A Hybrid Space Model for Misaligned Multi-modality Image Fusion

Yi Xiao[1*], Jia Wang[1*], Zhu Liu[1], Di Wang[2], Jinyuan Liu[1], Risheng Liu[1†]

[1]School of Software Technology, Dalian University of Technology, Dalian, China
[2]School of Computer Science and Artificial Intelligence, Civil Aviation University of China, Tianjin, China
xiaoyi@mail.dlut.edu.cn, jiawang0704@outlook.com, rsliu@dlut.edu.cn

Abstract

Infrared and visible image fusion aims to integrate complementary information, such as thermal saliency from infrared imagery and fine-grained texture details from visible imagery. However, real-world multi-modal misalignment and geometric deformation often introduce severe artifacts. Most existing methods focus on feature extraction within Euclidean space, thereby neglecting the inherent hierarchical structures embedded in multimodal representations. While Euclidean space excels at preserving local structural details and supporting efficient computation, hyperbolic space is naturally suited for modeling hierarchical relationships due to its geometric properties. Building upon these observations, this paper proposes a unified framework that jointly optimizes image registration and fusion through a dual-space architecture. This architecture synergistically combines the local fidelity of Euclidean geometry with the hierarchical modeling capability of hyperbolic geometry to enhance multimodal representation learning. Specifically, this paper introduces Hyperbolic Coupled Contrastive Learning Optimization (HCCLO), which aligns and optimizes the hierarchical structures of infrared and visible embeddings in hyperbolic space. Moreover, this paper designs a task-adaptive dual-space features fusion mechanism, which dynamically balances and fuses Euclidean local features with hyperbolic hierarchical representations, thereby improving adaptability for downstream tasks. Extensive experiments on misaligned multimodal datasets demonstrate that our method achieves state-of-the-art performance, while effectively capturing both spatial dependencies and hierarchical semantics.

Code — https://github.com/xiao-eee/HMMF.git

Introduction

Due to limitations in illumination conditions and sensor hardware, a single imaging modality can only capture partial scene information. Multi-modality image fusion aims to integrate complementary information from multiple source images to generate a more informative and perceptually enhanced fused image. Among various fusion tasks, infrared and visible image fusion (IVIF) has attracted extensive research attention owing to its unique advantages in combining thermal saliency and detailed texture. (Liu et al. 2022b) The resulting fused images offer more comprehensive scene representations and improved visual perception, thereby benefiting downstream computer vision applications such as semantic segmentation (Li et al. 2023a; Liu et al. 2023a), object detection (Liu et al. 2022a; Zhao et al. 2023a), scene understanding (Huang et al. 2020), and autonomous driving systems.

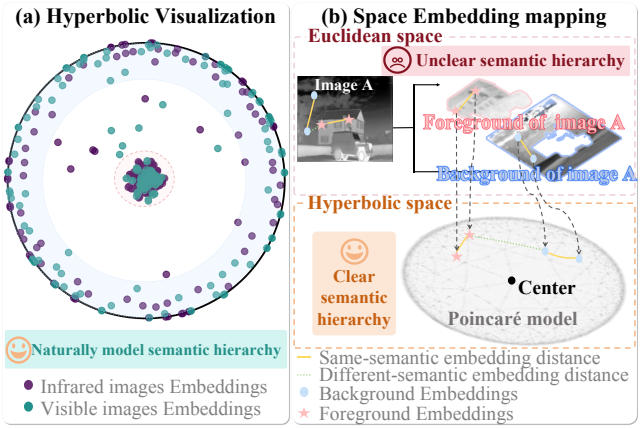


Figure 1: The superior capability of hyperbolic space in modeling semantic hierarchies. (a) Hyperbolic t-SNE visualizations of visible and infrared image embeddings reveal a clear hierarchical structure, particularly around the center and boundary regions of the Poincaré ball. (b) Hyperbolic visualizations of foreground and background features reveal a similar hierarchical organization, where features that are close in Euclidean space become distinctly separated in the Poincaré ball due to the semantic hierarchical capability of hyperbolic space.

Over the past decades, a variety of IVIF approaches have been proposed with the primary objective of enhancing fusion quality. Traditional methods are mainly categorized into five groups: Multi-Scale Transform (MST)-based methods (Li, Wu, and Kittler 2020), Sparse Representation (SR)-based methods (Liu et al. 2016), Subspace Decomposition-based methods (Lu et al. 2014), Saliency-Driven methods (Ma et al. 2017), and Optimization model-based methods (Ma et al. 2016), among others. However, these meth-

*These authors contributed equally.
†Corresponding author.

ods often rely on handcrafted designs and involve computationally intensive procedures. Recently, deep learning-based approaches have been introduced into IVIF, demonstrating significant progress. These methods (Ma et al. 2020; Zhang et al. 2020; Tang et al. 2023; Zhou et al. 2023; Zhang et al. 2021) focus on learning salient and representative features from the source modalities (Li et al. 2023b), achieving superior performance under various evaluation metrics.

However, most existing fusion methods (Zhao et al. 2023c, 2024b,a; Liu et al. 2024a; Li et al. 2024; Liu et al. 2025, 2024d) are designed for ideally aligned image pairs, and their performance degrades significantly in the presence of real-world geometric distortions. A common solution to the misaligned IVIF problem is the "register-then-fuse" pipeline (Gao et al. 2019; Wang et al. 2022), where image registration is performed as a preprocessing step prior to fusion. However, this sequential strategy suffers from a key limitation: fusion, treated as a downstream task, cannot provide feedback to improve the registration process. To overcome this limitation, several recent works (Xu et al. 2022; Xu, Yuan, and Ma 2023) propose two-stage joint training frameworks that optimize both registration and fusion simultaneously, leading to improved overall performance. In parallel, (Tang et al. 2022) design a shared encoder architecture with cross-modality feature consistency constraints, which facilitates the learning of modality-invariant features. Nevertheless, this approach tends to suppress modality-specific details, limiting the richness of the fused representations (Li et al. 2025).

In addition, the core of IVIF lies in preserving modality-specific complementary features while suppressing redundant or irrelevant information. In this paper, we revisit the task from a hierarchical semantic perspective (Liu et al. 2023b), noting that each modality inherently exhibits distinct structural hierarchies in which infrared (IR) images emphasize foreground thermal saliency and visible (VIS) images provide detailed background textures. To effectively integrate these hierarchical semantics, we explore hyperbolic representation learning, which inherently models tree-like structures due to its exponential volume growth and negative curvature properties. As illustrated in Fig.1(a), hyperbolic space captures the semantic hierarchy of cross-modality embeddings more effectively than Euclidean space. Moreover, as shown in Fig.1(b), features that are semantically distinct yet spatially close in Euclidean space become well-separated in the Poincaré ball model, enabling enhanced semantic discrimination. By leveraging curvature-aware representations, hyperbolic modeling achieves geometric consistency across modalities through hierarchy-preserving alignment. However, most existing approaches remain confined to Euclidean space. This representation efficiently preserves local structural information but fails to capture the hierarchical relationships that are essential for effective cross-modality fusion.

Motivated by the complementary strengths of hyperbolic and Euclidean geometries, a unified framework is proposed that jointly optimizes image registration and fusion within a dual-space architecture. In this framework, hyperbolic geometry effectively disentangles semantic hierarchies, while Euclidean space preserves local geometric fidelity. By explicitly modeling the cross-modal hierarchical dependencies inherent in infrared (IR) and visible (VIS) images, the proposed approach enhances fusion quality while maintaining semantic consistency, providing a robust foundation for downstream vision tasks. (Liu et al. 2024c) To further improve hierarchical representation learning and guide the dual-space optimization process, the Hyperbolic Coupled Contrastive Learning Optimization (HCCLO) module is introduced. This module aligns and optimizes cross-modal features in hyperbolic space, ensuring improved hierarchical consistency across modalities. Additionally, a task-adaptive fusion mechanism is proposed, dynamically combining Euclidean local features with hyperbolic hierarchical embeddings through learnable, task-sensitive weighting coefficients. This mechanism ensures flexible and robust adaptation to various downstream applications. The primary contributions of this work are as follows:

- A novel dual-space framework is presented for jointly optimizing multi-modality image registration and fusion. By leveraging the hierarchical modeling capabilities of hyperbolic space, this framework constructs a cross-modal hierarchical correlation structure between IR and VIS modalities, effectively mitigating edge ghosting and geometric distortions in misaligned image fusion.
- The HCCLO module is proposed as a contrastive learning strategy in hyperbolic space, enhancing hierarchical feature alignment across modalities and enforcing geometric consistency during joint registration and fusion.
- A task-adaptive dual-space features fusion mechanism is introduced, which integrates Euclidean local features with hyperbolic hierarchical features through dynamically learned weights, enabling robust and flexible adaptation to a wide range of downstream tasks.

Related Works

Registered IR-VIS Image Fusion

Image fusion techniques have evolved from traditional methods to deep learning-based approaches, achieving substantial improvements in feature extraction and fusion quality. Early deep fusion methods based on autoencoders (AEs) (Li and Wu 2018; Zhao et al. 2021; Liu et al. 2021) and convolutional neural networks (CNNs) (Ma et al. 2021; Li, Wu, and Kittler 2021) focused on learning shared representations from registered image pairs. However, these methods mainly focused on feature extraction but had difficulty preserving fine details. To address these limitations, Transformer-based models (Fu et al. 2022) have been introduced, demonstrating superior capabilities in modeling long-range dependencies and capturing global context. Combining the strengths of CNNs in local feature modeling and Transformers in global reasoning, hybrid CNN–Transformer architectures (Li et al. 2022) have emerged as the prevailing feature extraction approach in IVIF by integrating their complementary strengths.

Generative adversarial networks (GANs) have also been widely applied (Ma et al. 2019; Liu et al. 2022a; Ma et al.

2020; Zhang et al. 2021; Zhou et al. 2023), leveraging adversarial learning to enhance the realism and fidelity of fusion results. While early GAN-based methods employed separate discriminators for each modality, which limited their ability to model cross-modal dependencies, recent advances have introduced dual-discriminator frameworks (Ma et al. 2020; Zhang et al. 2021; Zhou et al. 2023) that better exploit complementary features and generate more semantically consistent fusion results.

In recent years, task-oriented fusion approaches (Liu et al. 2021, 2022a) have gained prominence by incorporating high-level semantic objectives, moving beyond simple visual quality to optimize for downstream tasks such as detection and classification. However, these methods typically assume that input images are well-aligned, which is rarely the case in real-world scenarios. Achieving accurate cross-modal registration remains highly challenging due to the substantial appearance disparities between infrared and visible images, often requiring additional alignment procedures to ensure reliable fusion.

Cross-Modality Image Registration and Fusion

The advancement of cross-modal image registration remains constrained by the lack of effective similarity metrics, particularly between infrared and visible modalities. Recently, several joint registration and fusion approaches (Li et al. 2025) have been proposed to address the misalignment problem in multi-modality image fusion. Representative methods include (Liu et al. 2022b; Wang et al. 2022; Xu, Yuan, and Ma 2023). The methods in (Gao et al. 2019; Wang et al. 2022) adopt a two-stage pipeline that separately trains registration and fusion modules. While this framework improves overall fusion performance, the first method overlooks the influence of cross-modal discrepancies on registration accuracy, and the second is limited by the quality of generated intermediate transformed images. More importantly, their separate optimization inhibits mutual enhancement between tasks. In contrast, (Xu, Yuan, and Ma 2023) proposes a two-stage registration framework with coarse-to-fine refinement, where the fine stage is jointly optimized with fusion. Meanwhile, (Tang et al. 2022) introduces a shared encoder to support cross-modal alignment, constructing a unified framework for registration, fusion, and semantic segmentation with bidirectional alignment and semantic awareness. Despite the rich and complementary nature of IR and VIS imagery, their shared features often lack fine detail. To address this, this paper introduces a hyperbolic geometry-based framework that leverages inter-modal complementarity and semantic hierarchies to achieve geometrically constrained alignment of features in hyperbolic space, enhancing the modeling of cross-modal hierarchical structures.

Hyperbolic Representation Learning

Hyperbolic space, characterized by constant negative curvature, causes metric distances to grow exponentially with radius. This property endows hyperbolic representations with a unique advantage in capturing hierarchical structures (Nickel and Kiela 2017; De Sa et al. 2018; Ramasinghe et al. 2024). In recent years, hyperbolic representation learning has been widely applied in natural language processing (Ganea, Bécigneul, and Hofmann 2018; Zhang and Gao 2021) and computer vision (Atigh et al. 2022; Wang et al. 2023), demonstrating superior performance in modeling hierarchical relationships and complex structured data. Furthermore, recent studies have leveraged hyperbolic learning to model inherent hierarchical properties across modalities in multimodal vision tasks (Ramasinghe et al. 2024; Kong et al. 2024).

In this work, this paper leverages hyperbolic space to compensate for the limitations of Euclidean geometry in modeling modality-specific hierarchical structures. By introducing hierarchical alignment constraints across modalities, our method preserves structural hierarchies while maintaining modality-specific characteristics. This design enhances both registration and fusion performance by providing a novel perspective grounded in hyperbolic geometry.

Proposed Method

Overall Framework

Fig.2 illustrates the overall workflow of the proposed method. Given a pair of misaligned infrared (IR) and visible (VIS) images, the goal of joint registration and fusion is to align the IR image with the VIS image and integrate complementary information from both modalities into a single fused image, ensuring that the resulting output is free from structural distortions and edge artifacts. The process is divided into four main stages:

First, a shared encoder is employed to extract modality-invariant features, reducing cross-modal discrepancies and mitigating alignment challenges between the infrared and visible images, while maintaining computational efficiency. This step prepares the images for the subsequent stages.

Second, a global transformation matrix predictor is introduced to map both modalities into a unified coordinate system. This step performs coarse registration between the misaligned IR and VIS images, providing an initial geometric transformation that facilitates finer alignment in subsequent stages.

Third, dual-space encoders are employed to extract both local and hierarchical features. The Euclidean encoder uses a Transformer-CNN architecture with a multi-window attention mechanism to capture fine-grained global features. In parallel, the hyperbolic encoder, acting as a complementary branch, enhances hierarchical information extraction through the proposed Hyperbolic Coupled Contrastive Learning Optimization (HCCLO), refining structural alignment with geometric constraints.

Fourth, the extracted hyperbolic hierarchical features and Euclidean local features are adaptively fused in a task-specific manner. This process employs two specialized decoders: the registration decoder utilizes a multi-scale dense registration subnetwork to achieve precise geometric alignment, while the fusion decoder incorporates a dual-channel attention mechanism to selectively emphasize complementary information. These decoders generate the final outputs: the registered infrared image and the fused image.

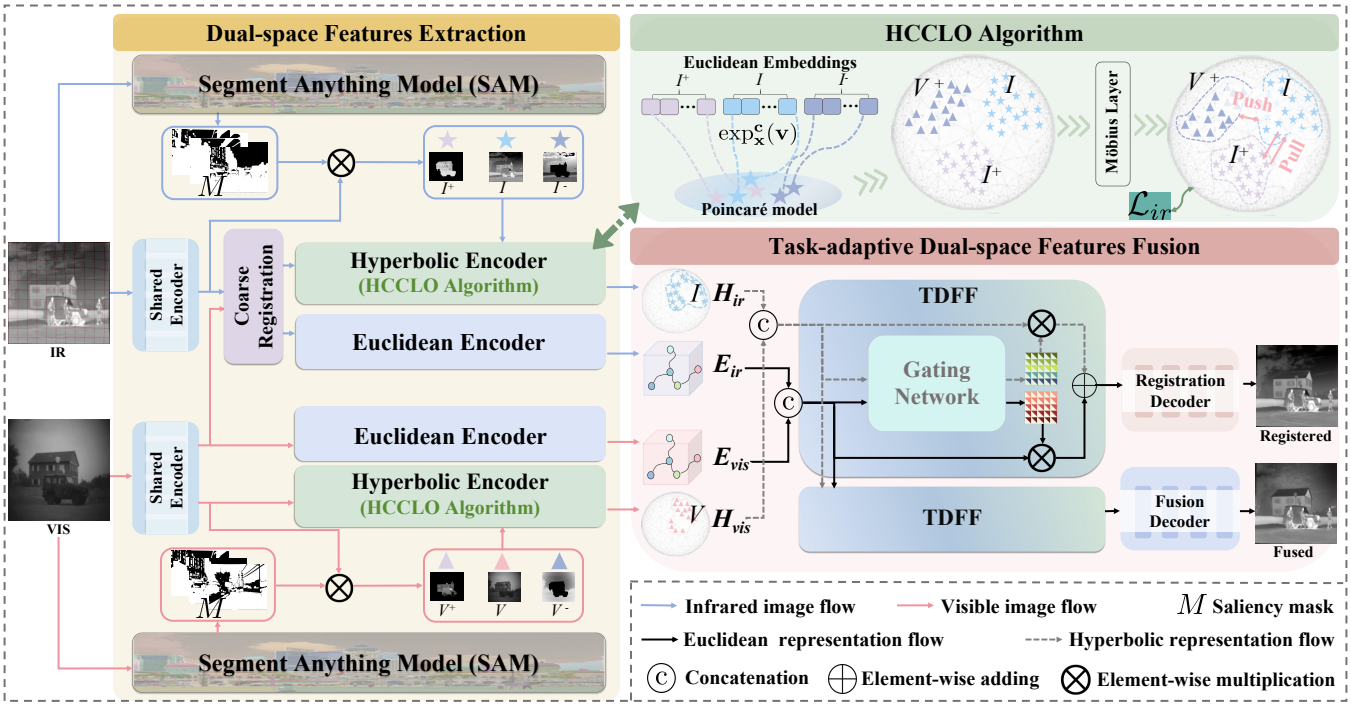


Figure 2: Overview of the proposed HMMF framework. It adopts a dual-space architecture that jointly optimizes multi-modality image registration and fusion. HCCLO aligns hyperbolic hierarchical representations between modalities, while TDFF adaptively combines Euclidean local and hyperbolic hierarchical features to enhance fusion and registration performance.

Unsupervised Hyperbolic Representation Learning

This section focuses on mapping features from Euclidean space to hyperbolic space. Concurrently, this paper proposes the Hyperbolic Coupled Contrastive Learning Optimization (HCCLO) to align and optimize this mapping process, effectively aligning cross-modal hierarchical structures via hyperbolic constrained optimization while preserving modality-specific discrepancies.

Poincaré Ball Model of Hyperbolic Geometry. Hyperbolic spaces are Riemannian manifolds characterized by constant negative curvature c, fundamentally differing from the zero-curvature geometry of Euclidean spaces. The Poincaré ball model represents hyperbolic space with the metric $g_p = (\lambda_x^2) g_e$, where the conformal factor is $\lambda_{c,x} = \frac{2}{1 - c||x||^2}$ and g_e denotes the Euclidean metric tensor. The curvature parameter c controls the curvature magnitude and the radius of the Poincaré ball. This metric causes regions near the boundary of the ball to be magnified, which makes the Poincaré ball model particularly effective for representing tree-like hierarchical structures with minimal distortion. In our hyperbolic encoder, the curvature parameter c is trainable and optimized during training, allowing the model to adaptively learn the optimal geometric structure for representing cross-modal hierarchical relationships. Due to the intrinsic curvature difference, vector operations defined in Euclidean space are fundamentally incompatible with hyperbolic geometry. Therefore, specialized operations are required as follows:

1) Möbius Addition: The Möbius addition is fundamental in hyperbolic geometry, as it preserves hyperbolic distance and maintains the non-Euclidean structure of the space. Given two points x and y in the hyperbolic space, their Möbius addition is defined as follows:

$$x \oplus y = \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2}, \quad (1)$$

where $\langle x, y\rangle$ represents their Euclidean inner product, $\|x\|$ and $\|y\|$ denote the Euclidean norms of x and y, respectively, and c is the curvature parameter that controls the negative curvature of the hyperbolic space. Möbius scalar multiplication and matrix multiplication are similarly defined to preserve the hyperbolic geometry.

2) Distance Function in the Poincaré Ball: The distance d_p between points x and y in the Poincaré ball is defined as:

$$d_p(x, y) = cosh^{-1}(1 + 2\frac{||x - y||^2}{(1 - ||x||^2)(1 - ||y||^2)}). \quad (2)$$

3) Exponential Map in the Poincaré Ball: The exponential map $exp_z(v)$ at a base point $z \in \mathbb{D}^n$ within the Poincaré ball model of hyperbolic space maps a tangent vector $v \in T_z\mathbb{D}^n$ to the manifold through the closed-form expression:

$$exp_z(v) = z \oplus (tanh(\frac{\lambda_z||v||}{2})\frac{v}{||v||}). \quad (3)$$

4) Logarithmic Map in the Poincaré Ball: The logarithmic map serves as the inverse operation to the exponential map. Given a point $y \in \mathbb{D}^n$ on the manifold, the goal is to determine the tangent vector $v \in T_z\mathbb{D}^n$ such that the geodesic

starting at $z \in \mathbb{D}^n$ and moving in the direction of v passes through y. Mathematically, this requirement is formalized as follows:

$$exp_z(v) = y \Rightarrow v = log_z(y). \tag{4}$$

Implementation of HCCLO. Due to the implicit nature of hierarchical structures in feature representations within hyperbolic space, this paper proposes HCCLO to explicitly align and optimize hyperbolic hierarchical features across modalities. While aligning cross-modal hierarchical relationships, HCCLO simultaneously preserves modality-specific discrepancies by leveraging hyperbolic geometric constraints. Since clear positive and negative samples are not readily defined for infrared and visible images, we exploit the intrinsic hierarchical semantics by dividing features into foreground and background regions. For infrared images, foreground salient thermal targets are of primary interest, whereas for visible images, the background vivid texture details are more critical. Inspired by (Liu et al. 2024b), this paper utilizes this hierarchical structure to construct positive and negative samples. Specifically, saliency masks (M) generated by SAM and their complements ($\overline{M}$) delineate foreground and background features, respectively. HCCLO employs two complementary contrastive loss components: the Infrared Branch Contrastive Loss ($\mathcal{L}_{ir}$) and the Visible Branch Contrastive Loss ($\mathcal{L}_{vis}$), facilitating effective hierarchical alignment. The detailed implementation steps of HCCLO are provided in Algorithm 1.

1) Infrared Branch Contrastive Loss: In the infrared (IR) features, highlighted regions correspond to foreground characteristics. However, since the corresponding visible (VIS) images often contain insufficient foreground information in these regions, they can be treated as negative samples relative to the IR features. This paper utilizes the mask M to extract foreground information from both modalities and impose constraints to enhance the discriminability of the IR foreground features. The detailed formulation is as follows:

$$\mathcal{L}_{ir} = max(d_p(I, I^+) - d_p(I, V^+) + \varepsilon, 0), \tag{5}$$

where I, I^+, V^+ denote the IR features, foreground features of IR features and foreground features of VIS features, respectively. ε represents a small constant added for numerical stability.

2) Visible Branch Contrastive Loss: Conversely, visible (VIS) features predominantly capture background patterns, while infrared (IR) features often lack detailed background information. This paper leverages these IR background deficiencies as negative samples to enhance the learning of VIS features. By employing the module $\overline{M}$, we extract cross-modal background correlations and further improve the robustness of visible features through constraint-based optimization. The detailed formulation is as follows:

$$\mathcal{L}_{vis} = max(d_p(V, V^-) - d_p(V, I^-) + \varepsilon, 0), \tag{6}$$

where V, V^-, I^- denote the VIS features, background features of VIS features and foreground features of IR features, respectively. ε is same as $\mathcal{L}_{ir}$.

Algorithm 1: HCCLO algorithm

Input: Training epochs T, Batch size B, IR features I, VIS features V, Saliency Mask M, Exponential map $exp_z(v)$, Infrared branch contrastive loss $\mathcal{L}_{ir}$, Visible branch contrastive loss $\mathcal{L}_{vis}$

Output: Aligned and optimized the hierarchical features of I and V in hyperbolic space

for $t = 1$ to T do **do**
  Let $I^- \leftarrow I * \overline{M}$
  Let $I^+ \leftarrow I * M$
  Let $V^- \leftarrow V * \overline{M}$
  Let $V^+ \leftarrow V * M$
  Let $I_H^-, I_H, I_H^+ \leftarrow exp_z(I^-), exp_z(I), exp_z(I^+)$
  Let $V_H^-, V_H, V_H^+ \leftarrow exp_z(V^-), exp_z(V), exp_z(V^+)$
  for b = 1 to B **do**
    Compute $\mathcal{L}_{ir}$ with I_H^+, I_H, V_H^+
    Compute $\mathcal{L}_{vi}$ with V_H^-, V_H, I_H^-
  end for
end for
return I_H and V_H

Task-adaptive Dual-Space Features Fusion (TDFF)

The TDFF module performs adaptive dual-space feature fusion through a learnable weighting mechanism, as illustrated in the lower-right corner of Fig.2. Given two input feature maps: hyperbolic hierarchical features ($\mathcal{H}_{ir+vi} \in \mathbb{R}^{H\times W\times 128}$) and Euclidean local features ($\mathcal{E}_{ir+vi} \in \mathbb{R}^{H\times W\times 128}$), the module generates task-adaptive fusion weights and produces the fused feature maps: $\mathcal{R}_{ir+vi} \in \mathbb{R}^{H\times W\times 128}$ for registration and $\mathcal{F}_{ir+vi} \in \mathbb{R}^{H\times W\times 128}$ for fusion via element-wise weighted summation.

Channel Compression Layer. The dimensionality reduction stage first compresses the input features. Both $\mathcal{H}_{ir+vi}$ and $\mathcal{E}_{ir+vi}$ pass through a convolutional layer (Conv), followed by a batch normalization layer (BN), and are then activated by a LeakyReLU function. This process is formally expressed as:

$$\mathbf{V}_H = \text{LeakyReLU}(\text{BN}(\text{Conv}_{1\times1}(\mathcal{H}_{ir+vi}))), \mathbf{V}_H \in \mathbb{R}^{\frac{C}{2}\times H\times W} \tag{7}$$

$$\mathbf{V}_E = \text{LeakyReLU}(\text{BN}(\text{Conv}_{1\times1}(\mathcal{E}_{ir+vi}))), \mathbf{V}_E \in \mathbb{R}^{\frac{C}{2}\times H\times W} \tag{8}$$

Adaptive Weight Generation. Fusion weights are predicted through concatenation of the compressed feature maps followed by a nonlinear mapping. The learned weights are generated as:

$$w = \text{Softmax}(\text{Conv}_{1\times1}([\mathbf{V}_H; \mathbf{V}_E])), \quad w \in \mathbb{R}^{2\times H\times W} \tag{9}$$

where w consists of two spatial weight maps satisfying the constraint:

$$\forall i, j \quad w_0^{(i,j)} + w_1^{(i,j)} = 1. \tag{10}$$

Feature Fusion. The final output is obtained by combining the input features using the learned weights w_0 and w_1:

$$\mathbf{F}_{fused} = w_0 \odot \mathcal{H}_{ir+vi} + w_1 \odot \mathcal{E}_{ir+vi}. \tag{11}$$

Method	Public.	TNO				M^3FD				$RoadScene$			
		QNCIE↑	MI ↑	VIF ↑	Q_{abf} ↑	QNCIE↑	MI ↑	VIF ↑	Q_{abf} ↑	QNCIE↑	MI ↑	VIF ↑	Q_{abf} ↑
UMFusion	IJCAI'22	0.8843	2.5894	0.6719	0.4360	0.8881	3.5366	0.6530	0.3915	0.8904	3.211	0.6491	0.4722
SuperFusion	IEEE JAS'22	0.8740	3.1075	0.7376	0.4448	0.8732	3.4986	0.7067	0.4931	0.8835	3.5226	0.7039	0.4797
MURF	TPAMI'23	0.8781	2.4225	0.6069	0.4710	0.8715	2.3800	0.5343	0.4464	0.8832	2.5953	0.5572	0.3738
CDDFuse	CVPR'23	0.8790	2.9976	0.8242	0.5279	0.8796	3.9168	0.8495	0.6331	0.8873	3.1442	0.6952	0.4893
IMF	TCSVT'24	0.8829	2.8478	0.7307	0.4849	0.8845	3.6485	0.7380	0.4288	0.8885	3.5067	0.7043	0.5080
CoCoNet	IJCV'24	0.8789	2.1740	0.6826	0.3081	0.8755	2.6599	0.7360	0.3884	0.8852	2.6726	0.5786	0.3693
MulFS-CAP	TPAMI'25	0.8653	1.8036	0.3141	0.2339	0.8661	2.4116	0.3756	0.4089	0.8707	2.2562	0.3700	0.2925
HMMF (Ours)	-	**0.9011**	**5.8358**	**1.0612**	**0.6216**	**0.8970**	**5.7963**	**0.9898**	**0.7029**	**0.9021**	**5.4088**	**0.9129**	**0.5529**

Table 1: The quantitative evaluation of the IVIF results with State-Of-The-Art Fusion Methods on the TNO, M^3FD and RoadScene datasets. The CrossRAFT algorithm is employed as the baseline registration model for both CDDFuse and CoCoNet.

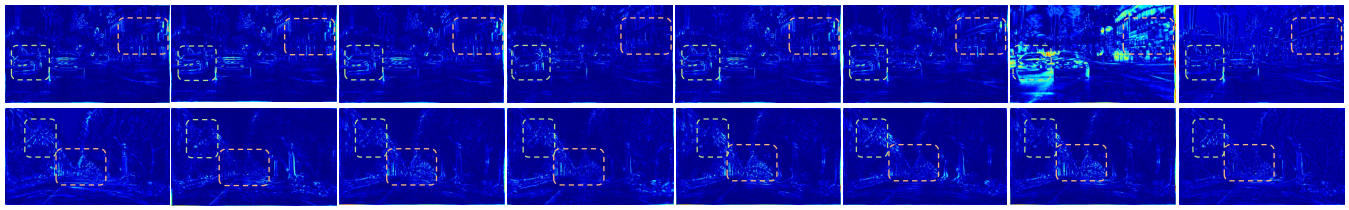


Figure 3: Error visualization of the registration results on the RoadScene dataset.

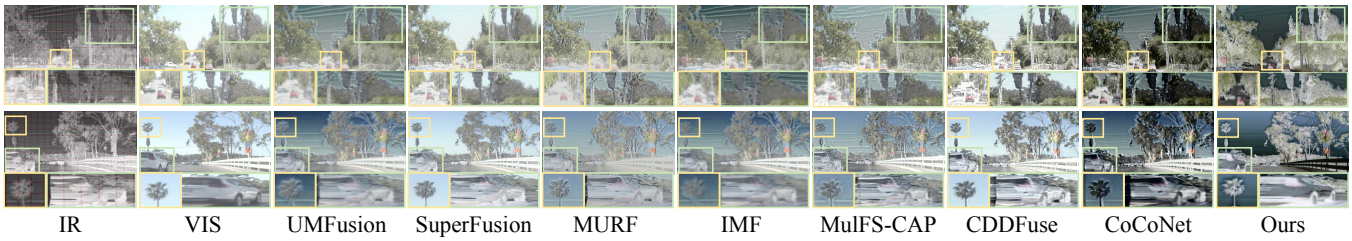


Figure 4: Fusion results visualization of state-of-the-art IR-VIS methods on the RoadScene dataset. CrossRAFT serves as the registration backbone for CDDFuse and CoCoNet.

Experiments

Experimental Setup Details

Dataset. We train and validate the proposed HMMF on misaligned datasets to simulate real-world operating conditions. Infrared (IR) sensors are prone to interference from internal temperature fluctuations and external hot airflow, resulting in misaligned IR images with non-rigid geometric distortions. We simulate non-rigid misalignment by applying affine and elastic transformations (translation: [0.0, 0.01] px; elastic: σ=32, kernel=97–101).

Implementation Details. Our framework is implemented in PyTorch. We train the model using 128×128 patches randomly sampled from RoadScene, M^3FD, and MSIFT datasets. Testing is conducted on RoadScene, M^3FD, and 24 TNO images. The model is trained for 1200 epochs with Adam optimizer (lr=1e-4, decayed every 300 epochs) on an NVIDIA V100 GPU.

To demonstrate the advantages of our method, we conduct qualitative and quantitative comparisons with several state-of-the-art multimodal registration and infrared-visible image fusion (IVIF) methods: CrossRAFT (Zhou, Tan, and Yan 2022), UMFusion (Wang et al. 2022), SuperFusion (Tang et al. 2022), MURF (Xu, Yuan, and Ma 2023), IMF (Wang et al. 2024), MulFS-CAP (Li et al. 2025), CDDFuse (Zhao et al. 2023b), and CoCoNet (Liu et al. 2024b). Since the latter two methods (CDDFuse and CoCoNet) do not provide integrated registration modules, we employ CrossRAFT as the pre-registration approach, which demonstrates superior registration performance as shown in Table 2. For the registration assessment, we adopt NCC and MI, while QNCIE, MI, VIF, and Q_{abf} are utilized for evaluating fusion performance. In all cases, higher metric values correspond to better results.

Comparison Results

Comparison of Registration Results. As shown in Table 2, our method achieves the highest scores in the NCC metric compared to other cross-modality registration methods. The registration error maps, visualized in Fig.3, illustrate the discrepancies between IR images registered by different methods and the target IR images. Through comparative analysis, our method consistently demonstrates superior registration accuracy.

Comparison of Fusion Results. As demonstrated in Table 1, our HMMF consistently ranks at the top across all

Method	Public.	*TNO*		*RoadScene*	
		NCC ↑	MI ↑	NCC ↑	MI ↑
CrossRAFT	AAAI'22	0.8839	1.9862	0.9246	2.2459
UMFusion	IJCAI'22	0.9263	2.2116	0.9316	2.278
SuperFusion	IEEE JAS'22	0.8564	1.7494	0.9346	**2.3572**
MURF	TPAMI'23	0.9104	2.1874	0.9231	2.2747
IMF	TCSVT'24	0.9439	**2.2456**	0.9445	2.2747
MulFS-CAP	TPAMI'25	0.9096	2.2012	0.9248	2.2930
HMMF(Ours)	-	**0.9440**	2.1166	**0.9538**	2.2952

Table 2: Quantitative evaluation of multi-modality image registration on the TNO and RoadScene datasets.

H/E	HCCLO	TDFF	*RoadScene*			
			$NCC\uparrow$	$MI\uparrow$	$QNCIE$	$Q_{abf}\uparrow$
H	×	×	0.9048	1.9615	0.8864	0.2251
H	✓	×	0.9175	2.0672	0.8868	0.2278
E	×	×	0.9020	1.8750	0.8865	0.2363
H+E	✓	✓	**0.9491**	**2.1767**	**0.9028**	**0.5318**

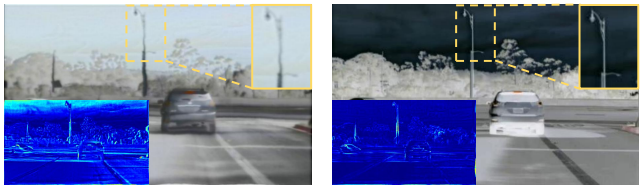
Table 3: Ablation study on the RoadScene dataset. H, Hyperbolic representations; E, Euclidean representations; H+E, dual-space representations;

four evaluation metrics, confirming its superior performance in fusing misaligned multimodal images. Notably, HMMF effectively eliminates ghosting artifacts caused by misalignment while preserving richer background textural details, resulting in fused images that better align with human visual perception.

Ablation Analysis

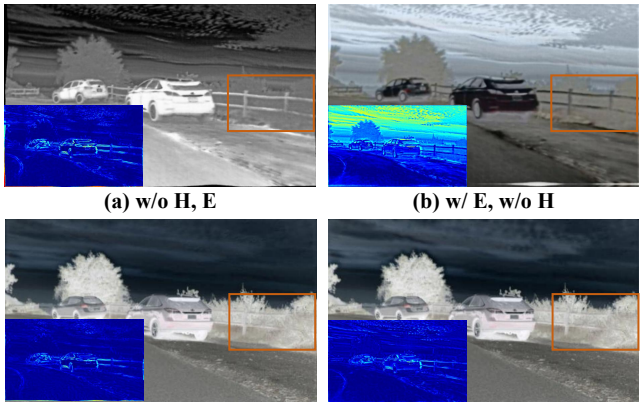
Effectiveness of HCCLO. We conduct ablation studies on the hyperbolic embeddings generated by HCCLO to validate its geometric effectiveness. As shown in Table 3, HCCLO consistently improves performance across all evaluation metrics. Qualitatively, it significantly reduces ghosting artifacts and enhances registration accuracy by better preserving structural continuity, as illustrated in Fig. 5.

Effectiveness of TDFF. Our ablation study compares two feature representations: Euclidean space features (E) and hyperbolic space features (H), along with a combined approach (E+H) where dynamic fusion weights are applied. As shown in Table 3, hyperbolic representation learning effectively captures intrinsic hierarchical relationships across modalities, while Euclidean space preserves high-frequency local details more efficiently. This paper further evaluates the effectiveness of the TDFF module by disabling its dynamic weighting mechanism. As illustrated in Fig.6, the TDFF module successfully integrates Euclidean local detail features with hyperbolic hierarchical features. Quantitative results in Table 3 demonstrate that our dual-space representation consistently outperforms the single-space baselines across all metrics, highlighting the synergistic benefits of TDFF.



(a) w/o HCCLO **(b) w/ HCCLO(Ours)**

Figure 5: Ablation analysis of the HCCLO.



(a) w/o H, E **(b) w/ E, w/o H**

(c) w/o E, w/ H **(d) w/ E, w/ H(TDFF)**

Figure 6: Ablation analysis of the TDFF.

Conclusion

This paper proposes a novel dual-space framework that jointly optimizes multi-modality image registration and fusion. The hyperbolic space leverages its powerful semantic hierarchy modeling capability to effectively capture hierarchical image features, thereby enhancing both registration accuracy and fusion quality. Concurrently, the framework exploits Euclidean local detail features for geometric precision alongside hyperbolic hierarchical features for semantic abstraction, adaptively balancing the distinct demands of registration and fusion tasks. To validate the robustness of the trained model, the deformation intensity of infrared images was intentionally increased, and its performance was rigorously tested under extreme geometric distortions. Future work will explore deeper integration of the relationships between registration and fusion, extending their application to downstream vision tasks.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No.62450072, U22B2052, 624B2033), the Distinguished Youth Funds of the Liaoning Natural Science Foundation (No.2025JH6/101100001), the Distinguished Young Scholars Funds of Dalian (No.2024RJ002), and the Fundamental Research Funds for the Central Universities.

References

Atigh, M. G.; Schoep, J.; Acar, E.; Van Noord, N.; and Mettes, P. 2022. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4453–4462.

De Sa, C.; Gu, A.; Ré, C.; and Sala, F. 2018. Representation Tradeoffs for Hyperbolic Embeddings. *Proceedings of machine learning research*, 80: 4460—4469.

Fu, Y.; Xu, T.; Wu, X.; and Kittler, J. 2022. PPT Fusion: Pyramid Patch Transformerfor a Case Study in Image Fusion. arXiv:2107.13967.

Ganea, O.; Bécigneul, G.; and Hofmann, T. 2018. Hyperbolic neural networks. *Advances in neural information processing systems*, 31.

Gao, C.; Gu, D.; Zhang, F.; and Yu, Y. 2019. ReCoNet: Real-Time Coherent Video Style Transfer Network. In Jawahar, C.; Li, H.; Mori, G.; and Schindler, K., eds., *Computer Vision – ACCV 2018*, 637–653. Cham: Springer International Publishing. ISBN 978-3-030-20876-9.

Huang, Z.; Lv, C.; Xing, Y.; and Wu, J. 2020. Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. *IEEE Sensors Journal*, 21(10): 11781–11790.

Kong, F.; Chen, Y.; Cai, J.; and Modolo, D. 2024. Hyperbolic Learning with Synthetic Captions for Open-World Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16762–16771.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, H.; Wu, X.-J.; and Kittler, J. 2020. MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29: 4733–4746.

Li, H.; Wu, X.-J.; and Kittler, J. 2021. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73: 72–86.

Li, H.; Yang, Z.; Zhang, Y.; Jia, W.; Yu, Z.; and Liu, Y. 2025. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, J.; Chen, J.; Liu, J.; and Ma, H. 2023a. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM international conference on multimedia*, 4471–4479.

Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; and Kasabov, N. K. 2023b. Learning a Coordinated Network for Detail-Refinement Multiexposure Image Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 713–727.

Li, J.; Zhu, J.; Li, C.; Chen, X.; and Yang, B. 2022. CGTF: Convolution-guided transformer for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–14.

Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, 270–288. Springer.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022a. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.

Liu, J.; Fan, X.; Jiang, J.; Liu, R.; and Luo, Z. 2021. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119.

Liu, J.; Li, X.; Wang, Z.; Jiang, Z.; Zhong, W.; Fan, W.; and Xu, B. 2024a. PromptFusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*.

Liu, J.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; and Fan, X. 2024b. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5): 1748–1775.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023a. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8081–8090.

Liu, J.; Zhang, B.; Mei, Q.; Li, X.; Zou, Y.; Jiang, Z.; Ma, L.; Liu, R.; and Fan, X. 2025. DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and Visible Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2226–2235.

Liu, R.; Gao, J.; Liu, X.; and Fan, X. 2024c. Learning with constraint learning: New perspective, solution strategy and various applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 5026–5043.

Liu, R.; Liu, X.; Zeng, S.; Zhang, J.; and Zhang, Y. 2023b. Hierarchical optimization-derived learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14693–14708.

Liu, R.; Liu, Z.; Liu, J.; Fan, X.; and Luo, Z. 2024d. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6594–6609.

Liu, R.; Ma, L.; Ma, T.; Fan, X.; and Luo, Z. 2022b. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5953–5969.

Liu, Y.; Chen, X.; Ward, R. K.; and Wang, Z. J. 2016. Image fusion with convolutional sparse representation. *IEEE signal processing letters*, 23(12): 1882–1886.

Lu, X.; Zhang, B.; Zhao, Y.; Liu, H.; and Pei, H. 2014. The infrared and visible image fusion algorithm based on target separation and sparse representation. *Infrared Physics & Technology*, 67: 397–407.

Ma, J.; Chen, C.; Li, C.; and Huang, J. 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31: 100–109.

Ma, J.; Tang, L.; Xu, M.; Zhang, H.; and Xiao, G. 2021. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–13.

Ma, J.; Xu, H.; Jiang, J.; Mei, X.; and Zhang, X.-P. 2020. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Transactions on Image Processing*, 29: 4980–4995.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48: 11–26.

Ma, J.; Zhou, Z.; Wang, B.; and Zong, H. 2017. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82: 8–17.

Nickel, M.; and Kiela, D. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.

Ramasinghe, S.; Shevchenko, V.; Avraham, G.; and Thalaiyasingam, A. 2024. Accept the Modality Gap: An Exploration in the Hyperbolic Space. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27253–27262.

Tang, H.; Liu, H.; Xu, D.; Torr, P. H. S.; and Sebe, N. 2023. AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4): 1972–1987.

Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; and Ma, J. 2022. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12): 2121–2137.

Wang, D.; Liu, J.; Fan, X.; and Liu, R. 2022. Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 3508–3515. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wang, D.; Liu, J.; Ma, L.; Liu, R.; and Fan, X. 2024. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, S.; Kang, Q.; She, R.; Wang, W.; Zhao, K.; Song, Y.; and Tay, W. P. 2023. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5176–5185.

Xu, H.; Ma, J.; Yuan, J.; Le, Z.; and Liu, W. 2022. RFNet: Unsupervised Network for Mutually Reinforcing Multi-modal Image Registration and Fusion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19647–19656.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10): 12148–12166.

Zhang, C.; and Gao, J. 2021. Hype-han: Hyperbolic hierarchical attention network for semantic embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3990–3996.

Zhang, H.; Yuan, J.; Tian, X.; and Ma, J. 2021. GAN-FM: Infrared and Visible Image Fusion Using GAN With Full-Scale Skip Connection and Dual Markovian Discriminators. *IEEE Transactions on Computational Imaging*, 7: 1134–1147.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118.

Zhao, W.; Xie, S.; Zhao, F.; He, Y.; and Lu, H. 2023a. MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding from Object Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13955–13965.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023b. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5906–5916.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024a. Equivariant Multi-Modality Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25912–25921.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023c. DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8082–8093.

Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; and Gool, L. V. 2024b. Image Fusion via Vision-Language Model. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhao, Z.; Xu, S.; Zhang, C.; Liu, J.; Zhang, J.; and Li, P. 2021. DIDFuse: deep image decomposition for infrared and visible image fusion. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20. ISBN 9780999241165.

Zhou, H.; Wu, W.; Zhang, Y.; Ma, J.; and Ling, H. 2023. Semantic-Supervised Infrared and Visible Image Fusion Via a Dual-Discriminator Generative Adversarial Network. *IEEE Transactions on Multimedia*, 25: 635–648.

Zhou, S.; Tan, W.; and Yan, B. 2022. Promoting single-modal optical flow network for diverse cross-modal flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3562–3570.