multispectral image matching. The evaluation is structured around the following key research questions:

- **Effectiveness of Windowing in IMHA:** To what extent does the proposed windowing strategy in IMHA enhance cross-spectral keypoint detection and matching compared to the baseline MultiPoint approach?
- **Adaptability of XPoint:** Can XPoint achieve robust matching and registration across diverse spectral domains, including VIS-TH, VIS-NIR, VIS-LWIR, and VIS-SAR?
- **Comparison to State-of-the-Art Methods:** How does XPoint perform relative to recent multispectral and general-purpose image matching methods, spanning both detector-based and detector-free paradigms?

Our experiments are conducted on several publicly available multispectral datasets that encompass a broad range of sensor types, spectral combinations, and scene contexts. To assess robustness, we follow the evaluation protocol introduced by Mikolajczyk et al. [144], where synthetic homographies are applied to simulate geometric transformations such as translation, scale, rotation, and perspective distortion.

We report three primary metrics: repeatability, matching score, and homography correctness. Additionally, we include Nearest Neighbor mean Average Precision (NN mAP) where applicable, specifically for evaluating the discriminative power of

learned descriptors in detector-based settings. Formal definitions and computation details of each metric are provided in Section 4.3.

To further analyze the effectiveness of key architectural components, we perform targeted ablation studies for both IMHA and XPoint. For IMHA, the study focuses on evaluating the impact of different window sizes used in the keypoint acceptance stage during homographic adaptation. For XPoint, the ablation examines several architectural choices, including: (i) the selection of the base keypoint detector (e.g., SURF vs. RIFT2), (ii) the encoder architecture (e.g., CNN, Swin Transformer, or VMamba), and (iii) the inclusion of the homography regression head. These component-wise evaluations allow us to isolate and understand the contribution of each design decision to the overall matching performance.

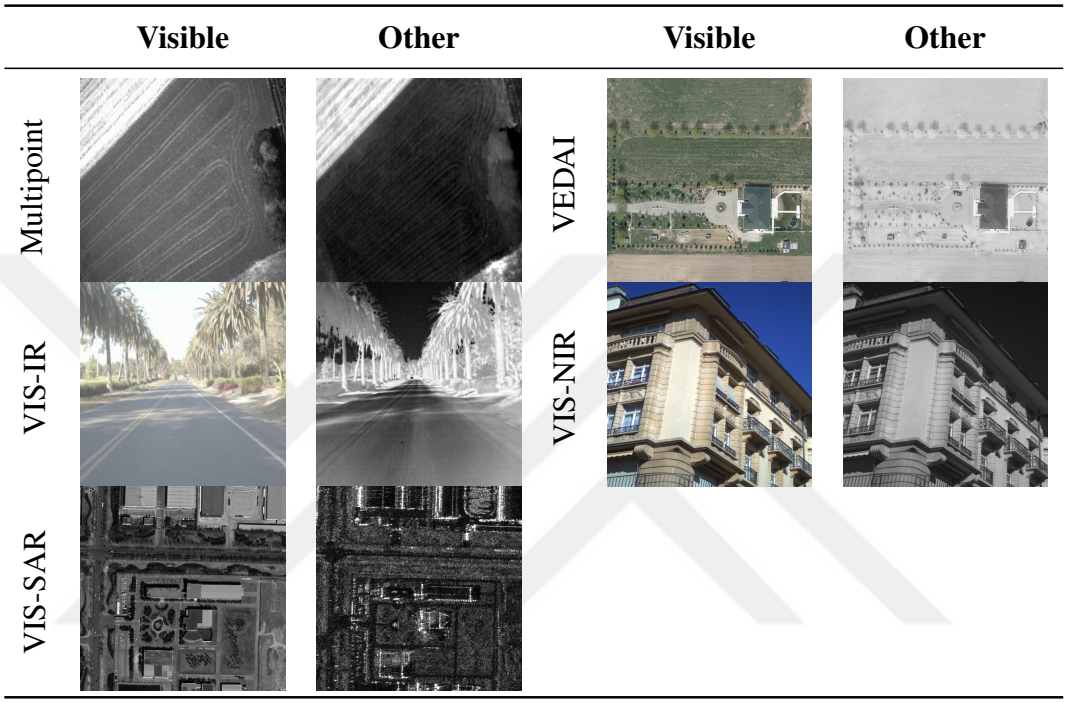
Finally, we provide qualitative results that illustrate the strengths and limitations of each method under various spectral and geometric challenges. These visual comparisons complement the quantitative analysis and offer insight into real-world performance in cross-spectral matching tasks.

4.2 Datasets

The experimental evaluation utilizes several multimodal datasets that span diverse spectral combinations and geographic contexts. Table 4.1 summarizes the datasets employed in this work, including their modalities, number of image pairs, and application domains.

The primary dataset is the **Multispectral (Optical–Thermal) Image Pair Dataset** from MultiPoint [1], which contains 9340 training and 4391 testing image pairs captured from fixed-wing UAV flights over agricultural areas. Each sample consists of aligned optical and thermal images (640×512 resolution), designed for robust keypoint detection and cross-spectral registration. This dataset serves as the base for pretraining.

Figure 4.1: *Sample image pairs from the five multispectral datasets used in this study. Each row presents an example image pair from a specific dataset, showcasing the diversity of spectral combinations and scene types: agricultural aerial imagery in MultiPoint (VIS-TH), urban aerial scenes in VEDAI (VIS-NIR), driving scenarios in VIS-IR (VIS-LWIR), architectural scenes in VIS-NIR (VIS-NIR), and satellite imagery in VIS-SAR (VIS-SAR). Visible spectrum images are shown in the left columns, while their corresponding multispectral counterparts are displayed on the right.*



The VEDAI dataset [139] includes 1045 training and 200 testing image pairs in RGB and near-infrared (NIR) modalities. It focuses on small vehicle detection in urban and suburban environments in Utah, USA, and supports fine-tuning for real-world object detection under spectral and structural variations.

Additionally, we incorporate datasets from the ReDFeat benchmark [80], which consist of three challenging image matching subsets:

- **VIS-NIR**: 345 training and 128 testing VIS-NIR image pairs with a resolution of 983×686, captured from diverse scenes such as forests, urban streets, old buildings, and indoor environments.

- **VIS-IR**: Comprising 221 training and 47 testing VIS-LWIR image pairs (average size: 533×321), this dataset includes static daytime building scenes and dynamic nighttime road scenes.
- **VIS-SAR**: A large-scale set of 2011 training and 424 testing image pairs of VIS-SAR imagery with 512×512 resolution. These satellite-based images capture rural and town scenes, providing a benchmark for evaluating geometry- and illumination-invariant registration.

To ensure consistency across modalities and datasets, all input images are converted to grayscale (if applicable), resized to 256×256, and normalized to the range $[0, 1]$ before being used for training or evaluation.

Together, these datasets enable a comprehensive evaluation of the model's generalization capability across aerial, satellite, and ground-based multimodal settings. Figure 4.1 illustrates example image pairs from each dataset for qualitative comparison.

Table 4.1: *Summary of the multimodal datasets used for training and evaluation.*

Dataset	Modalities	Train Pairs	Test Pairs	Domain
MultiPoint [1]	Visible–Thermal	9340	4391	Agriculture (Aerial)
VEDAI [139]	Visible–NIR	1045	200	Urban (Aerial)
VIS-NIR [80]	Visible–NIR	345	128	Mixed (Ground)
VIS-IR [80]	Visible–LWIR	221	47	Road Scenes (Ground)
VIS-SAR [80]	Visible–SAR	2011	424	Satellite Remote Sensing

4.3 Evaluation Protocol

In the evaluation phase, both the MultiPoint-based model enhanced with the proposed windowing strategy (see Section 3.2) and the XPoint architecture (see Section 3.3) are evaluated using the protocol by Mikolajczyk et al. [144]. This evaluation involves computing keypoints and their descriptors. Random homographies are generated by applying geometric transformations sampled from the following ranges: translation ($\pm 5\%$),

scale ($\pm 10\%$), rotation ($\pm 90°$), and symmetric perspective distortion ($\pm 5\%$) along both width and height. These transformations create warped test sets for performance assessment.

For detector-based methods [118, 1, 80], keypoints and descriptors are extracted and matched using a bidirectional nearest neighbor strategy. Detector-free methods [106, 103, 37, 40, 97], by contrast, directly estimate pixel-level correspondences without detecting keypoints or computing descriptors. Metrics such as Repeatability, Matching Score, and Homography Correctness are used consistently for evaluating both the MultiPoint-based model with windowing (see Section 3.2) and the XPoint architecture (see Section 3.3). MAGSAC [43] is applied in all cases to remove outliers and refine correspondences, using a reprojection threshold of 2 for homography estimation. While NN mAP is used to evaluate descriptor distinctiveness for the detector-based model, it is not applied in the evaluation of XPoint. This is because XPoint is evaluated alongside detector-free methods, which do not produce explicit keypoints or descriptors but instead predict dense pixel-level correspondences. In such cases, descriptor-level metrics like NN mAP are not meaningful or applicable.

Repeatability measures the consistency of keypoint detection across transformations. A keypoint is considered repeatable if a corresponding keypoint in the other image lies within a fixed threshold under homographic mapping. Repeatability is defined as:

$$\operatorname{Corr}(\mathbf{x}_i) = \left(\min_{j \in \{1,\ldots,N_2\}} \|\mathbf{x}_i - \hat{\mathbf{x}}_j\| \right) \leq \varepsilon \tag{4.1}$$

$$\operatorname{Rep} = \frac{1}{N_1 + N_2} \left(\sum_i \operatorname{Corr}(\mathbf{x}_i) + \sum_j \operatorname{Corr}(\mathbf{x}_j) \right) \tag{4.2}$$

where N_1 and N_2 are the number of detected keypoints in the first and second images, respectively, and $\varepsilon = 5$ pixels.

Nearest-Neighbor Mean Average Precision (NN mAP) evaluates descriptor discriminativeness by computing the mean Average Precision (mAP) across optical and

thermal modalities. Precision and recall are computed over multiple descriptor distance thresholds, and the final score is the average of the AUCs for both modalities:

$$\text{NNmAP} = \frac{1}{2}(\text{AP}_{\text{opt}} + \text{AP}_{\text{th}}) \tag{4.3}$$

This metric is applicable only to methods that produce explicit descriptors, making it relevant for detector-based models. Therefore, **NN mAP is not reported for XPoint**, as it is evaluated alongside detector-free methods, which do not extract discrete keypoints or descriptors but instead directly predict dense pixel-level correspondences. In such cases, NN-based descriptor evaluation is not meaningful.

Matching Score quantifies the percentage of correct correspondences among all matches and is defined as:

$$\text{Matching Score} = \frac{\text{True Positives}}{\text{Total Matches}} \tag{4.4}$$

This metric is computed symmetrically over overlapping regions, using a 5-pixel threshold to determine correctness.

Homography Correctness evaluates the quality of the estimated homography by comparing the average corner alignment error between ground truth and predicted transformations. It is defined as:

$$\text{HC} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\frac{1}{4}\sum_{j=1}^{4} d_{ij} \leq \varepsilon\right), \quad d_{ij} = \left\|c'_{ij} - \hat{c}'_{ij}\right\|_2, \tag{4.5}$$

where d_{ij} is the Euclidean distance between the ground truth and predicted corner locations. A threshold $\varepsilon \in \{1, \ldots, 10\}$ pixels is used to evaluate the accuracy of the estimated transformation.

4.4 Experiments of IMHA

This section presents the experimental validation of the proposed Improved Multispectral Homographic Adaptation (IMHA) strategy. We first introduce the training configuration and loss formulation, followed by evaluations on the MultiPoint and VEDAI

datasets. Finally, we investigate the effect of the window size in the keypoint acceptance stage.

4.4.1 Self-Supervision and Training

In this study, we do not modify the model architecture introduced in MultiPoint [1]. Instead, we propose an enhanced strategy for generating pseudo ground truth labels through Improved Multispectral Homographic Adaptation (IMHA). To evaluate its impact, we retrain the original MultiPoint model using pseudo-labels derived from our IMHA method, which replaces the original multiplicative fusion rule with a spatially tolerant windowing-based acceptance criterion. Training is conducted in a self-supervised manner using the MultiPoint dataset, which contains 9340 optical–thermal image pairs for training and 4391 pairs for testing. At each iteration, an image pair is randomly selected, and one image is synthetically warped using a random homography composed of translation, rotation, scaling, and perspective distortion. The model is then trained to produce consistent keypoint and descriptor outputs between the original and warped views.

The pseudo ground truth keypoints for supervision are generated using the IMHA strategy with a window size of 5×5, as discussed in Section 3.2. Random crops of size 256×256 are applied as data augmentation. The training runs for 1200 epochs using the Adam optimizer with a batch size of 16 and a learning rate of 0.001. All other components and hyperparameters remain unchanged from the original MultiPoint implementation.

4.4.2 Loss Functions

The loss formulation in IMHA follows the original MultiPoint design [1], focusing on keypoint repeatability and descriptor consistency. The total loss comprises two components: an interest point detection loss $\mathcal{L}_p$ and a descriptor loss $\mathcal{L}_d$, and is defined as:

$$\mathcal{L} = \mathcal{L}_p(\mathcal{X}, Y) + \mathcal{L}_p(\mathcal{X}', Y') + \lambda \mathcal{L}_d(\mathcal{D}, \mathcal{D}', S) \tag{4.6}$$

The formulation of $\mathcal{L}_p$ and $\mathcal{L}_d$ mirrors that of the XPoint framework (see Section 3.3.3.3), incorporating 65-way classification with a dustbin channel and a contrastive descriptor loss based on hinge margins. All remaining training parameters and settings follow the configuration in MultiPoint [1].

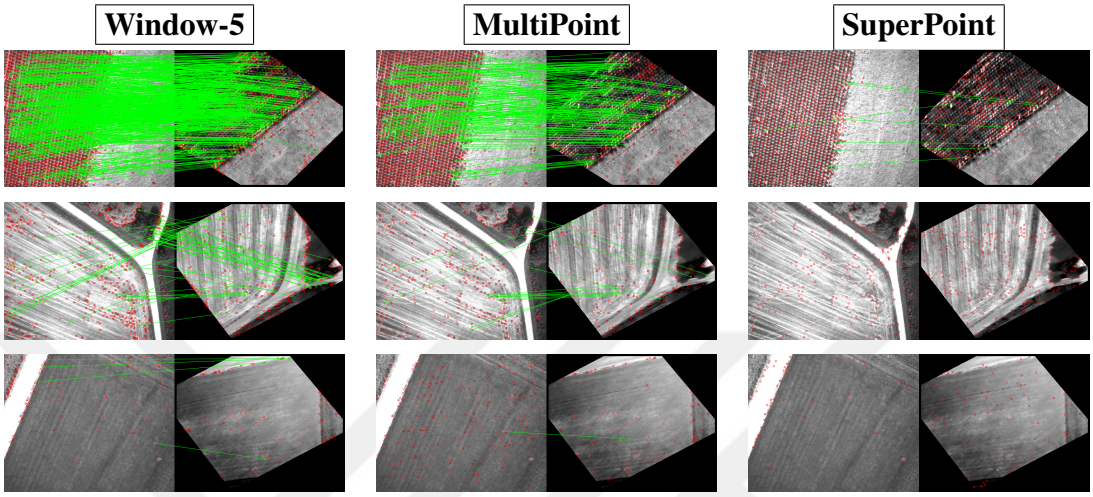
4.4.3 Evaluation on MultiPoint's Optical–Thermal Dataset

We begin by evaluating IMHA on the MultiPoint test set. Table 4.2 reports quantitative results, while Figure 4.2 provides qualitative comparisons. The proposed **Window-5** model outperforms all baseline methods across all key metrics. Notably, the Repeatability and Matching Score improve significantly over **MultiPoint**, indicating more stable and consistent keypoints. The NN mAP and number of keypoints (Nkp) also increase, showing that the windowing strategy yields more discriminative descriptors and denser detections. These improvements translate to better performance in homography estimation accuracy at all pixel thresholds. Qualitative examples in Figure 4.2 further support these findings, where **Window-5** produces visually correct matches under modality shift, especially in textured regions.

Table 4.2: *Evaluation metrics on the MultiPoint test set with random homographies applied.*

Method	Repeat.	Nkp	NN mAP	M. Score	$\epsilon = 2$	$\epsilon = 5$	$\epsilon = 10$
SURF	0.222	550	0.006	0.022	0.007	0.043	0.022
SIFT	0.267	635	0.021	0.035	0.064	0.171	0.217
SuperPoint	0.163	255	0.071	0.032	0.015	0.089	0.148
MultiPoint	0.281	424	0.271	0.134	0.125	0.507	0.667
Window-5	**0.329**	**647**	**0.303**	**0.179**	**0.140**	**0.522**	**0.705**

Figure 4.2: *Qualitative results on the MultiPoint dataset. In each matched image pair, the left image is optical and the right is thermal. Green lines show correct matches within a 4-pixel threshold. Both* ***Window-5*** *and* ***MultiPoint*** *succeed when enough texture is present, while* ***SuperPoint*** *struggles under large cross-spectral transformations. All methods degrade under extreme modality differences (bottom row).*



4.4.4 Generalization Results on VEDAI Dataset

To assess generalizability, we evaluate the model on the VEDAI [139] dataset, which contains 1245 optical–thermal image pairs. We use 1045 pairs for training and 200 for testing. Table 4.3 shows that the **Window-5** model again outperforms the baseline MultiPoint across all metrics. While SIFT and SuperPoint perform competitively in some cases, our method achieves a significantly higher NN mAP and Matching Score, reflecting improved descriptor reliability and cross-domain robustness. These results underscore the adaptability of the windowing strategy to new domains beyond the original training distribution.

4.4.5 Ablation Study: Effect of Window Size

To determine the optimal spatial tolerance for the keypoint acceptance mechanism, we conduct an ablation study comparing three window sizes: 3, 5, and 7. As shown in Table 4.4, the 5×5 window achieves the best balance between repeatability and precision,

Table 4.3: *Evaluation metrics on the VEDAI test set with random homographies applied.*

Method	Repeat.	Nkp	NN mAP	M. Score	$\epsilon = 2$	$\epsilon = 5$	$\epsilon = 10$
SIFT	0.455	**471**	0.245	0.185	**0.775**	**0.874**	**0.915**
SuperPoint	**0.567**	217	0.365	0.256	0.392	0.548	0.608
MultiPoint	0.392	184	0.463	0.277	0.377	0.618	0.714
Window-5	0.548	361	**0.713**	**0.446**	0.462	0.789	**0.915**

outperforming other sizes across all key metrics including Matching Score, NN mAP, and Homography Correctness at pixel thresholds of $\epsilon = 2, 5, 10$. This result confirms that moderate spatial tolerance enhances cross-spectral keypoint robustness without sacrificing localization accuracy.

Table 4.4: *Comparison of the effects of different window sizes in evaluation metrics.*

	Repeatability	Nkp	NN mAP	M. Score	$\epsilon = 2$	$\epsilon = 5$	$\epsilon = 10$
Window-3	0.325	476	0.247	0.151	0.118	0.479	0.659
Window-5	**0.329**	**647**	**0.303**	**0.179**	**0.140**	**0.522**	**0.705**
Window-7	0.302	480	0.165	0.121	0.094	0.418	0.601

In conclusion, the proposed IMHA method significantly improves multispectral keypoint detection and matching, particularly in visible–thermal settings. The windowing-based homographic adaptation strategy not only increases repeatability and matching precision but also generalizes well to unseen datasets. The ablation study confirms the importance of carefully selecting the spatial acceptance window size, with 5×5 providing the most effective balance.

While IMHA is tailored for visible–thermal domains, broader spectral differences such as VIS-NIR, VIS-LWIR, and VIS-SAR present additional challenges. To address this, we now introduce and evaluate **XPoint** 3.3, a fully modular and self-supervised framework capable of robust multimodal registration beyond the visible–thermal spectrum.

4.5 Experiments of XPoint

While XPoint can estimate homography directly, simulations reveal that an indirect approach (Fig. 3.4(C)) using keypoints detected by our model yields more robust results. XPoint computes keypoints and descriptors, which can be matched using traditional methods like nearest neighbor search or deep learning-based approaches like LightGlue [53]. In this work, we use the traditional pipeline for its simplicity and compatibility across benchmarks in detector-based methods, including our method. Outliers are removed using MAGSAC [43] to estimate homography parameters. Benchmarks include detector-based methods, RIFT2[118], MultiPoint[1] and RedFeaT[80] and detector-free methods, ASpanFormer[106], DKM[103], LoFTR[37], RoMa[40] and XoFTR[97]. For our method, we test two variants, "XPoint_001" and "XPoint_015", which have keypoint detection threshold values of 0.001 and 0.015, respectively. The evaluation results for all datasets are presented in Figures 4.3 and 4.4.

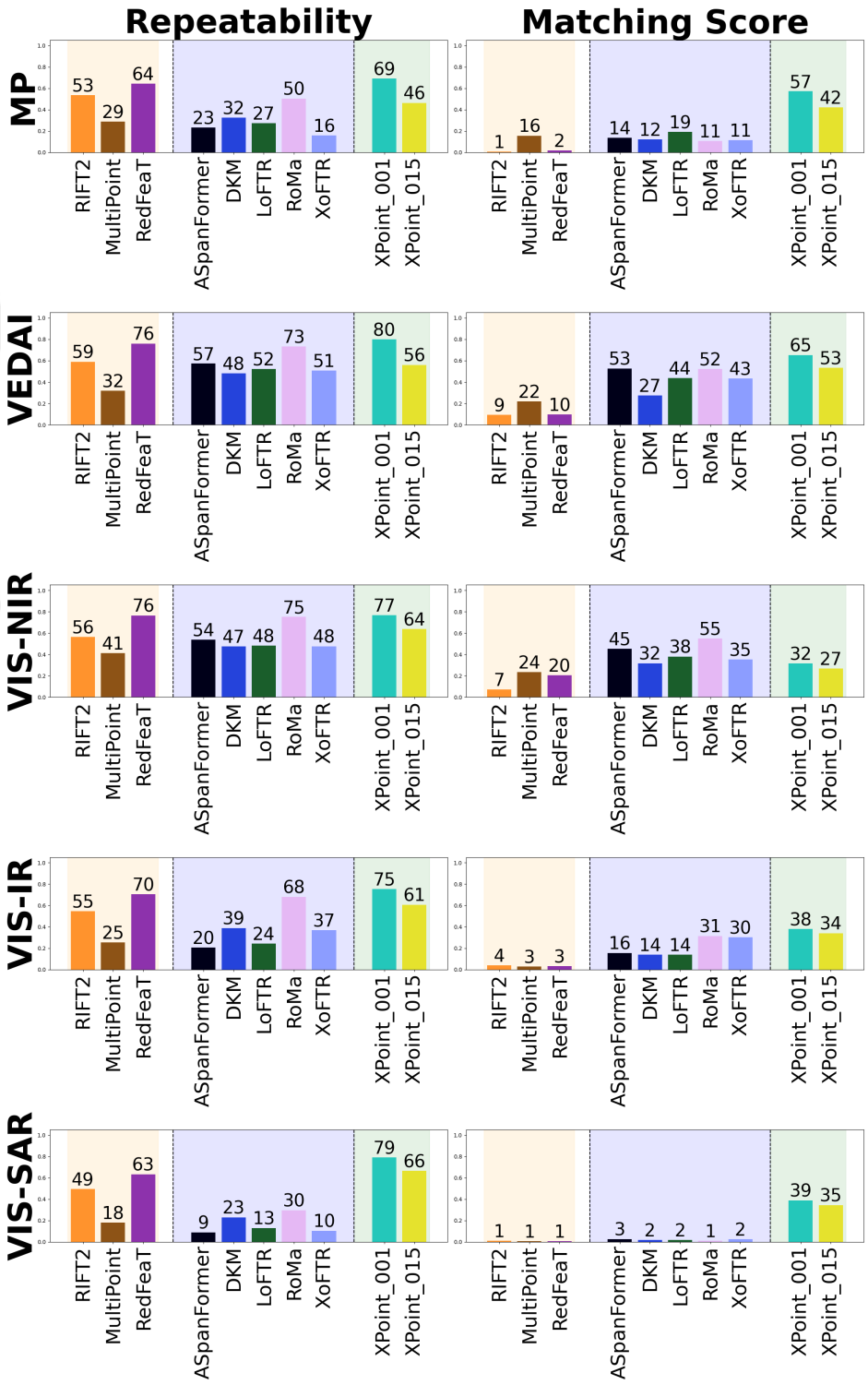
We begin by evaluating the effectiveness of XPoint in detecting repeatable keypoints and producing robust correspondences. This section presents comparisons with both detector-based and detector-free baselines across multiple modalities.

4.5.1 Feature Matching Performance

Figure 4.3 highlights the feature matching performance of detector-based methods, RIFT2[118], MultiPoint[1] and RedFeaT[80] and detector-free methods, ASpanFormer[106], DKM[103], LoFTR[37], RoMa[40] and XoFTR[97] and our proposed method's variants XPoint_001 and XPoint_015 on five different datasets. In this figure, Repeatability and Matching Score are selected as primary evaluation metrics and calculated at pixel threshold of 5.

Repeatability. Figure 4.3, row "Repeatability", presents the Repeatability scores across various datasets, with each dataset occupying a separate column. Our proposed

Figure 4.3: *Repeatability (left) and Matching Score (right) for detector-based (RIFT2, MultiPoint, ReDFeat), detector-free (ASpanFormer, DKM, LoFTR, RoMa, XoFTR) and proposed XPoint variants across five multispectral datasets: MP (VIS-TH), VEDAI (VIS-NIR), VIS-NIR (VIS-NIR), VIS-IR (VIS-LWIR) and VIS-SAR (VIS-SAR). XPoint_001 exhibits the highest repeatability and matching consistency on all datasets except VIS-NIR, where detector-free methods marginally outperform.*



XPoint_001 variant consistently outperforms established models such as RoMa [40] and RedFeaT [80] across all datasets. In contrast, the XPoint_015 variant demonstrates slightly lower performance. These findings highlight the robustness of XPoint in ensuring reliable keypoint detection, even across diverse spectral domains. Notably, XPoint excels in identifying candidate keypoints suitable for matching, showcasing its ability to effectively adapt to the nonlinear radiation distortion (NRD) effects inherent to multispectral imagery.

Matching Score. Figure 4.3, row "Matching Score", showcases the Matching Scores across multiple datasets. Our XPoint_001 variant consistently outperforms other methods with a significant margin across all datasets, except in the VIS-NIR setting. This superior performance underscores the robustness of our approach in identifying reliable matches, particularly in challenging multispectral scenarios such as Optical-Thermal and VIS-SAR. Both XPoint_001 and XPoint_015 variants demonstrate exceptional performance, surpassing other methods in most cases. This indicates that the repeatable keypoints detected by our method are also highly matchable, highlighting the strength of our descriptor representation.

In the VIS-NIR setting, however, we are surpassed by detector-free methods. This dataset includes a wide variety of scenes, such as countryside, fields, forests, indoor environments, mountains, historic buildings, streets, urban areas, and water bodies. Given the limited training data of only 345 image pairs, matching the performance of detector-free methods—trained on large datasets with extensive scene diversity similar to VIS-NIR—becomes challenging. Nonetheless, our approach still outperforms all detector-based methods, including RedFeaT, the previous state-of-the-art in this setting. This achievement highlights the competitive advantage of our method among detector-based approaches, even with limited training data.

In summary, XPoint_001 consistently outperforms competing methods in repeatability and matching score, particularly in challenging modalities such as VIS–SAR and Optical–Thermal. While detector-free methods show advantages on datasets similar to their training domains, XPoint demonstrates strong adaptability even with limited data.

4.5.2 Homography Correctness and Image Registration Performance

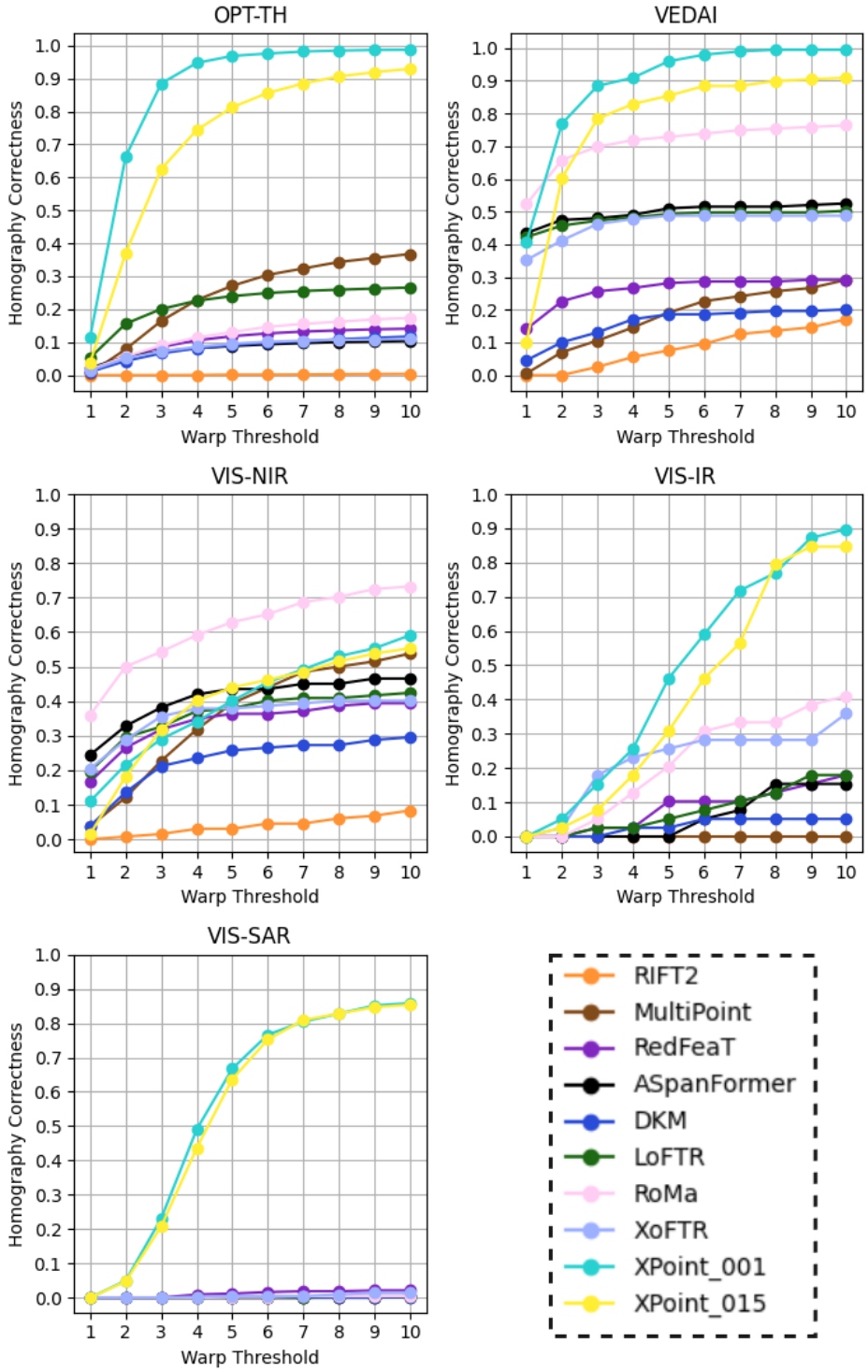
Figure 4.4 presents the Homography Correctness (HC) results across various datasets. The HC metric, as defined in Section 4.3, captures the proportion of image pairs for which the mean corner alignment error remains within a specified pixel threshold. This metric effectively quantifies the geometric accuracy of the estimated homography matrices.

Across the majority of datasets, the proposed XPoint_001 variant consistently outperforms other methods by a substantial margin, demonstrating robust geometric alignment capabilities. The performance further improves as the pixel error threshold is relaxed, reflecting the accuracy and reliability of the estimated transformations.

An exception is observed in the VIS-NIR setting, where XPoint_001 performs on par with detector-free methods like RoMa. This can be attributed to the extensive scene diversity in the VIS-NIR dataset, which includes countryside, fields, forests, buildings, and urban environments. Such diversity closely matches the training data of many detector-free methods, providing them with a domain-specific advantage.

In contrast, the XPoint_001 variant exhibits exceptional performance on more challenging datasets, such as OPT-TH and VIS-SAR, where it consistently surpasses all competing methods by a wide margin. This outcome highlights the robustness of XPoint_001 in accurately estimating homography matrices, even under significant spectral variation. The qualitative examples in Figure 4.8 and Figure 4.9 further illustrate this

Figure 4.4: *Homography correctness versus warp threshold (ϵ) for detector-based (RIFT2, MultiPoint, ReDFeat), detector-free (ASpanFormer, DKM, LoFTR, RoMa, XoFTR), and proposed XPoint variants across five benchmarks: OPT-TH (VIS-TH), VEDAI (VIS-NIR), VIS-NIR, VIS-IR (VIS-LWIR), and VIS-SAR.*



advantage, demonstrating precise image registration even in complex multimodal conditions where competing methods often struggle due to poor keypoint localization and matching accuracy.

These findings reinforce the effectiveness of the proposed approach in both feature matching and accurate homography estimation, making it particularly suitable for real-world applications requiring precise geometric consistency.

4.5.3 System Runtime and Keypoint Statistics

We evaluate runtime and keypoint statistics for all methods, including XPoint variants, on an NVIDIA RTX 2080Ti GPU. Figure 4.5 compares the average keypoints detected and runtime (seconds) across multiple datasets. XPoint_001 detect an average of 637 keypoints in 1.41 seconds per image pair, while XPoint_015 detect 356 keypoints in 0.49 seconds, reflecting a trade-off between keypoint density and processing speed. Compared to other methods, XPoint variants strike a favorable balance between runtime and detection. For instance, RoMa achieves high keypoint counts (788) but averages a longer runtime (1.07 seconds), while MultiPoint is faster (0.04 seconds) but detects fewer keypoints (73). This balance makes XPoint suitable for real-time applications, with configurable variants like XPoint_001 and XPoint_015 adapting to different precision or computational needs.

4.5.4 Ablation Study

4.5.4.1 Analysis of Architectural Modifications

To evaluate the impact of different components within the XPoint architecture, an ablation study was conducted on the Optical-Thermal dataset (see Table 4.5). The performance metrics considered include Repeatability, Matching Score, and Homography Correctness (HC) at pixel error thresholds of $\varepsilon = 3$ and $\varepsilon = 5$. In all experiments, the keypoint detection threshold was fixed at 0.015.

Figure 4.5: *Comparison of average runtime (seconds) and keypoint counts across datasets for various methods.*

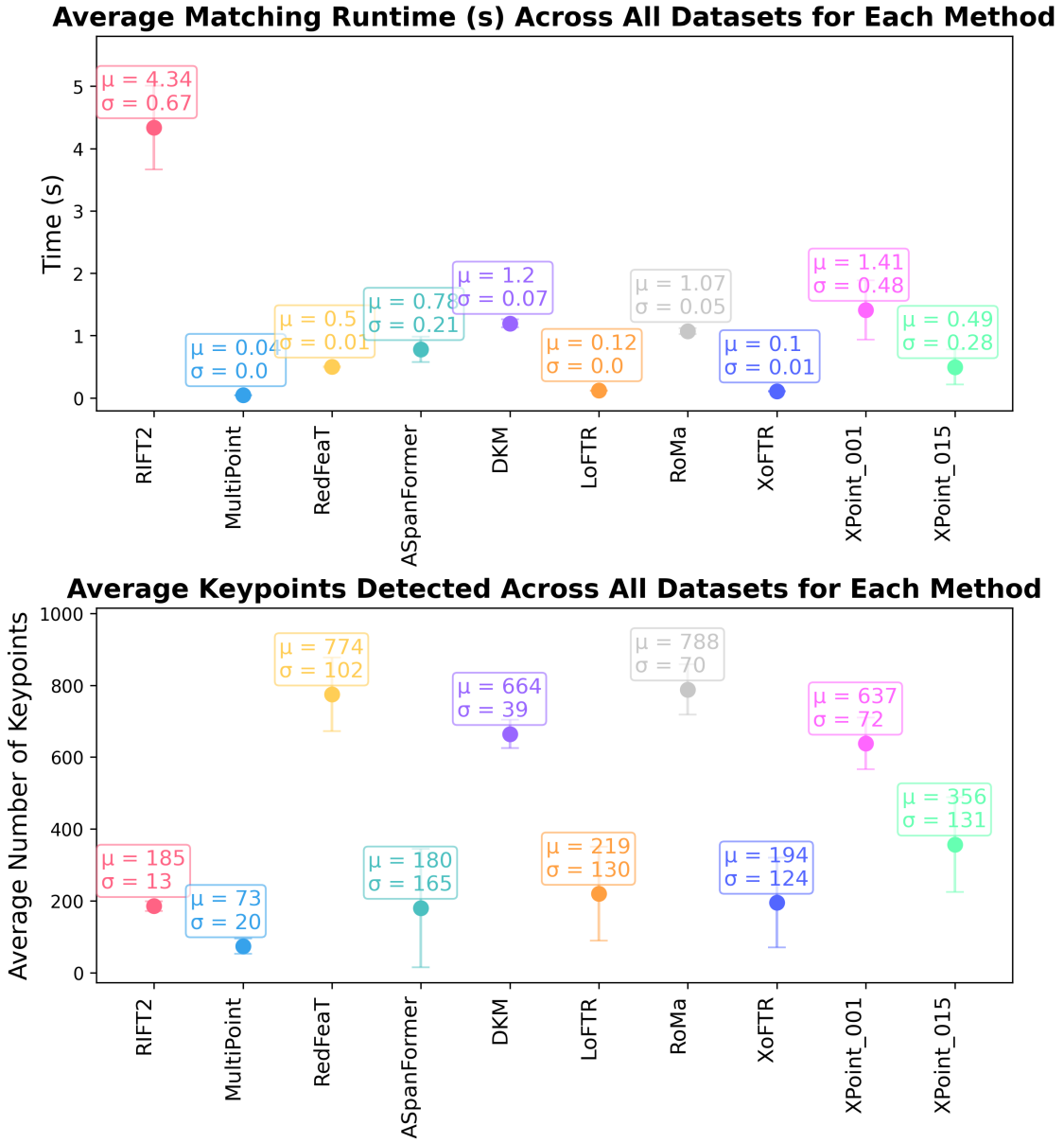


Table 4.5: *Ablation Study of XPoint Architectural Configurations on the Optical-Thermal Dataset.*

	Homographic Adaptation		Architecture	
Configuration	**Acceptance**	**Base Detector**	**Encoder**	**Homography Head**
RIFT2	-	-	-	-
Base	Gaussian	SURF	CNN	✗
XP1	Windowing [18]	SURF	CNN	✗
XP2	Proposed W.	SURF	CNN	✗
XP3	Proposed W.	RIFT2	CNN	✗
XP4	Gaussian	SURF	Swin-T	✗
XP5	Windowing	SURF	Swin-T	✗
XP6	Windowing	SURF	Swin-T	✓
XP7	Proposed W.	RIFT2	Swin-T	✓
XPoint	Proposed W.	RIFT2	VMamba-T	✓
	Performance Metrics			
Method	**Repeatibility**	**Matching Score**	$\varepsilon = 3$	$\varepsilon = 5$
RIFT2	0.534	0.006	0	0.001
Base	0.286	0.156	0.164	0.271
XP1	0.311	0.149	0.181	0.293
XP2	0.346	0.145	0.201	0.350
XP3	0.439	0.122	0.257	0.451
XP4	0.323	0.340	0.256	0.374
XP5	0.379	0.344	0.296	0.482
XP6	0.391	0.373	0.344	0.521
XP7	0.462	0.401	0.573	0.794
XPoint	**0.469**	**0.426**	**0.631**	**0.813**

4.5.4.2 RIFT2 vs. Base Model

The *RIFT2* configuration demonstrates a significant advantage in Repeatability, achieving a score of 0.534, compared to the *Base* model's 0.286. This highlights RIFT2's strength in detecting repeatable keypoints. However, this comes at the cost of descriptor distinctiveness, as reflected in its significantly lower Matching Score (0.006 vs. 0.156 for the Base model). This result indicates that while RIFT2 excels in keypoint detection, it struggles to accurately match descriptors across different modalities, primarily due to the limitations of handcrafted descriptors.

4.5.4.3 Impact of Keypoint Acceptance Techniques

Incorporating the original windowing method in *XP1* increased Repeatability from 0.286 to 0.311 and slightly improved HC. However, the small decline in Matching Score suggests that while more repeatable keypoints were detected, the distinctiveness of these points remained a challenge due to the limitations of the CNN encoder.

Implementing the proposed windowing technique in *XP2* further improved Repeatability to 0.346 and enhanced HC. The continued slight impact on Matching Score reflects the challenge of extracting distinctive descriptors as the number of detected keypoints increases.

4.5.4.4 Effectiveness of the Base Detector

Replacing the SURF detector with RIFT2 in the *XP3* configuration led to substantial gains, increasing Repeatability to 0.439 and significantly boosting HC. However, the Matching Score decreased to 0.122, indicating that higher keypoint density can challenge the CNN encoder's ability to extract distinctive features. This result reinforces the role of RIFT2 as a powerful base detector when combined with a learning-based approach, effectively balancing keypoint density and repeatability.

4.5.4.5 Impact of Encoder Architecture

Upgrading to a Swin Transformer encoder in *XP4* resulted in improvements across all metrics, with a notable increase in Matching Score to 0.340. This demonstrates the Swin-T encoder's superior ability to extract distinct and robust descriptors. Reintroducing windowing in *XP5* further boosted performance, increasing Repeatability to 0.379 and improving HC at both $\varepsilon = 3$ and $\varepsilon = 5$, highlighting the benefit of combining advanced feature extraction with effective keypoint acceptance strategies.

4.5.4.6 Contribution of the Homography Head

Adding a homography head in *XP6* provided a consistent boost across all metrics, raising Repeatability to 0.391 and Matching Score to 0.373. The improved HC scores confirm the homography head's role in enhancing geometric consistency and refining keypoints for more accurate alignment.

4.5.4.7 Synergistic Effects of Combined Components

Combining RIFT2 as the base detector, the proposed windowing method, the Swin-T encoder, and the homography head in *XP7* resulted in substantial gains across all metrics. The Repeatability reached 0.462, the Matching Score increased to 0.401, and HC scores improved to 0.573 and 0.794 at $\varepsilon = 3$ and $\varepsilon = 5$, respectively. This configuration effectively integrates multiple complementary components, resulting in superior keypoint learning and matching performance.

4.5.4.8 Superior Performance with VMamba-T Encoder

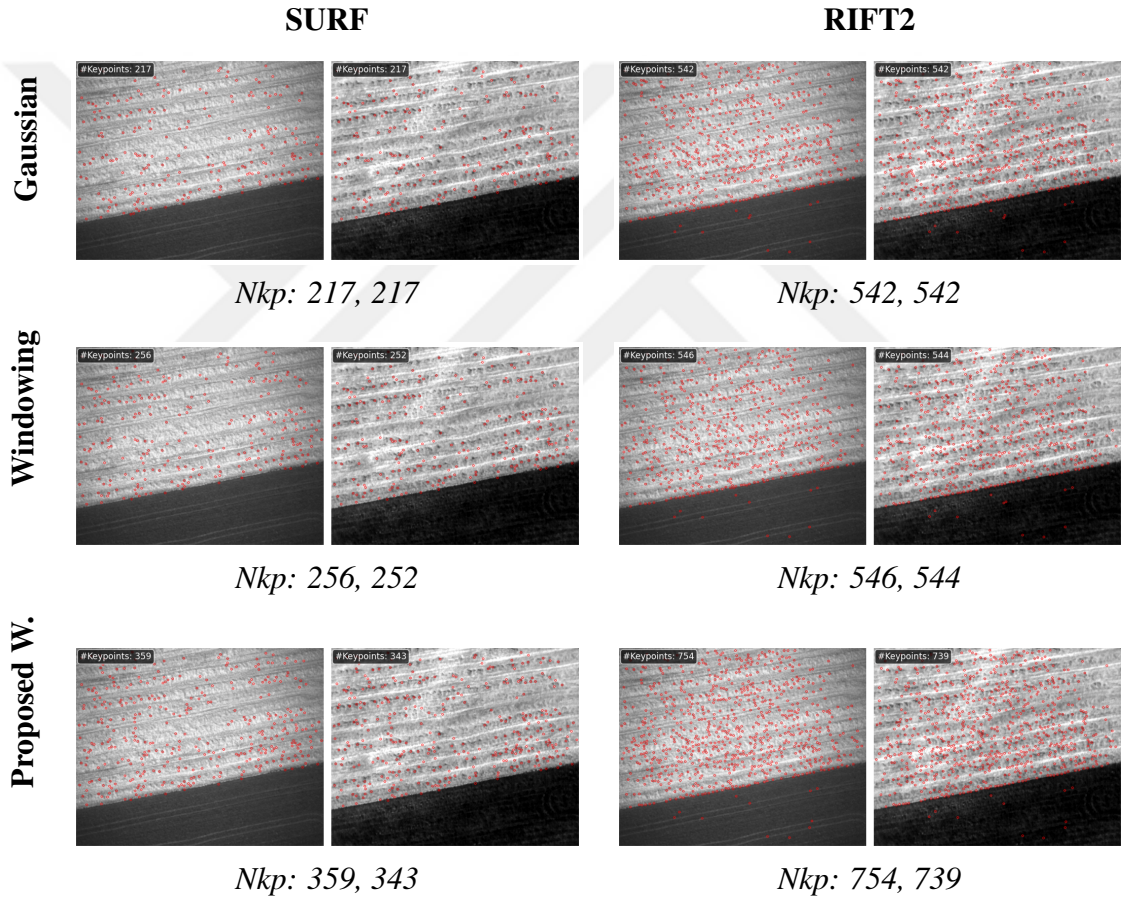
The final *XPoint* model incorporates the VMamba-T encoder, achieving the highest metrics across all evaluated configurations. Specifically, it reached a Repeatability of 0.469, a Matching Score of 0.426, and HC scores of 0.631 and 0.813. This underscores the VMamba-T encoder's superior capability in cross-modal matching through advanced feature extraction, confirming its effectiveness in challenging multimodal contexts.

4.5.4.9 Summary of Ablation Findings

This ablation study demonstrates that each architectural component contributes meaningfully to the performance of XPoint. The windowing-based keypoint acceptance strategy improves repeatability by enabling more consistent detection across challenging inputs. Using RIFT2 as a base detector enhances keypoint localization in cross-modal

settings. Upgrading the encoder to Swin-T or VMamba-T significantly boosts descriptor quality, while adding a homography regression head further refines alignment and improves geometric consistency. Combined, these improvements lead to better Repeatability, Matching Score, and Homography Correctness.

Figure 4.6: *Visual comparison of keypoint detection results for optical–thermal image pairs using different base detectors (SURF vs. RIFT2) and homographic adaptation strategies. Rows correspond to different aggregation techniques (Gaussian, weighted windowing, and proposed unweighted windowing), and columns indicate the base detector used. Nkp denotes the number of detected keypoints in each image.*



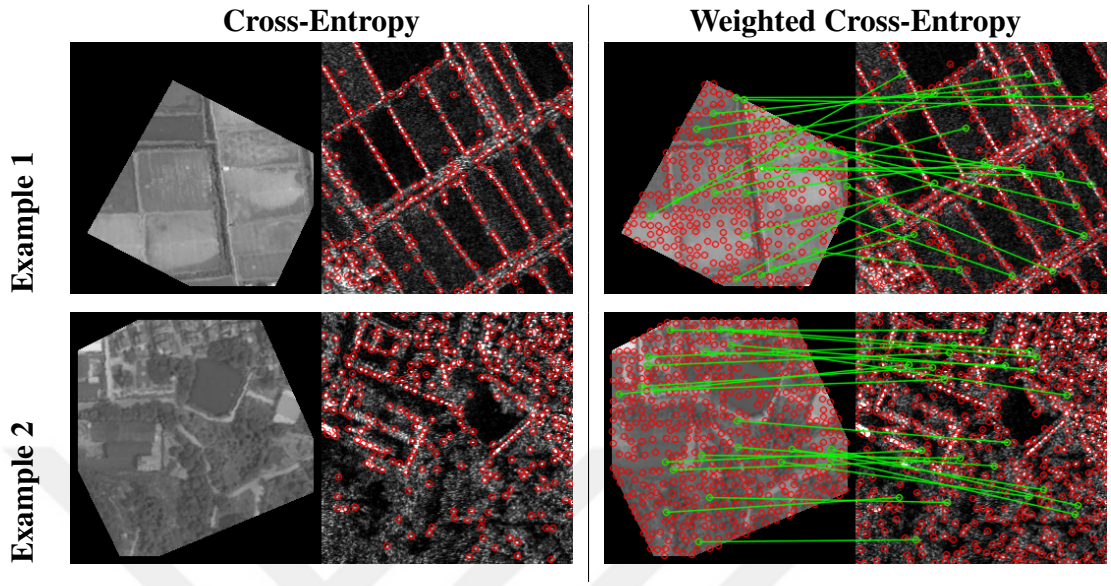
4.5.4.10 Effect of Base Detector and Proposed Windowing Technique in Homographic Adaptation Stage

This section evaluates the impact of different base detectors and keypoint acceptance techniques on keypoint detection in multispectral pairs, comparing SURF and RIFT2 detectors with three adaptation methods: Gaussian, windowing, and our proposed windowing technique (see Table 4.6). The Gaussian method generates a keypoint heatmap by performing element-wise multiplication on image pairs, followed by smoothing with a Gaussian filter. However, it lacks adaptability to spectral variations. The windowing technique [18] improves flexibility by matching keypoints within a defined window, but applies a binary acceptance criterion. Our proposed method further refines this by treating keypoints probabilistically and accumulating observations across homographies, resulting in more robust keypoint detection adaptable to both spectral and viewpoint changes. Among detectors, RIFT2 outperforms SURF across all adaptation methods, owing to its ability to address nonlinear radiation distortion (NRD) and reduce spectral domain gaps. Combining RIFT2 with our proposed technique produces a dense, reliable keypoint set, proving optimal for cross-spectral matching. In summary, this pairing enhances keypoint density and distribution across multispectral images, effectively handling spectral and viewpoint variability for improved cross-spectral performance.

4.5.4.11 Handling Keypoint Sparsity Problem using Weighted Cross-Entropy Loss

In the experiments shown in Table 4.7, two examples are presented in each row, with two loss functions compared in the columns: regular Cross-Entropy (left) and Weighted Cross-Entropy (right). VIS-SAR, trained with standard cross-entropy, struggles with class imbalance, as evident in the images where one spectrum is densely filled with keypoints while the other contains none, resulting in zero matches between the two modalities. This imbalance limits the model's ability to predict accurate correspondences

Figure 4.7: *Visual comparison between Cross-Entropy and Weighted Cross-Entropy loss strategies.*



between the spectra, diminishing the matching score. To address this issue, a weighted cross-entropy loss was employed during training. This approach assigns lower weights to the losses of overrepresented classes, such as the dustbin class in the VIS-SAR setting, promoting more balanced predictions across both spectra, as demonstrated in the examples. The result is an improvement in keypoint distribution and matching accuracy between the modalities. Weighted loss mitigates the negative impact of class imbalance and increases the robustness of the model in cross-spectral matching tasks.

4.5.5 Summary of Experimental Results

Across five datasets, XPoint demonstrates strong adaptability, outperforming both traditional and state-of-the-art learning-based methods in repeatability, matching score, and homography estimation. The modular design enables flexibility between runtime and performance. Ablation studies validate the contribution of each architectural component, highlighting the importance of robust keypoint generation, transformer-based descriptors, and supervised geometric consistency.

Figure 4.8: *Qualitative results on three datasets—OPT-TH, VEDAI, and VIS-NIR—using four methods: XPoint (ours), ReDFeat, RoMa, and XoFTR. Top row per method shows MAGSAC-filtered keypoint matches; bottom row shows warped image alignment.*

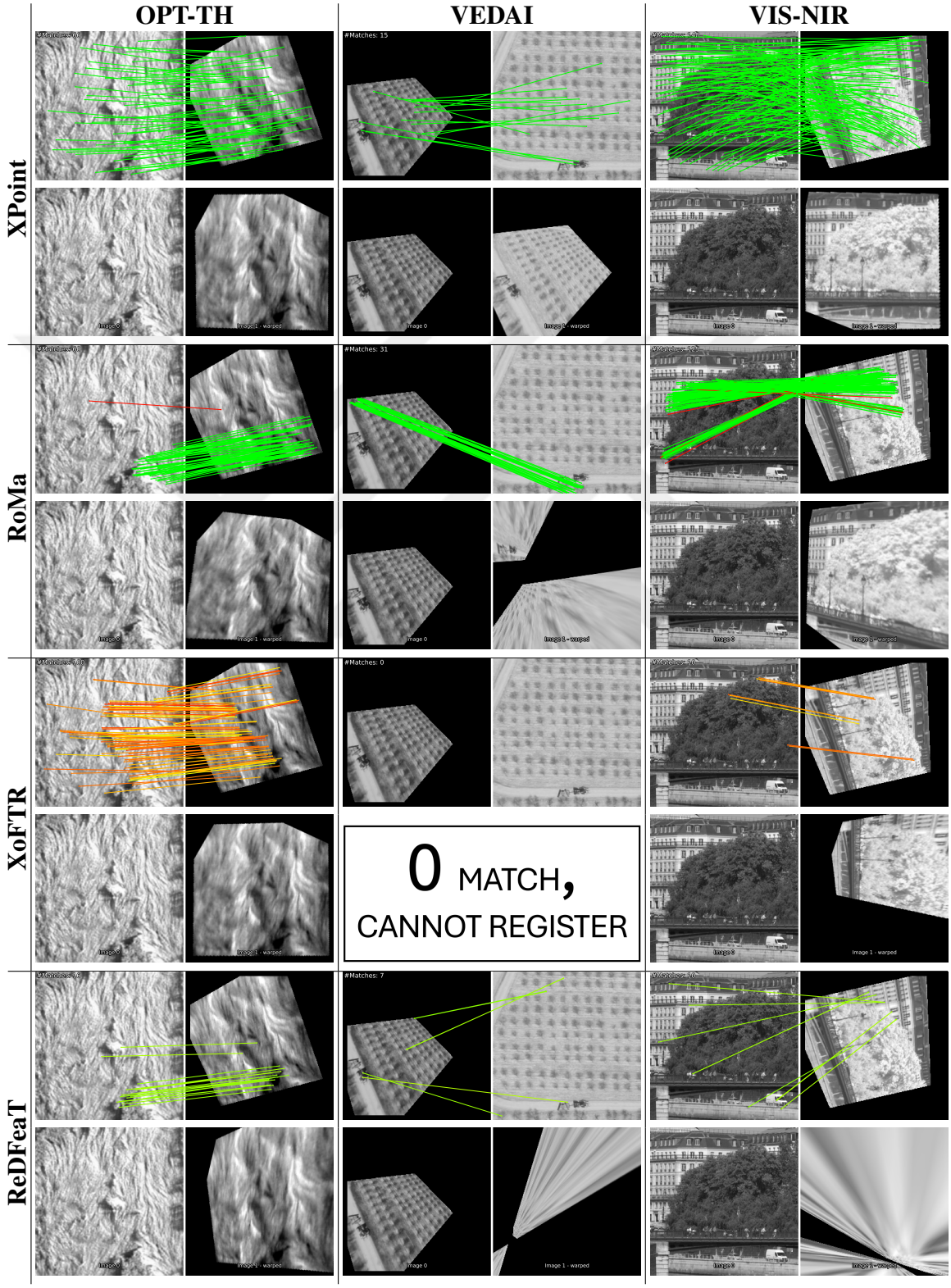
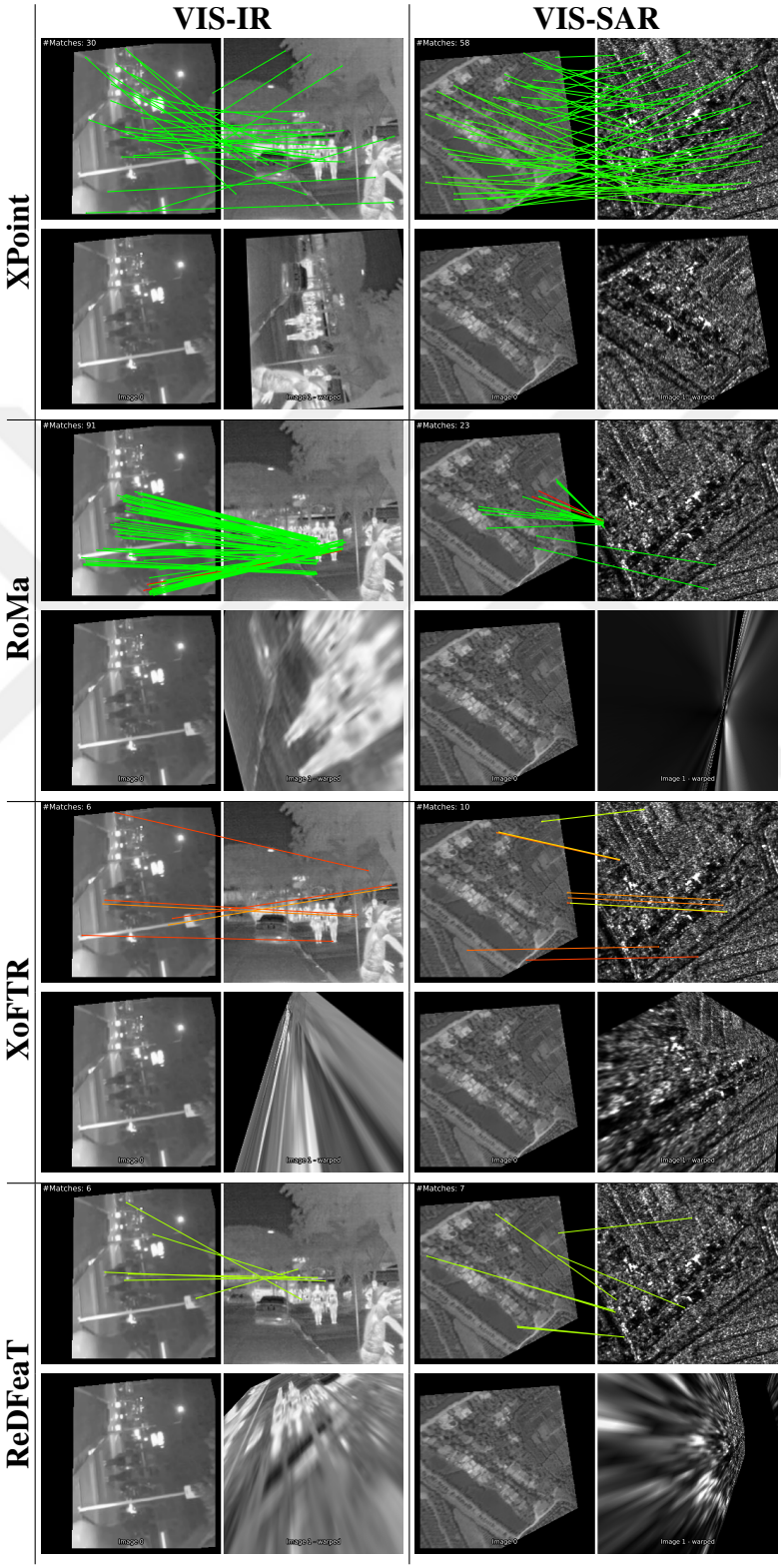


Figure 4.9: *Qualitative results on two datasets—VIS-IR and VIS-SAR—using the same four methods. Match quality and registration accuracy are compared using MAGSAC-filtered keypoints and resulting image warps.*



5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

This thesis presented a self-supervised deep learning framework for multispectral image matching, addressing long-standing challenges posed by nonlinear radiation distortions (NRDs), spectrum-invariant keypoint localization, and the scarcity of annotated datasets in real-world multimodal scenarios. To this end, we introduced two principal contributions.

First, we proposed an Improved Multispectral Homographic Adaptation strategy that enhances pseudo-ground-truth keypoint generation under spectral and viewpoint variability. This strategy employs a spectrum-aware windowing mechanism that augments repeatability and localization accuracy in feature detection, even across modality shifts such as visible–thermal or visible–SAR.

Second, we introduced XPoint, a fully self-supervised image matching architecture utilizing a pretrained VMamba encoder and modular decoder heads for keypoint detection, descriptor generation, and homography estimation. XPoint circumvents the need for dense annotations or expensive supervision (e.g., depth or pose data), enabling robust alignment using only aligned image pairs. Evaluations on five publicly available cross-spectral datasets demonstrate competitive or superior performance in feature matching and registration metrics, alongside notable computational efficiency.

Together, these components offer a scalable and label-efficient paradigm for robust multispectral matching in complex, real-world environments.

5.2 Future Work

Multi-modal image matching—aimed at aligning images captured from heterogeneous sensors or under varying environmental conditions—remains a challenging task. Disparities in modality, resolution, illumination, and viewpoint hinder the development of

generalizable solutions. While recent deep learning methods, particularly self-supervised approaches, have shown promise, current architectures often lack adaptability and scalability to unseen scenarios or diverse matching difficulties.

Future work could explore adaptive deep learning frameworks grounded in self-supervised meta-learning. The goal is to build modular architectures that quickly adapt to new modalities with minimal supervision. These systems could dynamically adjust computational complexity and matching strategies based on task difficulty, estimated by a learned model that assesses modality gaps, content differences, and geometric variation. An iterative pipeline may further enhance efficiency—starting with lightweight matching and escalating to more complex modules only when needed. This could be especially beneficial for real-time applications such as UAV navigation or medical image fusion. Another direction is rapid cross-modal adaptation. Incorporating meta-learning could enable the use of modality-agnostic base models, fine-tuned via few-shot self-supervised learning. This reduces the need for extensive retraining when facing novel image types.

Finally, the proposed methods have potential for deployment in practical domains: in remote sensing for environmental monitoring and disaster response; in medicine for precise CT-MRI alignment; and in autonomous systems for robust multimodal fusion across visual, thermal, and LiDAR data to ensure safer navigation in challenging conditions.

REFERENCES

[1] F. Achermann, A. Kolobov, D. Dey, T. Hinzmann, J. J. Chung, R. Siegwart, and N. Lawrance, "MultiPoint: Cross-spectral registration of thermal and optical aerial imagery," in *Conference on Robot Learning (CoRL)*, 2020.

[2] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," 2017.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, p. 346–359, 01 2008.

[4] Z. Kang, J. Yang, Z. Yang, and S. Cheng, "A review of techniques for 3d reconstruction of indoor environments," *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 330, 2020.

[5] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE intelligent vehicles symposium (IV)*, pp. 963–968, Ieee, 2011.

[6] Y. Zhao, S. Xu, S. Bu, H. Jiang, and P. Han, "Gslam: A general slam framework and benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1110–1120, 2019.

[7] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[8] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

[9] J. Wang, Y. Zhong, Y. Dai, S. Birchfield, K. Zhang, N. Smolyanskiy, and H. Li, "Deep two-view structure-from-motion revisited," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 8953–8962, 2021.

[10] M. Muzammul, A. Algarni, Y. Y. Ghadi, and M. Assam, "Enhancing uav aerial image analysis: Integrating advanced sahi techniques with real-time detection models on the visdrone dataset," *IEEE Access*, vol. 12, pp. 21621–21633, 2024.

[11] L. Miller, C. Pelletier, and G. I. Webb, "Deep learning for satellite image time-series analysis: A review," *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[12] Z. Hou, Y. Liu, and L. Zhang, "Pos-gift: A geometric and intensity-invariant feature transformation for multimodal images," *Information Fusion*, vol. 102, p. 102027, 2024.

[13] 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*, vol. 82, pp. 28–42, 2022.

[14] P. Chhabra, N. K. Garg, and M. Kumar, "Content-based image retrieval system using orb and sift features," *Neural Computing and Applications*, vol. 32, no. 7, pp. 2725–2733, 2020.

[15] G. Shamsipour, S. Fekri-Ershad, M. Sharifi, and A. Alaei, "Improve the efficiency of handcrafted features in image retrieval by adding selected feature generating layers of deep convolutional neural networks," *Signal, image and video processing*, vol. 18, no. 3, pp. 2607–2620, 2024.

[16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

[17] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE international conference on computer vision*, pp. 1385–1392, 2013.

[18] İ. C. Yağmur and H. F. Ates, "Improved homographic adaptation for keypoint generation in cross-spectral registration of thermal and optical imagery," in *Image and Signal Processing for Remote Sensing XXIX*, vol. 12733, pp. 67–73, SPIE, 2023.

[19] H. Zhou, T. Sattler, and D. W. Jacobs, "Evaluating local features for day-night matching," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 724–736, Springer, 2016.

[20] Y. Yao, Y. Zhang, Y. Wan, X. Liu, X. Yan, and J. Li, "Multi-modal remote sensing image matching considering co-occurrence filter," *IEEE Transactions on Image Processing*, vol. 31, pp. 2584–2597, 2022.

[21] B. Fan, Y. Yang, W. Feng, F. Wu, J. Lu, and H. Liu, "Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features," *IEEE Transactions on Multimedia*, 2022.

[22] H. Zhou, J. Ma, C. C. Tan, Y. Zhang, and H. Ling, "Cross-weather image alignment via latent generative model with intensity consistency," *ieee transactions on image processing*, vol. 29, pp. 5216–5228, 2020.

[23] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, 2014.

[24] Y. Zhang, G. Ma, and J. Wu, "Air-ground multi-source image matching based on high-precision reference image," *Remote Sensing*, vol. 14, no. 3, p. 588, 2022.

[25] T. Rouček, A. S. Amjadi, Z. Rozsypálek, G. Broughton, J. Blaha, K. Kusumam, and T. Krajník, "Self-supervised robust feature matching pipeline for teach and repeat navigation," *Sensors*, vol. 22, no. 8, p. 2836, 2022.

[26] X. Liu, Y. He, Y.-M. Cheung, X. Xu, and N. Wang, "Learning relationship-enhanced semantic graph for fine-grained image–text matching," *IEEE Transactions on Cybernetics*, pp. 1–14, 2022.

[27] K. Zhang, Z. Mao, A.-A. Liu, and Y. Zhang, "Unified adaptive relevance distinguishable attention network for image-text matching," *IEEE Transactions on Multimedia*, vol. 25, pp. 1320–1332, 2023.

[28] S. Long, S. C. Han, X. Wan, and J. Poon, "Gradual: Graph-based dual-modal representation for image-text matching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3459–3468, 2022.

[29] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Information Fusion*, vol. 73, pp. 22–71, 2021.

[30] Y. Zhang and R. K. Mishra, "Remote sensing image fusion: A survey of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 2, no. 4, pp. 8–22, 2010.

[31] M. Schmitt and X. X. Zhu, "A comprehensive system for automatic fusion of multi-sensor data for sar and optical image alignment," *Remote Sensing*, vol. 8, no. 3, p. 243, 2016.

[32] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in Medicine & Biology*, vol. 46, no. 3, p. R1, 2001.

[33] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[34] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[35] A. Wu, W.-S. Zheng, and J.-H. Lai, "Rgb-infrared cross-modality person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5380–5389, 2017.

[36] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999.

[37] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," *CVPR*, 2021.

[38] D. Zhao, Y. Yang, Z. Ji, and X. Hu, "Rapid multimodality registration based on mm-surf," *Neurocomputing*, vol. 131, pp. 87–97, 2014.

[39] J. Lange, R. O'Shaughnessy, and M. Rizzo, "Rapid and accurate parameter inference for coalescing, precessing compact binaries," 2018.

[40] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," 2023.

[41] I. C. Yagmur, H. F. Ates, and B. K. Gunturk, "Xpoint: A self-supervised visual-state-space based architecture for multispectral image registration," 2024.

[42] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, p. 381–395, June 1981.

[43] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "Magsac++, a fast, reliable and accurate robust estimator," 2019.

[44] Q. Huang, X. Guo, Y. Wang, H. Sun, and L. Yang, "A survey of feature matching methods," *IET Image Process.*, vol. 18, pp. 1385–1410, 2024.

[45] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Upper Saddle River, NJ: Pearson Education, 3rd ed., 2008.

[46] J. Yuan and D. Zhang, "An adaptive method for image matching based on normalized cross-correlation," *Pattern Recognition Letters*, vol. 27, no. 9, pp. 1002–1009, 2006.

[47] P. M. Teillet, "Image correction for radiometric effects in remote sensing," *International Journal of Remote Sensing*, vol. 7, no. 12, pp. 1637–1651, 1986.

[48] M. M. Rahman, G. J. Hay, and I. Couloigner, "An assessment of polynomial regression techniques for the relative radiometric normalization (rrn) of high-resolution multi-temporal airborne thermal infrared (tir) imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 95, pp. 1–13, 2014.

[49] S. Liu, J. Lian, X. Zhan, C. Liu, Y. Tian, and H. Duan, "Automatically eliminating seam lines with poisson editing in complex relative radiometric normalization mosaicking scenarios," *arXiv preprint arXiv:2106.07441*, 2021.

[50] Y. Xu, Q. Xu, X. Xie, and X. Huang, "Local feature matching using deep learning: A survey," *Pattern Recognition*, 2024.

[51] Y. Liao, Y. Di, K. Zhu, H. Zhou, M. Lu, Y. Zhang, Q. Duan, and J. Liu, "Local feature matching from detector-based to detector-free: a survey," *Applied Intelligence*, vol. 54, pp. 3954–3989, 2024.

[52] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

[53] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local Feature Matching at Light Speed," in *ICCV*, 2023.

[54] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the Alvey Vision Conference*, pp. 23.1–23.6, Alvety Vision Club, 1988. doi:10.5244/C.2.23.

[55] J. Revaud, P. Weinzaepfel, C. D. Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: Repeatable and reliable detector and descriptor," 2019.

[56] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.net: Keypoint detection by handcrafted and learned cnn filters," 2019.

[57] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," 2016.

[58] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," pp. 2564–2571, 11 2011.

[59] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2828–2837, 2017.

[60] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, and H. Liao, "Learning two-view correspondences and geometry using order-aware network," *International Conference on Computer Vision (ICCV)*, 2019.

[61] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6128–6136, 2017.

[62] F. R. J. M. Anastasiya Mishchuk, Dmytro Mishkin, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proceedings of NeurIPS*, Dec. 2017.

[63] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan, *GeoDesc: Learning Local Descriptors by Integrating Geometry Constraints*, p. 170–185. Springer International Publishing, 2018.

[64] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "Sosnet: Second order similarity regularization for local descriptor learning," 2019.

[65] X. Zhao, X. Wu, W. Chen, P. C. Y. Chen, Q. Xu, and Z. Li, "Aliked: A lighter keypoint and descriptor extraction network via deformable transformation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[66] Z. Luo, T. Shen, L. Zhou, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Contextdesc: Local descriptor augmentation with cross-modality context," 2019.

[67] H. Xie, Y. Zhang, J. Qiu, X. Zhai, X. Liu, Y. Yang, S. Zhao, Y. Luo, and J. Zhong, "Semantics lead all: Towards unified image registration and fusion from a semantic perspective," *Information Fusion*, vol. 98, p. 101835, 2023.

[68] X. Chen, C. Fu, M. Tie, C.-W. Sham, and H. Ma, "Affnet: An attention-based feature-fused network for surface defect segmentation," *Applied Sciences*, vol. 13, no. 11, 2023.

[69] P. Truong, S. Apostolopoulos, A. Mosinska, S. Stucky, C. Ciller, and S. D. Zanet, "Glampoints: Greedily learned accurate match points," 2020.

[70] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning feature descriptors using camera pose supervision," 2024.

[71] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint detection and description of local features," 2019.

[72] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng, and Z. He, "RF-Net: An End-To-End Image Matching Network Based on Receptive Field," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8124–8132, June 2019. ISSN: 2575-7075.

[73] A. Bhowmik, S. Gumhold, C. Rother, and E. Brachmann, "Reinforced Feature Points: Optimizing Feature Detection and Description for a High-Level Task," pp. 4948–4957, 2020.

[74] A. Barroso-Laguna, Y. Verdie, B. Busam, and K. Mikolajczyk, "HDD-Net: Hybrid Detector Descriptor with Mutual Interactive Learning," in *Computer Vision – ACCV 2020* (H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, eds.), (Cham), pp. 500–516, Springer International Publishing, 2021.

[75] U. S. Parihar, A. Gujarathi, K. Mehta, S. Tourani, S. Garg, M. Milford, and K. M. Krishna, "RoRD: Rotation-Robust Descriptors and Orthographic Views for Local Feature Matching," Mar. 2022. arXiv:2103.08573 [cs].

[76] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Aslfeat: Learning local features of accurate shape and localization," *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[77] F. Xue, I. Budvytis, and R. Cipolla, "SFD2: Semantic-guided Feature Detection and Description," June 2023. arXiv:2304.14845 [cs].

[78] O. Wiles, S. Ehrhardt, and A. Zisserman, "Co-attention for conditioned image matching," 2021.

[79] K. Li, L. Wang, L. Liu, Q. Ran, K. Xu, and Y. Guo, "Decoupling makes weakly supervised local feature better," 2022.

[80] Y. Deng and J. Ma, "Redfeat: Recoupling detection and description for multimodal feature learning," *IEEE Transactions on Image Processing*, vol. 32, pp. 591–602, 2022.

[81] J. Sun, L. Ji, and J. Zhu, "Shared coupling-bridge scheme for weakly supervised local feature learning," vol. 26, pp. 1200–1212.

[82] H. Chen, Z. Luo, J. Zhang, L. Zhou, X. Bai, Z. Hu, C.-L. Tai, and L. Quan, "Learning to match features with seeded graph matching network," *International Conference on Computer Vision (ICCV)*, 2021.

[83] Y. Shi, J.-X. Cai, Y. Shavit, T.-J. Mu, W. Feng, and K. Zhang, "Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching," 2023.

[84] R. Pautrat, I. Suárez, Y. Yu, M. Pollefeys, and V. Larsson, "GlueStick: Robust image matching by sticking points and lines together," in *International Conference on Computer Vision (ICCV)*, 2023.

[85] Z. Kuang, J. Li, M. He, T. Wang, and Y. Zhao, "Densegap: Graph-structured dense correspondence learning with anchor points," 2022.

[86] J. Liu and X. Zhang, "Drc-net: Densely connected recurrent convolutional neural network for speech dereverberation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 166–170, 2022.

[87] G. Bökman and F. Kahl, "A case for using rotation invariant features in state of the art feature matchers," in *CVPRW*, 2022.

[88] S. Tang, J. Zhang, S. Zhu, and P. Tan, "Quadtree attention for vision transformers," 2022.

[89] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Matchformer: Interleaving attention in transformers for feature matching," in *Asian Conference on Computer Vision*, 2022.

[90] K. T. Giang, S. Song, and S. Jo, "Topicfm: Robust and interpretable topic-assisted feature matching," *arXiv preprint arXiv:2207.00328*, 2022.

[91] D. Huang, Y. Chen, S. Xu, Y. Liu, W. Wu, Y. Ding, C. Wang, and F. Tang, "Adaptive assignment for geometry aware local feature matching," 2023.

[92] T. Xie, K. Dai, K. Wang, R. Li, and L. Zhao, "Deepmatcher: A deep transformer-based network for robust and accurate local feature matching," 2023.

[93] K. Dai, T. Xie, K. Wang, Z. Jiang, R. Li, and L. Zhao, "Oamatcher: An overlapping areas-based network with label credibility for robust and accurate feature matching," *Pattern Recognition*, vol. 147, p. 110094, Mar. 2024.

[94] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient LoFTR: Semi-dense local feature matching with sparse-like speed," in *CVPR*, 2024.

[95] Q. Wu, H. Jiang, Y. Ding, L. Luo, J. Xie, and J. Yang, "Diff-reg v2: Diffusion-based matching matrix estimation for image matching and 3d registration," 2025.

[96] X. Yan, L. Lin, N. J. Mitra, D. Lischinski, D. Cohen-Or, and H. Huang, "Shapeformer: Transformer-based shape completion via sparse representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6239–6249, 2022.

[97] Ö. Tuzcuoğlu, A. Köksal, B. Sofu, S. Kalkan, and A. A. Alatan, "Xoftr: Cross-modal feature matching transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4275–4286, 2024.

[98] X. Lu and S. Du, "Jamma: Ultra-lightweight local feature matching with joint mamba," 2025.

[99] I. Melekhov, A. Tiulpin, T. Sattler, M. Pollefeys, E. Rahtu, and J. Kannala, "DGC-Net: Dense geometric correspondence network," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[100] P. Truong, M. Danelljan, and R. Timofte, "Glu-net: Global-local universal network for dense flow and correspondences," 2021.

[101] P. Truong, M. Danelljan, L. V. Gool, and R. Timofte, "Learning accurate dense correspondences and when to trust them," 2021.

[102] P. Truong, M. Danelljan, R. Timofte, and L. V. Gool, "Pdc-net+: Enhanced probabilistic dense correspondence network," 2021.

[103] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, "Dkm: Dense kernelized feature matching for geometry estimation," 2022.

[104] U. Efe, K. G. Ince, and A. A. Alatan, "Dfm: A performance baseline for deep feature matching," 2021.

[105] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "Cotr: Correspondence transformer for matching across images," 2021.

[106] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. Mckinnon, Y. Tsin, and L. Quan, "Aspanformer: Detector-free image matching with adaptive span transformer," in *European Conference on Computer Vision*, pp. 20–36, Springer, 2022.

[107] Y. Di, Y. Liao, K. Zhu, H. Zhou, Y. Zhang, Q. Duan, J. Liu, and M. Lu, "Mivi: multi-stage feature matching for infrared and visible image," vol. 40, p. 1839–1851, May 2023.

[108] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," 2016.

[109] F. E. Nowruzi, R. Laganiere, and N. Japkowicz, "Homography estimation from image pairs with hierarchical convolutional networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 904–911, 2017.

[110] C. Wang, X. Wang, X. Bai, Y. Liu, and J. Zhou, "Self-supervised deep homography estimation with invertibility constraints," *Pattern Recognition Letters*, vol. 128, pp. 355–360, 2019.

[111] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," in *RA-L*, IEEE, 2018.

[112] M. Huo, Z. Zhang, X. Ren, and X. Yang, "Abhe: All attention-based homography estimation," 2023.

[113] S.-Y. Cao, R. Zhang, L. Luo, B. Yu, Z. Sheng, J. Li, and H.-L. Shen, "Recurrent homography estimation using homography-guided image warping and focus transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9833–9842, June 2023.

[114] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[115] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1879–1888, June 2022.

[116] B. Hou, J. Ren, and W. Yan, "Unsupervised multi-scale-stage content-aware homography estimation," *Electronics*, vol. 12, no. 9, 2023.

[117] N. e. a. Pielawski, "CoMIR: Contrastive multimodal image representation for registration," 2020. arXiv:2006.06325 [cs].

[118] J. Li, P. Shi, Q. Hu, and Y. Zhang, "Rift2: Speeding-up rift with a new rotation-invariance technique," 2023.

[119] X. Jiang, J. Ren, Z. Li, X. Zhou, D. Liang, and X. Bai, "Minima: Modality invariant image matching," 2024.

[120] X. He, H. Yu, S. Peng, D. Tan, Z. Shen, H. Bao, and X. Zhou, "Matchanything: Universal cross-modality image matching with large-scale pre-training," 2025.

[121] "Multimodal retinal image registration using a fast principal component analysis hybrid-based similarity measure," in *ICIP*, 2013.

[122] X. e. a. Shen, "Multi-modal and multi-spectral registration for natural images," in *ECCV*, (Cham), pp. 309–324, Springer, 2014.

[123] Y. Ye and L. Shen, "HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-1, pp. 9–16, 2016.

[124] Q. e. a. Zeng, "Real-time adaptive visible and infrared image registration based on morphological gradient and c_sift," *Journal of Real-Time Image Processing*, vol. 17, pp. 1103–1115, 2020.

[125] F. e. a. Dellinger, "SAR-SIFT: A sift-like algorithm for sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 453–466, 2015.

[126] R. Wang, Y. Shi, and W. Cao, "GA-SURF: A new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra," *Pattern Recognition Letters*, vol. 127, pp. 11–17, 2019.

[127] Z. e. a. Fu, "A local feature descriptor based on combination of structure and texture information for multispectral image matching," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, pp. 100–104, 2019.

[128] Q. Wu and S. Zhu, "Multispectral image matching method based on histogram of maximum gradient and edge orientation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[129] Q. e. a. Xiong, "Multispectral image matching considering absolute phase consistency gradient and log-polar description," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[130] Y. Yao, Y. Zhang, Y. Wan, X. Liu, X. Yan, and J. Li, "Multi-modal remote sensing image matching considering co-occurrence filter," *IEEE Transactions on Image Processing*, vol. 31, pp. 2584–2597, 2022. Publisher: IEEE.

[131] D. Quan, H. Wei, S. Wang, R. Lei, B. Duan, Y. Li, B. Hou, and L. Jiao, "Self-distillation feature learning network for optical and sar image registration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[132] Y. Liu, W. He, and H. Zhang, "GRiD: Guided refinement for detector-free multimodal image matching," *IEEE Transactions on Image Processing*, vol. 33, pp. 5892–5906, 2024.

[133] M. Elsaeidy, M. E. Erkol, B. K. Güntürk, and H. F. Ateş, "Infrared-to-optical image translation for keypoint-based image registration," in *SIU*, pp. 1–4, 2022.

[134] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[135] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutualinformation," vol. 16, pp. 14–22, 07 1996.

[136] L. Dai, Q. Guan, and H. Liu, "Robust image registration of printed circuit boards using improved sift-pso algorithm," *The Journal of Engineering*, vol. 2018, no. 16, pp. 1793–1797, 2018.

[137] C. A. Aguilera, A. D. Sappa, and R. Toledo, "Lghd: A feature descriptor for matching across non-linear intensity variations," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 178–181, IEEE, 2015.

[138] E. B. Baruch and Y. Keller, "Joint detection and matching of feature points in multimodal images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6585–6593, 2021.

[139] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery : A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, 03 2015.

[140] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, June 2022.

[141] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017.

[142] Y. Li, W. Pei, and Z. He, "Srhen: Stepwise-refining homography estimation network via parsing geometric correspondences in deep latent space," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3063–3071, 2020.

[143] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[144] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, pp. 1615–30, 11 2005.