Text-Guided Channel Perturbation and Pre-Trained Knowledge Integration for Unified Multi-Modality Image Fusion

Xilai Li, Xiaosong Li*, Weijun Jiang

College of Physics and Optoelectronic Engineering, Foshan University, Foshan 528225, China
20210300236@stu.fosu.edu.cn, lixiaosong@buaa.edu.cn, 20230300130@stu.fosu.edu.cn

Abstract

Multi-modality image fusion enhances scene perception by combining complementary information. Unified models aim to share parameters across modalities for multi-modality image fusion, but large modality differences often cause gradient conflicts, limiting performance. Some methods introduce modality-specific encoders to enhance feature perception and improve fusion quality. However, this strategy reduces generalisation across different fusion tasks. To overcome this limitation, we propose a **U**nified multi-modality image fusion framework based on Channel **P**erturbation and Pre-trained Knowledge Integration (**UP-Fusion**). To suppress redundant modal information and emphasize key features, we propose the Semantic-Aware Channel Pruning Module (SCPM), which leverages the semantic perception capability of a pre-trained model to filter and enhance multi-modality feature channels. Furthermore, we proposed the Geometric Affine Modulation Module (GAM), which uses original modal features to apply affine transformations on initial fusion features to maintain the feature encoder modal discriminability. Finally, we apply a Text-Guided Channel Perturbation Module (TCPM) during decoding to reshape the channel distribution, reducing the dependence on modality-specific channels. Extensive experiments demonstrate that the proposed algorithm outperforms existing methods on both multi-modality image fusion and downstream tasks.

Code — https://github.com/ixilai/UP-Fusion

Introduction

Multi-modality image fusion (MMIF) aims to combine complementary features from different imaging modalities to produce more comprehensive and informative representations (Zhang et al. 2021; Jie et al. 2025). For example, infrared and visible image fusion (IVIF) (Li et al. 2024a; Zhao et al. 2023a; Liu et al. 2024c; Wang et al. 2025b,a) merges fine texture details from visible light with thermal radiation information from infrared imaging, improving the robustness of machine vision under complex lighting conditions. Similarly, medical image fusion (MEIF) (Xu et al. 2024; James and Dasarathy 2014) integrates data from Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single Photon Emission Computed Tomography (SPECT), enabling a more complete understanding of pathological features. Beyond these applications, MMIF also shows strong potential in downstream tasks such as semantic segmentation and object detection.

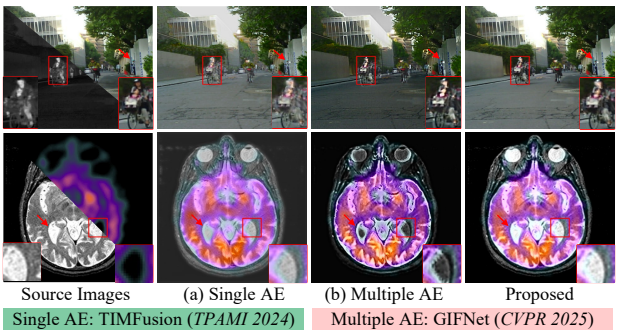

Figure 1: Comparison of the Single AE algorithm (Liu et al. 2024d), Multiple AE algorithm (Cheng et al. 2025), and the proposed method across different MMIF tasks.

Recently, unified MMIF models have attracted increasing attention. A typical approach within this trend is based on a unified autoencoder (AE) structure (Cheng, Xu, and Wu 2023; Yang et al. 2025; Zhu et al. 2024). In such methods, pre-fused features are typically obtained by channel-level concatenation or direct summation of multi-modal inputs, which are then passed through a shared encoder–decoder network for optimization. Alternatively, shared weights are employed to extract features from different modalities simultaneously. Some methods enhance adaptability to MMIF tasks by incorporating strategies such as continual learning (Xu et al. 2020) and meta-learning (Liu et al. 2024d). The single-encoder structure allows models to share feature extractors and decoder parameters across modalities, leading to strong cross-task generalisation. However, these approaches often lack explicit modeling of inter-modal interaction mechanisms, which limits their fusion performance. As shown in Fig. 1, although the single-AE method can handle multiple MMIF tasks, its fusion quality remains subop-

*Corresponding author

timal. The second class of methods adopts modality-specific encoder architectures (Liu et al. 2024b; Bai et al. 2025; Zhao et al. 2024). These approaches design separate encoders for each modality to extract modality-specific representations, which are then concatenated or fused before being processed by a unified decoder to generate the final fused image. This strategy effectively preserves the distinct feature expressiveness of each modality during the encoding stage. In addition, the method (Li et al. 2023a) introduces inter-modal interaction modules between encoders to enable cross-modal feature guidance, further improving fusion performance. However, modality-specific encoders have inherent limitations. Since each encoder tends to overfit the feature distribution of its corresponding modality during training, the model often suffers from poor generalization when applied to unseen modal combinations or datasets with different distributions. Consequently, many MMIF methods require training separate weights for different tasks rather than adopting a unified set of parameters. As shown in Fig. 1, when a multi-AE algorithm trained on IVIF data is applied to the MEIF task, it tends to introduce erroneous background details instead of extracting salient and informative features.

To balance the generalization ability of a single AE with the superior fusion quality of modality-specific AEs, we propose a unified MMIF framework based on text-guided channel perturbation with pre-trained knowledge integration (UP-Fusion). Considering the high redundancy and weak complementarity in multi-modal data, and the tendency for feature extraction to cause substantial channel expansion, we design a Semantic-Aware Channel Pruning Module (SCPM). This module leverages a Squeeze-and-Excitation (SE) block to model channel response strength and integrates the global semantic awareness of the pre-trained ConvNeXt model, jointly guiding the identification of salient channels. Subsequently, to enhance the representation and fusion of modality-specific features, we propose the Geometric Affine Modulation Module (GAM). This module performs structural modulation of pre-fused features through affine transformations. In the decoding phase, we propose the Text-Guided Channel Perturbation Module (TCPM), which utilises semantic text to guide the channel selection and alignment of multi-modality features after first using channel attention for further filtering, enabling the features to remove modal markers and improving the generalisation ability of the model. Our main contributions are as follows:

- We propose a unified multi-modality image fusion framework based on text-guided channel perturbation and pre-trained knowledge, effectively reducing modal redundancy while enhancing unified modeling and generalization across cross-modal features.
- We propose the Semantic-Aware Channel Pruning Module, which integrates semantic awareness with channel attention to filter redundant channels and strengthen salient features. Additionally, we propose the Geometric Affine Modulation Module to perform affine transformations on fusion representations based on modality-specific features.
- Extensive experiments demonstrate that the proposed method significantly outperforms existing task-specific and unified fusion models in both infrared-visible and medical image fusion tasks, while also exhibiting superior performance in downstream applications.

Related Works

Multi-Modality Image Fusion

Recent advances in MMIF have primarily centered on deep learning-based frameworks (Yue et al. 2023; Liu et al. 2024a; Bai et al. 2024). Mainstream approaches can be broadly divided into two categories: autoencoder (AE)-based methods and generative model-based methods. The first group includes models built upon Convolutional Neural Networks (CNNs) (Xu et al. 2020), Transformers (Ma et al. 2022), Mamba architectures (Xie et al. 2024), and their hybrid variants (Zhao et al. 2023a), whereas the second group primarily relies on Generative Adversarial Networks (GANs) (Liu et al. 2022a) and diffusion models (Zhao et al. 2023b). For example, Liu et al. (2024b) proposed a coupled contrastive learning network for MMIF. This method effectively preserves salient information from different modalities and reduces redundancy by incorporating a contrastive constraint mechanism into the loss function. Zhao et al. (2024) proposed an equivariant MMIF framework that integrates the equivariant prior of natural images into the training process to improve fusion performance. To address the instability of GAN-based methods, Zhao et al. (2023b) proposed an MMIF algorithm using the Denoising Diffusion Probabilistic Model, framing the fusion task as a conditional generation problem.

In addition, some researchers have extended the MMIF technique to downstream tasks (Liu et al. 2025; Li et al. 2023c)(e.g., detection and segmentation) and complex scenarios (Li et al. 2024b; Huang et al. 2024; Yi et al. 2024; Li et al. 2025; Lv et al. 2025) (e.g., adverse weather and low light). Specifically, Wu et al. (2025) distilled semantic knowledge from the Segment Anything Model into the fusion network, leveraging high-level semantic priors to guide the fusion process. Similarly, Bai et al. (2025) proposed a learnable task-guided fusion loss, enabling the fusion results to adapt effectively to various downstream tasks. Meanwhile, some studies focus on improving the generalisation and applicability of fusion methods in complex scenes. For example, Li et al. (2024b) proposed a MMIF method for adverse weather by designing an integrated restoration–fusion model, enabling feature restoration and interaction simultaneously. Yi et al. (2024) introduced a degradation-aware interactive fusion network that leverages text guidance to enhance fusion performance.

Despite recent progress in MMIF and its application in downstream tasks and complex scenarios, the design of unified model architectures remains overlooked. Existing methods typically rely on learning strategies to unify weights, such as U2Fusion (Xu et al. 2020) with continual learning and GIFNet (Cheng et al. 2025) with task-specific constraints. **In contrast, we explore the construction of unified models from the perspectives of single and modality-specific AEs, offering new insights to advance the field.**

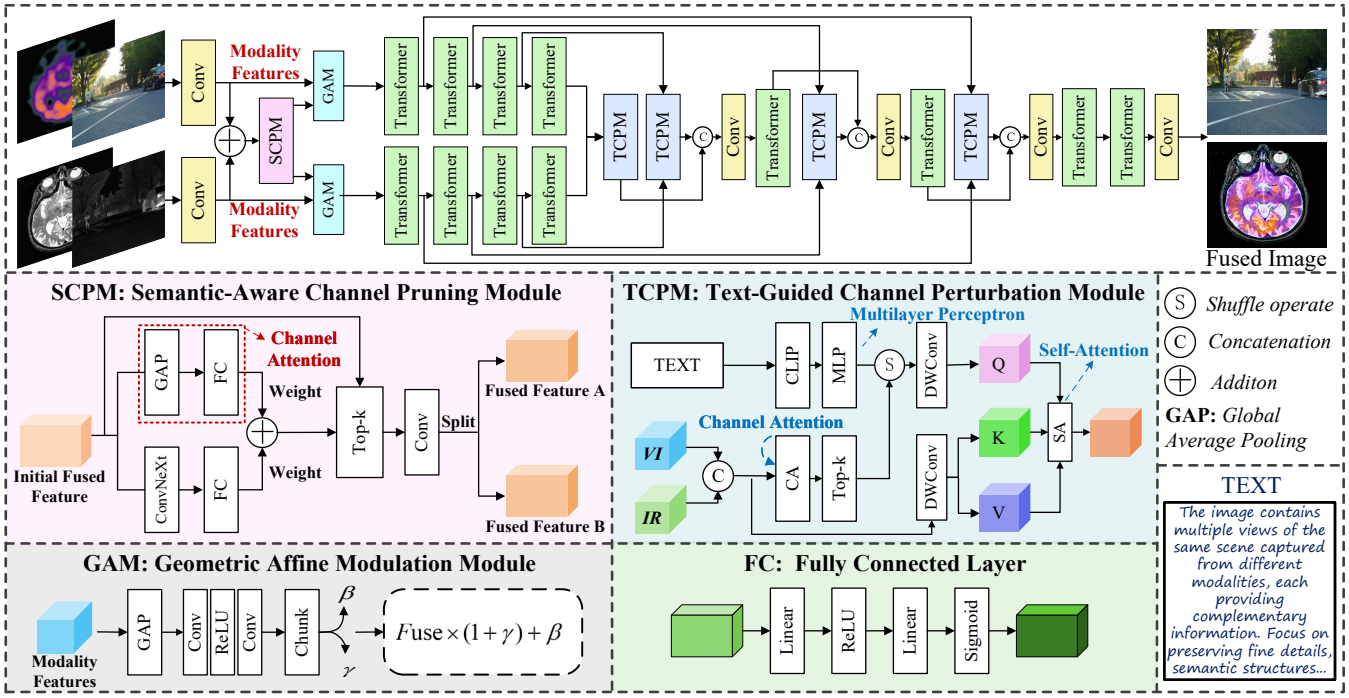

Figure 2: The overall framework of the proposed unified multi-modality image fusion framework.

Proposed Method

Overview

As shown in Fig. 2, the proposed algorithm adopts an AE-based architecture built upon Transformer blocks (Zamir et al. 2022). The network comprises a 4-layer encoder and a 4-layer decoder, where the number of Transformer blocks in the encoder progressively increases as $[4, 6, 6, 8]$, along with a corresponding rise in attention heads ($[1, 2, 4, 8]$), enabling gradual spatial resolution reduction while expanding channel capacity. In the encoder stage, multi-modal inputs I^A and I^B are first processed through convolutional layers. The proposed SCPM is then applied to remove redundant channels while retaining salient ones, followed by channel expansion to restore the original dimensionality. To incorporate modality-specific cues, we introduce the GAM module, which performs affine transformations on the pre-fused features under the guidance of the original modal representations. The dual-branch encoder extracts the modulated multi-modality features independently and performs channel-level interaction using the proposed TCPM at the deepest layer. The decoder, composed of four Transformer blocks, progressively reconstructs the fused image. At each stage, the corresponding encoder features are integrated and refined through TCPM, then concatenated with the upsampled features. Finally, channel compression via convolution ensures stepwise reconstruction of the fused image.

Semantic-Aware Channel Pruning Module

In MMIF tasks, redundant features often cause modal bias or feature interference. This redundancy suppresses complementary information, ultimately diminishing the semantic expressiveness and visual fidelity of the fused image. To address this, we propose the Semantic-Aware Channel Pruning Module (SCPM), which adaptively preserves the most representative salient features for subsequent encoding by integrating a channel attention mechanism with global semantic guidance.

Specifically, we first compute the channel importance weight ω_C using the channel attention mechanism. Then, the pre-trained ConvNeXt backbone (Liu et al. 2022b) is employed to extract semantic features from the initial fused representation, producing semantic weight ω_S through a linear mapping. The final channel weight ω_F is obtained by weighted fusion of ω_C and ω_S,

$$\omega_F = \omega_C + \alpha \cdot \sigma(\omega_S) \tag{1}$$

where α is a learnable parameter for balancing the weights and σ is the Sigmoid activation function. According to ω_F, SCPM adopts the Top-k strategy to retain the top 70% of the significant channels, and expands back to the original channel dimensions by 1×1 convolution. Finally, the output features are divided into two subsets $Fuse^M$, $M \in [A, B]$ by channel partitioning operation, which are sent to GAM for further processing.

Geometric Affine Modulation Module

Existing bimodal encoder-based approaches typically introduce modality-specific features directly for feature extraction and interaction. Although this strategy improves fusion performance, it tends to produce modality-dependent algorithms, as the encoder overfits to modality-specific features and thus generalizes poorly to unseen MMIF tasks. To over-

come this limitation, we propose the GAM, which modulates fused features from a geometric perspective to mitigate the dependency of modality.

Specifically, we first perform global average pooling on the original multi-modality features to obtain F_G^A and F_G^B, respectively. Subsequently, the mapping is performed by a fully-connected network of two-layer 1×1 convolutions to obtain the scaling factor γ and the offset term β of the affine transformation:

$$[\gamma^M, \beta^M] = \mathrm{Conv}_{1\times1}\big(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(F_G^M))\big) \quad (2)$$

Then, we perform affine modulation of the fused features:

$$F_O^M = Fuse^M \cdot (1+\gamma) + \beta \quad (3)$$

By modulating with two parameters—translation and scaling—GAM enables the fusion features refined by SCPM to adjust their spatial distribution according to the global geometric characteristics of the original modality.

Text-Guided Channel Perturbation Module

After obtaining feature representations of different modalities through a multi-modality encoder, existing methods typically perform a weighted combination and input the result into the decoding module for reconstruction. This approach offers simplicity and fast convergence on modality-specific tasks. However, because this operation is inherently unlearnable, it relies heavily on the modality-specific data distribution during training. Consequently, the decoding process cannot adapt dynamically when the model is applied to different multimodal image fusion tasks. To address this limitation, inspired by Tian et al. (2025), we introduce the concept of channel perturbation into MMIF and propose the TCPM. TCPM uses channel selection and rearrangement, guided by textual semantics, to perturb modal feature distribution during decoding.

Specifically, multi-modality features are first concatenated along the channel dimension, and the importance of each channel is evaluated using a channel attention mechanism. Then, a Top-k strategy selects the top 50% most significant channels. To alleviate gradient disturbance caused by channel rearrangement, a 1×1 convolution is applied to expand the channel number to twice the original, enhancing feature redundancy and expressiveness. Simultaneously, to realize text-guided channel perturbation, CLIP-encoded (Radford et al. 2021) text features are passed through a two-layer linear mapping network to produce bootstrap weights matching the channel dimension. These weights are used to generate a channel rearrangement index (Tian et al. 2025), guiding the fused features to undergo semantically relevant channel rearrangement. To further enhance interaction between perturbed and original features, a self-attention mechanism (Tian et al. 2025) is employed, where the perturbed features serve as the query (Q), and the original features provide the key (K) and value (V). The resulting output is processed through a feed-forward network.

Loss Function

During training, we used two loss functions: gradient loss and L1 loss, to preserve image details and structural information. The gradient loss is defined as:

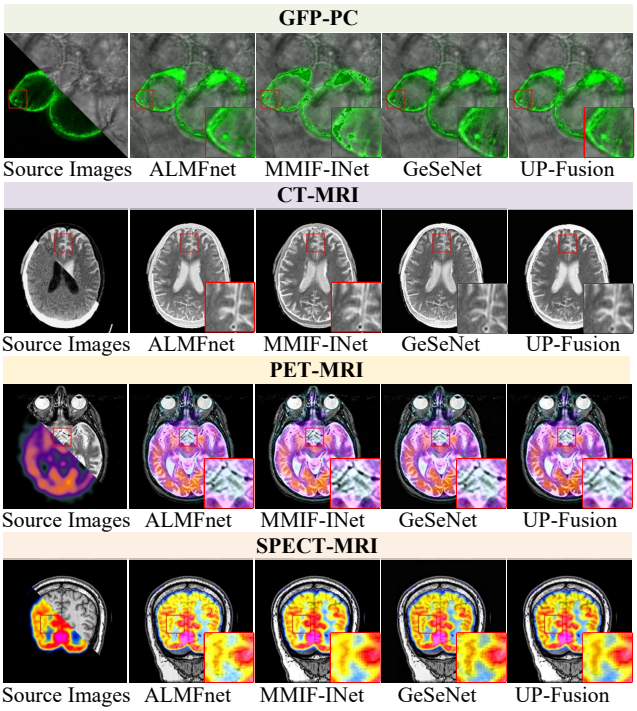

Figure 3: Comparison of the proposed algorithm and medical image fusion methods on medical image tasks.

$$L_{\text{grad}} = \frac{1}{HW}\left\|\nabla Fuse - \max\left(\nabla I_A, \nabla I_B\right)\right\|_1 \quad (4)$$

where H and W are the length and width of the image, respectively, $Fuse$ is the final fusion result, and I_A and I_B are the multi-modality source images, respectively. The L1 loss is calculated as follows.

$$L_{l_1} = \frac{1}{HW}\sum_{M\in\{A,B\}}\left\|Fuse - I_M\right\|_1 \quad (5)$$

The total loss L_{T} can be expressed as $L_{\text{T}} = L_{\text{grad}} + L_{l_1}$

Experiments

Experimental Setup Details

Dataset To evaluate the generality of the proposed method, we conducted experiments on two tasks: infrared-visible image fusion and medical image fusion. For the former, three mainstream datasets were used: MSRS (Tang et al. 2022), LLVIP (Jia et al. 2021), and M3FD (Liu et al. 2022a). For the latter, we used multi-modality brain medical images from Harvard Medical School[1], covering three typical fusion scenarios: CT-MRI, PET-MRI, and SPECT-MRI. In addition, we tested generalisation on a mainstream database[2] containing green fluorescent protein (GFP) and phase-contrast (PC) images.

[1]http://www.med.harvard.edu/aanlib/home.html
[2]http://data.jic.ac.uk/Gfp/

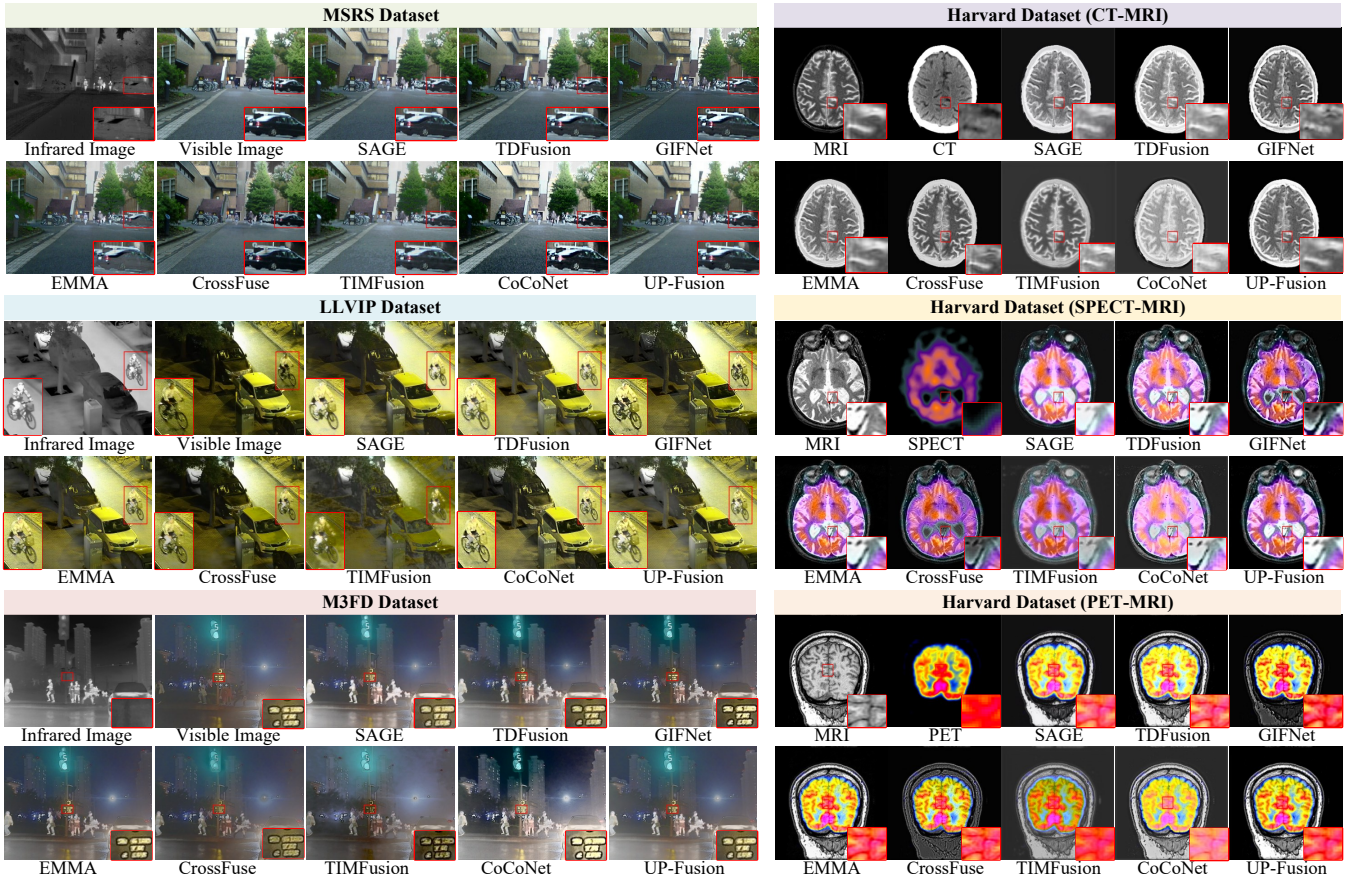

Figure 4: Comparison of different methods on infrared and visible image fusion, and medical image fusion tasks.

Metrics We used five metrics to objectively evaluate fusion performance: Nonlinear Correlation Information Entropy (Q_{NCIE}), Phase Congruency-based Image Fusion Metric (Q_P), Visual Information Fidelity (VIF), Structural Similarity Index Measure ($SSIM$), and Normalized Weighted Performance Metric ($Q^{AB/F}$) (Liu et al. 2011).

Comparison methods We compared with 10 state-of-the-art MMIF methods, including three IVIF methods (CrossFuse (Li and Wu 2024), SAGE (Wu et al. 2025), TDFusion (Bai et al. 2025)), three MEIF methods (ALMFNet (Mu et al. 2023), MMIF-INet (He et al. 2025), GeSeNet (Li et al. 2023b)), and four general MMIF methods (GIFNet (Cheng et al. 2025), EMMA (Zhao et al. 2024), TIMFusion (Liu et al. 2024d), CoCoNet (Liu et al. 2024b)).

Training Details The proposed method is trained only on the LLVIP dataset and uses a single weight for all MMIF tasks. We conducted training for a total of 100 epochs. The images were randomly cropped to a size of 192×192 pixels, with a batch size of 2. The initial learning rate was set to 0.0001 and gradually decayed to 0.00001 using a cosine decay schedule. The optimization was performed using the Adam optimizer. All experiments are conducted in a PyTorch 2.1.1 environment on a server equipped with four GeForce RTX 3090 GPUs.

Qualitative Comparison

Fig. 3 shows the fusion results of the proposed algorithm (UP-Fusion) compared with existing MEIF methods on brain images from Harvard Medical School and GFP-PC images. Although trained only on IVIF data, our method achieves excellent performance on medical image tasks, accurately extracting bone structures in CT images and detailed textures in MRI images. It also outperforms medical-specific fusion methods in color fidelity. These results indicate that the proposed channel perturbation mechanism effectively reduces modality dependence and significantly enhances the cross-task ability of the proposed method.

Fig. 4 presents the visual results of the proposed algorithm compared with three IVIF methods and four general MMIF algorithms on both IVIF and MEIF tasks. Methods such as SAGE, TDFusion, GIFNet, EMMA, CrossFuse, and CoCoNet, like ours, are trained only on infrared-visible images and generally perform well on IVIF tasks. However, when applied to medical image fusion, SAGE and CoCoNet lose detail, while GIFNet and CrossFuse introduce excessive black background regions, degrading fusion quality. Although TDFusion and EMMA remain stable on medical images, they exhibit contrast loss on IVIF tasks. In contrast, the proposed method achieves excellent performance across multiple IVIF datasets and medical fusion tasks.

Datasets		MSRS					LLVIP				
Methods	Pub.	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$	$Q^{AB/F}\uparrow$	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$	$Q^{AB/F}\uparrow$
SAGE	*CVPR 25*	0.8123	0.5210	0.4359	0.4691	0.6242	0.8065	0.4128	0.3590	0.4223	0.5631
TDFusion	*CVPR 25*	0.8159	0.5529	0.4257	0.5010	0.6770	0.8082	0.4406	0.3577	0.4485	0.6252
GIFNet	*CVPR 25*	0.8048	0.3314	0.2781	0.4192	0.4542	0.8058	0.3399	0.2126	0.3855	0.4515
EMMA	*CVPR 24*	0.8110	0.4407	0.3689	0.4613	0.5947	0.8075	0.4183	0.3301	0.4339	0.5969
CrossFuse	*INF 24*	0.8142	0.3564	0.3246	0.3931	0.4943	0.8129	0.4095	0.3398	0.4396	0.6246
TIMFusion	*PAMI 24*	0.8051	0.1913	0.1834	0.2834	0.4078	0.8042	0.1361	0.2018	0.2629	0.2754
CoCoNet	*IJCV 24*	0.8061	0.2925	0.1669	0.2987	0.3547	0.8062	0.4007	0.2434	0.3799	0.5081
UP-Fusion	w/o	0.8167	0.5671	0.4587	0.5074	0.6859	0.8092	0.5042	0.3817	0.4636	0.6864
Datasets		M3FD					Harvard				
Methods	Pub.	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$	$Q^{AB/F}\uparrow$	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$	$Q^{AB/F}\uparrow$
SAGE	*CVPR 25*	0.8065	0.4813	0.4110	0.4909	0.5974	0.8058	0.3374	0.2263	0.2054	0.4244
TDFusion	*CVPR 25*	0.8065	0.4748	0.4041	0.5231	0.6270	0.8070	0.4961	0.2859	0.4518	0.6410
GIFNet	*CVPR 25*	0.8061	0.4002	0.2918	0.4710	0.5243	0.8054	0.2784	0.1689	0.1942	0.4195
EMMA	*CVPR 24*	0.8109	0.5341	0.3829	0.4787	0.6032	0.8068	0.5154	0.2745	0.3029	0.6489
CrossFuse	*INF 24*	0.8140	0.4122	0.3807	0.4634	0.5844	0.8072	0.3626	0.2390	0.2232	0.5645
TIMFusion	*PAMI 24*	0.8055	0.2926	0.2680	0.3928	0.4912	0.8056	0.3671	0.2299	0.1600	0.3364
CoCoNet	*IJCV 24*	0.8049	0.3242	0.1967	0.3288	0.3447	0.8056	0.3900	0.2068	0.1872	0.4750
UP-Fusion	w/o	0.8119	0.5190	0.4582	0.5163	0.6275	0.8074	0.5665	0.3190	0.3639	0.6814

Table 1: Quantitative comparison of different methods across multiple multi-modality image fusion datasets. The marked red indicates the best score, and the marked blue indicates the second score.

Datasets		GFP-PC					Harvard				
Methods	Pub.	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$	$Q^{AB/F}\uparrow$	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$	$Q^{AB/F}\uparrow$
ALMFnet	*TCSVT 23*	0.8102	0.4813	0.3077	0.3140	0.5839	0.8068	0.5434	0.3003	0.5777	0.6780
GeSeNet	*TNNLS 23*	0.8146	0.4740	0.3049	0.3202	0.6044	0.8066	0.5576	0.2900	0.5425	0.7058
MMIF-INet	*INF 25*	0.8078	0.3274	0.2448	0.2966	0.4949	0.8071	0.5591	0.2958	0.2599	0.7155
UP-Fusion	w/o	0.8170	0.4826	0.3193	0.3339	0.5914	0.8074	0.5665	0.3190	0.3639	0.6814

Table 2: Quantitative comparison of medical image fusion methods on a medical image dataset. The marked red indicates the best score, and the marked blue indicates the second score.

Methods	$Q_{NCIE}\uparrow$	$Q_P\uparrow$	$VIF\uparrow$	$SSIM\uparrow$
w/o SCPM	0.8052	0.5343	0.3046	0.2645
w/o GAM	0.8071	0.5488	0.3151	0.3478
w/o TCPM	0.8054	0.5221	0.3016	0.2824
w/o CA (TCPM)	0.8068	0.5057	0.2983	0.2632
w/o ConvNeXt	0.8072	0.5478	0.3018	0.2807
w/o CA (SCPM)	0.8071	0.5441	0.3052	0.3218
UP-Fusion	0.8074	0.5665	0.3190	0.3639

Table 3: Quantitative comparison of different ablation results. The marked red indicates the best score.

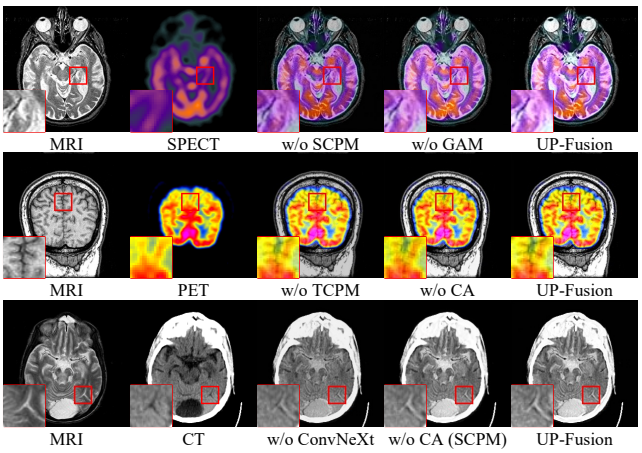

Figure 5: Different ablation results on medical image fusion.

Quantitative Comparison

Tabs. 1 and 2 present the quantitative comparison results of the proposed algorithm on IVIF and MEIF tasks, respectively. Table 1 evaluates the performance of the proposed method against IVIF and general MMIF algorithms on IVIF datasets and examines its generalisation to MEIF tasks. Table 2 compares the proposed algorithm with MEIF-specific methods on MEIF tasks, including GFP-PC images.

As shown in Tab. 1, the proposed algorithm achieves the optimum in multiple metrics on the MSRS, LLVIP, and M3FD datasets. In the MEIF task, it also ranks first in four metrics and second in one, showing superior cross-modal generalisation ability. Tab. 2 further shows that the proposed algorithm outperforms fusion methods designed for medical images in several metrics.

Ablation Studies

Ablation studies were performed to evaluate each module, with experiments carried out primarily on the MEIF task.

Methods	Pub.	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color cone	Bump	mIoU
SAGE	*CVPR 25*	98.38	88.8	69.53	70.37	57.13	71.21	86.06	65.08	76.9	75.94
TDFusion	*CVPR 25*	98.51	90.37	74.43	71.9	65.02	74.39	83.87	65.94	78.85	78.14
GIFNet	*CVPR 25*	98.4	88.98	71.13	71.92	59.19	74.11	78.05	65.3	64.62	74.63
EMMA	*CVPR 24*	98.53	90.36	74.65	71.61	64.61	74.49	83.73	65.61	76.36	77.77
CrossFuse	*INF 24*	98.11	86.89	65.17	69.65	49.31	69.85	75.76	65.56	59.36	70.74
TIMFuison	*PAMI 24*	97.76	82.87	66.34	60.74	40.07	64.48	77.93	54.77	55.1	66.67
CoCoNet	*IJCV 24*	98.12	85.83	68.66	70.47	52.57	70.05	74.03	60.89	53.9	70.5
UP-Fusion	w/o	98.56	90.49	73.47	71.94	64.02	74.05	85.63	66.1	80.22	78.28

Table 4: Quantitative comparison of segmentation accuracy for different algorithms on the MSRS dataset.

Methods	Pub.	People	Car	Bus	Lamp	Motorcycle	Truck	mAP@0.5	mAP@[0.5:0.95]
SAGE	*CVPR 25*	0.823	0.918	0.923	0.817	0.704	0.832	0.836	0.532
TDFusion	*CVPR 25*	0.831	0.919	0.919	0.806	0.715	0.826	0.836	0.535
GIFNet	*CVPR 25*	0.805	0.907	0.919	0.785	0.685	0.783	0.814	0.506
EMMA	*CVPR 24*	0.819	0.9	0.891	0.727	0.649	0.787	0.796	0.498
CrossFuse	*INF 24*	0.716	0.907	0.901	0.792	0.699	0.807	0.804	0.503
TIMFuison	*PAMI 24*	0.733	0.906	0.907	0.783	0.656	0.779	0.794	0.499
CoCoNet	*IJCV 24*	0.764	0.899	0.913	0.748	0.664	0.743	0.789	0.485
UP-Fusion	w/o	0.825	0.922	0.927	0.823	0.728	0.819	0.841	0.541

Table 5: Quantitative comparison of detection accuracy for different algorithms on the M3FD dataset.

Analysis of SCPM SCPM suppresses redundant features and reduces modality dependence in the encoder. As shown in Fig. 5, removing SCPM markedly degrades contrast and fine structural details in medical images. Correspondingly, information and structure–preservation metrics (e.g., Q_{NCIE} and $SSIM$; Table 3) drop noticeably, underscoring SCPM's role in improving fusion quality.

Analysis of GAM To prevent the direct introduction of modality-specific information, GAM modulates the fused features after SCPM processing from a geometric perspective. Ablation experiments show that removing GAM results in fused features lacking modality guidance, leading to insufficient cross-modal interaction. As shown in Fig. 5, although the model retains some generalisation ability, black background information in SPECT is incorrectly fused. Similarly, the metrics in Tab. 3 also decrease.

Analysis of TCPM TCPM introduces channel perturbation during the decoding stage to help the model adapt to dynamic feature variations. As shown in Fig. 5 and Tab. 3, removing TCPM leads to a significant decline in performance on medical images, with fusion results incorporating incorrect modality information, thereby reducing overall quality.

Analysis of Channel Attention We introduce the channel attention module (CA) in both SCPM and TCPM. In SCPM, CA is used to compress channels and extract salient features to construct the initial fusion results, while in TCPM, it filters out low-significance channels to prevent redundancy from affecting subsequent reconstruction. The ablation results in Fig. 5 and Tab. 3 show that the removal of CA leads to texture blurring, and the detail expression is weakened.

Analysis of Pre-Trained Model In SCPM, we introduce the pre-trained ConvNeXt model to provide global semantic guidance. Ablation experiments show that removing ConvNeXt significantly reduces cross-task performance. This is because the CA module in SCPM relies on training data from a specific modality, and without the guidance of pre-trained knowledge, it may incorrectly select channels when generalising to unseen tasks.

Downstream Task Experiments

We conducted experiments on two downstream tasks: object detection and semantic segmentation. Specifically, BANet (Peng et al. 2021) was employed for semantic segmentation on the MSRS dataset, and YOLOv7 (Wang, Bochkovskiy, and Liao 2023) was used for object detection on the M3FD dataset. All fusion methods were evaluated under the same experimental settings. As shown in Tab. 4, the proposed algorithm achieves the best segmentation accuracy and mIoU across multiple categories. Tab. 5 further shows that our method also leads in two key detection metrics, mAP@0.5 and mAP@[0.5:0.95], confirming its superior performance.

Conclusion

In this paper, we propose a multi-modality image fusion method based on channel perturbation and pre-training knowledge. Specifically, we design the SCPM to preserve salient features by combining channel attention with global semantic guidance. The GAM is introduced to mitigate the loss of generalisation caused by directly injecting modality information into the encoder. Finally, the TCPM is applied in the decoding stage to improve adaptability by rearranging feature channels. Extensive experiments on fusion tasks and downstream applications validate the effectiveness of the proposed method.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 62201149), the Basic and Applied Basic Research of Guangdong Province (No. 2023A1515140077), the Natural Science Foundation of Guangdong Province (No.2024A1515011880), the Research Fund of Guangdong-Hong Kong-Macao Joint Laboratory for Intelligent Micro-Nano Optoelectronic Technology (No. 2020B1212030010).

References

Bai, H.; Zhang, J.; Zhao, Z.; Wu, Y.; Deng, L.; Cui, Y.; Feng, T.; and Xu, S. 2025. Task-driven Image Fusion with Learnable Fusion Loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7457–7468.

Bai, H.; Zhao, Z.; Zhang, J.; Wu, Y.; Deng, L.; Cui, Y.; Jiang, B.; and Xu, S. 2024. ReFusion: Learning Image Fusion from Reconstruction with Learnable Loss Via Meta-Learning. *International Journal of Computer Vision*, 1–21.

Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Tang, Z.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; and Kittler, J. 2025. One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28102–28112.

Cheng, C.; Xu, T.; and Wu, X.-J. 2023. MUFusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92: 80–92.

He, D.; Li, W.; Wang, G.; Huang, Y.; and Liu, S. 2025. MMIF-INet: Multimodal medical image fusion by invertible network. *Information Fusion*, 114: 102666.

Huang, J.; Li, X.; Tan, H.; Yang, L.; Wang, G.; and Yi, P. 2024. DeDNet: Infrared and visible image fusion with noise removal by decomposition-driven network. *Measurement*, 237: 115092.

James, A. P.; and Dasarathy, B. V. 2014. Medical image fusion: A survey of the state of the art. *Information fusion*, 19: 4–19.

Jia, X.; Zhu, C.; Li, M.; Tang, W.; and Zhou, W. 2021. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3496–3504.

Jie, Y.; Xu, Y.; Li, X.; Zhou, F.; Lv, J.; and Li, H. 2025. FS-Diff: Semantic guidance and clarity-aware simultaneous multimodal image fusion and super-resolution. *Information Fusion*, 121: 103146.

Li, H.; and Wu, X.-J. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103: 102147.

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. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, X.; Li, X.; Tan, T.; Li, H.; and Ye, T. 2025. UMCFuse: A Unified Multiple Complex Scenes Infrared and Visible Image Fusion Framework. *IEEE Transactions on Image Processing*.

Li, X.; Li, X.; Ye, T.; Cheng, X.; Liu, W.; and Tan, H. 2024a. Bridging the gap between multi-focus and multi-modal: a focused integration framework for multi-modal image fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1628–1637.

Li, X.; Liu, W.; Li, X.; Zhou, F.; Li, H.; and Nie, F. 2024b. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *arXiv preprint arXiv:2402.02090*.

Li, X.; Zou, Y.; Liu, J.; Jiang, Z.; Ma, L.; Fan, X.; and Liu, R. 2023c. From text to pixels: a context-aware semantic synergy solution for infrared and visible image fusion. *arXiv preprint arXiv:2401.00421*.

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.; 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.; Wu, G.; Liu, Z.; Wang, D.; Jiang, Z.; Ma, L.; Zhong, W.; and Fan, X. 2024c. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.; 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, Z.; Blasch, E.; Xue, Z.; Zhao, J.; Laganiere, R.; and Wu, W. 2011. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. *IEEE transactions on pattern analysis and machine intelligence*, 34(1): 94–109.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Lv, G.; Gao, X.; Dong, A.; Wei, Z.; and Cheng, J. 2025. SL-Fusion: a structure-aware infrared and visible image fusion network for low-light scenes. *IEEE Transactions on Circuits and Systems for Video Technology*.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Mu, P.; Wu, G.; Liu, J.; Zhang, Y.; Fan, X.; and Liu, R. 2023. Learning to search a lightweight generalized network for medical image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 5921–5934.

Peng, C.; Tian, T.; Chen, C.; Guo, X.; and Ma, J. 2021. Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation. *Neural Networks*, 137: 188–199.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92.

Tian, X.; Liao, X.; Liu, X.; Li, M.; and Ren, C. 2025. Degradation-Aware Feature Perturbation for All-in-One Image Restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28165–28175.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475.

Wang, Y.; Lin, Y.; He, X.; Zheng, H.; Yan, K.; Fan, L.; Huang, Y.; and Ding, X. 2025a. Learning Diffusion High-Quality Priors for Pan-Sharpening: A Two-Stage Approach With Time-Aware Adapter Fine-Tuning. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, Z.; Zhang, J.; Guan, T.; Zhou, Y.; Li, X.; Dong, M.; and Liu, J. 2025b. Efficient Rectified Flow for Image Fusion. *Advances in Neural Information Processing Systems*.

Wu, G.; Liu, H.; Fu, H.; Peng, Y.; Liu, J.; Fan, X.; and Liu, R. 2025. Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17882–17891.

Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 502–518.

Xu, Y.; Li, X.; Jie, Y.; and Tan, H. 2024. Simultaneous tri-modal medical image fusion and super-resolution using conditional diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 635–645. Springer.

Yang, B.; Jiang, Z.; Pan, D.; Yu, H.; Gui, G.; and Gui, W. 2025. LFDT-Fusion: A latent feature-guided diffusion Transformer model for general image fusion. *Information Fusion*, 113: 102639.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27026–27035.

Yue, J.; Fang, L.; Xia, S.; Deng, Y.; and Ma, J. 2023. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32: 5705–5720.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; and Ma, J. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76: 323–336.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023a. 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. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25912–25921.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023b. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8082–8093.

Zhu, P.; Sun, Y.; Cao, B.; and Hu, Q. 2024. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7099–7108.