

# Multi-Modal Image Fusion via Intervention-Stable Feature Learning

Xue Wang<sup>1,2</sup>, Zheng Guan<sup>1\*</sup>, Wenhua Qian<sup>1\*</sup>, Chengchao Wang,<sup>1</sup> Runzhuo Ma<sup>3</sup>

<sup>1</sup> School of Information Science and Engineering, Yunnan University

<sup>2</sup> School of Artificial Intelligence, Nanyang Normal University

<sup>3</sup> Department of Electrical and Electronic Engineering, Hong Kong Polytechnic University

gz\_627@sina.com    whqian@ynu.edu.cn

## Abstract

Multi-modal image fusion integrates complementary information from different modalities into a unified representation. Current methods predominantly optimize statistical correlations between modalities, often capturing dataset-induced spurious associations that degrade under distribution shifts. In this paper, we propose an intervention-based framework inspired by causal principles to identify robust cross-modal dependencies. Drawing insights from Pearl’s causal hierarchy, we design three principled intervention strategies to probe different aspects of modal relationships: i) complementary masking with spatially disjoint perturbations tests whether modalities can genuinely compensate for each other’s missing information, ii) random masking of identical regions identifies feature subsets that remain informative under partial observability, and iii) modality dropout evaluates the irreplaceable contribution of each modality. Based on these interventions, we introduce a Causal Feature Integrator (CFI) that learns to identify and prioritize intervention-stable features maintaining importance across different perturbation patterns through adaptive invariance gating, thereby capturing robust modal dependencies rather than spurious correlations. Extensive experiments demonstrate that our method achieves SOTA performance on both public benchmarks and downstream high-level vision tasks. The [Code](#) can be available.

## 1. Introduction

Multi-modal image fusion (MMIF) aims to integrate complementary information from different sensing modalities into a unified representation that is more informative and reliable than any individual modality alone [11, 18, 19, 30, 40, 54]. In infrared and visible image fusion (IVIF), a sub-task of MMIF, texture-rich structural detail from the visible spectrum is fused with semantic and thermal cues from infrared sensing. The fused output typically exhibits higher perceptual quality and richer scene, and is more resilient un-

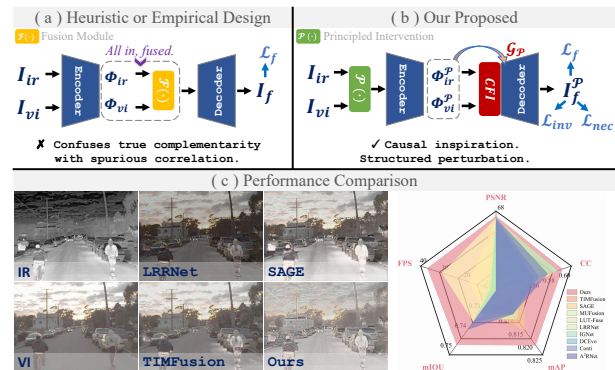


Figure 1. Comparison with SOTA method in training framework and performance. (a) General designs of existing methods, which often rely on empirical trial-and-error to fit all source features. (b) Our framework actively probes modal dependencies through structured interventions. (c) The superiority of our method is validated by its advantages in static metrics (CC, PSNR), efficiency analysis (FPS), semantic segmentation (mIoU), object detection (mAP), and qualitative comparisons.

der degradation conditions, thereby supporting downstream high-level vision tasks [5, 14, 15, 29]. As a result, MMIF has become a key component in applications such as security monitoring, autonomous driving, and medical imaging [16, 31, 33, 49].

The evolution of deep learning has catalyzed remarkable progress in MMIF. Current SOTA methods leverage sophisticated architectures, from CNN-based dual-stream networks to Transformer-based global attention mechanisms, to model complex inter-modal relationships [13, 23, 48, 53]. These approaches typically formulate fusion as an optimization problem that maximizes statistical dependencies between modalities, employing various losses to preserve intensity, gradient, and semantic information from source images. Recent advances further incorporate task-specific guidance, adversarial training, and diffusion models to enhance fusion quality [2, 6, 19, 30, 33].

Despite these empirical successes, current MMIF meth-

ods share a fundamental limitation: they learn from observational data without distinguishing genuine complementary relationships from spurious statistical regularities. This correlation-centric paradigm creates several critical challenges. When thermal signatures consistently co-occur with specific visible patterns in training data, models capture these statistical associations regardless of whether they reflect meaningful dependencies or dataset artifacts. Consequently, features are selected based on co-occurrence frequency rather than their actual contribution to fusion quality. This superficial learning leads to brittle models that fail when deployment conditions differ from training distributions, as the spurious correlations they rely upon no longer hold. The root of these issues lies in the passive nature of observational learning. Models trained solely on input-output pairs cannot determine whether observed correlations are causal or coincidental. This echoes a fundamental distinction in machine learning: while correlation reveals patterns in data, understanding causation requires active intervention. Pearl’s causal hierarchy [24] formalizes this insight through three levels of reasoning: association (*observing patterns*), intervention (*manipulating variables*), and counterfactual (*imagining alternatives*). **Current MMIF methods operate exclusively at the association level, missing crucial insights that intervention-based reasoning could provide.** However, applying causal principles to MMIF is nontrivial: unlike standard supervised tasks, fusion lacks explicit supervision for feature preservation, and modality complementarity often violates common independence assumptions. These properties demand a careful design of causal reasoning to the fusion setting.

Motivated by this gap, we propose a novel framework that leverages causal insights to design principled interventions for robust multi-modal fusion. We draw inspiration from causal principles to systematically probe modal dependencies through structured perturbations. ***Our core idea is that features genuinely important for fusion should maintain their relevance across different intervention patterns, while spurious correlations should break down under systematic perturbation.*** We instantiate this idea through three complementary intervention strategies, each designed to test specific hypotheses about modal relationships. **Complementary masking** applies spatially disjoint perturbations to different modalities, testing whether one modality can genuinely compensate for missing information in the other. This reveals true cross-modal complementarity as opposed to redundant encoding. **Random masking** corrupts identical regions across modalities to identify feature combinations that remain informative despite partial observability, highlighting robust local dependencies. **Modality dropout** completely removes individual modalities to quantify their irreplaceable contributions, preventing

over-reliance on any single information source.

Building on these interventions, we introduce the Causal Feature Integrator (CFI), which identifies intervention-invariant features through adaptive invariance gating. Unlike conventional attention mechanisms that weight features based on statistical salience, CFI explicitly models regions that exhibit consistent importance across diverse perturbation patterns. This mechanism enables the network to prioritize robust cross-modal dependencies while suppressing spurious correlations that arise from dataset biases. Our key contributions are summarized as follows:

- We propose a systematic intervention framework for MMIF that moves beyond passive correlation learning to actively probe modal dependencies, inspired by principles from causal reasoning to identify robust fusion patterns.
- We design three principled intervention strategies that test complementary aspects of modal relationships: cross-modal compensation, local sufficiency, and global necessity, each addressing specific failure modes in correlation-based fusion.
- We introduce the CFI with learnable invariance gating, which explicitly identifies and aggregates features that remain stable across interventions, enabling more robust and interpretable fusion decisions.
- Through extensive experiments on multiple benchmarks and downstream tasks, we demonstrate that intervention-based training produces models with superior generalization, particularly under challenging conditions where correlation-based methods fail.

## 2. Related Work

### 2.1. Learning-based MMIF

Deep learning has driven substantial progress in multi-modal feature processing through adaptive feature extraction, fusion, and reconstruction [4, 18, 32, 34, 35, 38]. Classical autoencoder-based methods rely on trained encoders and decoders to ensure effective feature representation, while enforcing cross-modal complementarity via carefully designed fusion rules [17, 28, 42, 43]. Representative methods such as CDDFuse [54] integrate Transformer and CNN architectures to jointly capture global structures and local details. Similarly, SwinFusion [23] adopts an end-to-end architecture that combines Transformer and CNN modules, enabling adaptive feature fusion and reducing reliance on manually engineered heuristics. Beyond these architectures, generative fusion methods such as TarDAL [16] formulate fusion as an adversarial game between the fused image and source modalities, introducing task-level supervision and optimizing fusion under multi-task objectives. More recently, diffusion models have further improved fusion quality [47, 55]. For example, Mask-DiFuser [30] leverages generative diffusion to produce high-quality

multi-modal fused images, while ControlFusion [31] combines text conditioning with diffusion to achieve controllable, degradation-aware fusion.

## 2.2. Causal-Inspired Learning

The distinction between correlation and causation has motivated numerous works to incorporate causal principles into machine learning, improving generalization, robustness, and interpretability. Causal reasoning has achieved notable success in tasks such as low-light enhancement [50, 51], self-supervised representation learning [46], and domain generalization [22, 52]. These methods typically employ intervention-based strategies, counterfactual augmentation, or structural causal models to identify stable relationships that generalize beyond training distributions. In multi-modal learning, causal perspectives have been explored to achieve robust feature learning by modeling inter-modal relationships [7, 45] and employing intervention techniques to mitigate spurious correlations [21, 25, 26].

## 3. Method

This section presents our intervention-based fusion framework, using IVIF as the primary instantiation with natural extensions to other multi-modal fusion scenarios.

### 3.1. Problem Formulation

Let  $I_{vi} \in \mathbb{R}^{3 \times H \times W}$  and  $I_{ir} \in \mathbb{R}^{1 \times H \times W}$  denote the visible and infrared input images, with  $I_f \in \mathbb{R}^{1 \times H \times W}$  representing the fused output. Conventional fusion methods model this process through statistical optimization:

$$I_f = \mathcal{F}(I_{ir}, I_{vi}) + \mathbf{n}_f, \quad (1)$$

where  $\mathcal{F}(\cdot, \cdot)$  represents the learned fusion mapping and  $\mathbf{n}_f$  captures model uncertainty. These approaches minimize reconstruction losses to capture statistical dependencies, treating all correlations as potentially informative. However, this passive learning paradigm cannot distinguish genuine modal complementarity from spurious co-occurrences arising from dataset biases.

To address this limitation, we propose a different perspective inspired by causal reasoning principles. Consider the underlying data generation process: a latent scene  $S$  generates observations through different modalities  $S \rightarrow \{\Theta_{ir}, \Theta_{vi}\} \rightarrow I_f$ , where  $\Theta_{ir}$  and  $\Theta_{vi}$  represent modality-specific features. External factors such as lighting conditions and sensor characteristics create spurious correlations that do not reflect true modal dependencies. To identify robust fusion patterns, we need to move beyond passive observation to active probing. Drawing inspiration from Pearl’s causal hierarchy [24], we recognize three levels of understanding that could benefit fusion: observing

correlations (*association*), testing under perturbations (*intervention*), and reasoning about alternatives (*counterfactual*). Our key insight is that features genuinely important for fusion should maintain their relevance under principled perturbations, while spurious correlations should degrade. This motivates our intervention-based approach: systematically perturbing inputs through structured masking operations  $\mathcal{M}$  to identify intervention-stable features.

### 3.2. Principled Intervention Design

We design three complementary intervention strategies, each testing specific hypotheses about modal relationships and addressing different failure modes of correlation-based fusion.

#### Intervention 1: Testing Cross-Modal Compensation.

True modal complementarity means modalities can compensate for each other’s missing information, not just co-occur statistically. Based on this principle, we apply spatially disjoint perturbations through complementary masking. We generate non-overlapping masks  $\mathcal{M}^v, \mathcal{M}^i \in \{0, 1\}^{H \times W}$  satisfying  $\mathcal{M}^v \cap \mathcal{M}^i = \mathbf{O}$ , where each mask randomly occludes multiple spatial regions. The intervention produces:

$$I_f^c = \mathcal{F}(I_{ir} \odot \mathcal{M}^i, I_{vi} \odot \mathcal{M}^v), \quad (2)$$

where  $\odot$  denotes element-wise multiplication. If the model successfully reconstructs the scene despite disjoint corruptions, it demonstrates genuine cross-modal compensation rather than within-modal memorization. *This intervention specifically probes whether information from modality A can functionally substitute for missing content in modality B.*

**Intervention 2: Identifying Locally Sufficient Features.** Robust fusion should tolerate partial observability, maintaining quality even when local regions are corrupted. To identify such locally sufficient feature combinations, we apply identical random masks to both modalities:

$$I_f^r = \mathcal{F}(I_{ir} \odot \mathcal{M}^r, I_{vi} \odot \mathcal{M}^r), \quad (3)$$

where  $\mathcal{M}^r \in \{0, 1\}^{H \times W}$  randomly occludes identical spatial locations across modalities. Features that preserve fusion quality despite these perturbations represent robust local dependencies essential for scene understanding. *This intervention reveals which spatial regions contain sufficient information for high-quality fusion, independent of their statistical frequency in training data.*

**Intervention 3: Quantifying Modal Necessity.** Over-reliance on single modalities is a common failure mode when models exploit spurious correlations. To ensure balanced modal utilization, we perform complete modality dropout:

$$I_f^i = \mathcal{F}(I_{ir} \odot \mathbf{O}, I_{vi}), \quad I_f^v = \mathcal{F}(I_{ir}, I_{vi} \odot \mathbf{O}). \quad (4)$$

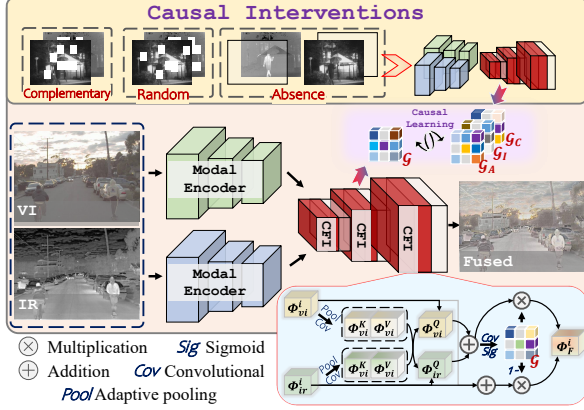


Figure 2. The framework of the proposed method. It employs a U-Net-like Siamese architecture that incorporates CFI within the decoder to enable robust multi-modal fusion. By leveraging three complementary intervention strategies, the model learns to identify intervention-stable features that represent genuine cross-modal complementarity rather than spurious statistical co-occurrences arising from dataset biases.

This extreme intervention quantifies each modality’s irreplaceable contribution by measuring performance degradation under complete removal. *By enforcing substantial quality loss when either modality is absent, we prevent the model from learning shortcuts that ignore complementary information.*

These three interventions work synergistically: complementary masking ensures genuine cross-modal interaction, random masking identifies robust local patterns, and modality dropout prevents degenerative solutions. Together, they guide the model toward learning intervention-stable features that reflect true modal dependencies.

### 3.3. Network Architecture

Our architecture, illustrated in Figure 2, implements the intervention framework through a U-Net-based design with specialized fusion modules. The network consists of parallel encoders for feature extraction and a decoder with our proposed Causal Feature Integrator (CFI) for intervention-aware fusion.

**Feature Extraction.** Two weight-sharing encoders process  $I_{vi}$  and  $I_{ir}$  independently, each containing three convolutional blocks with progressive downsampling. This generates multi-scale representations  $\{\Theta_1^v, \Theta_2^v, \Theta_3^v\}$  and  $\{\Theta_1^i, \Theta_2^i, \Theta_3^i\}$  that capture both fine details and semantic abstractions. The parallel design preserves modality-specific characteristics while enabling subsequent cross-modal reasoning.

**Causal Feature Integrator (CFI).** The CFI module learns to identify and prioritize intervention-stable features through adaptive fusion. For features  $\{\Theta_k^i, \Theta_k^v\}$  at scale  $k$ ,

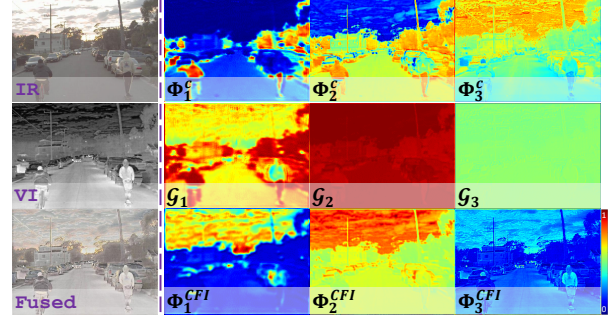


Figure 3. An illustrative example of feature visualization. Through the learnable invariance gates, the model progressively focuses on intervention-stable features across successive iterations, prioritizing regions that maintain consistent importance under perturbations to ensure effective integration of complementary modalities.

CFI performs three operations: cross-modal attention for information exchange, invariance gating for stability assessment, and adaptive fusion for feature integration.

First, we compute bidirectional cross-modal attention to capture complementary relationships. We generate *queries*, *keys*, and *values* through  $1 \times 1$  convolutions:

$$\begin{aligned} Q_k^v, K_k^v, V_k^v &= \text{Conv}_{1 \times 1}(\Theta_k^v), \\ Q_k^i, K_k^i, V_k^i &= \text{Conv}_{1 \times 1}(\Theta_k^i). \end{aligned} \quad (5)$$

To manage computational cost, we pool keys and values to compact representations of size  $r \times r$  where  $r \ll H, W$ . This pooling strategy preserves three critical properties: *i*) spatial continuity without hard boundaries, *ii*) global receptive fields essential for handling spatial misalignment between modalities, and *iii*) efficient scaling to high-resolution multi-scale features. Cross-modal attention then computes:

$$\begin{aligned} \Theta_k^{v \rightarrow i} &= \text{Attention}(Q_k^v, \text{Pool}(K_k^i), \text{Pool}(V_k^i)), \\ \Theta_k^{i \rightarrow v} &= \text{Attention}(Q_k^i, \text{Pool}(K_k^v), \text{Pool}(V_k^v)), \end{aligned} \quad (6)$$

where  $\Theta_k^{v \rightarrow i}$  captures infrared information relevant to visible queries and vice versa. This bidirectional mechanism enables the network to reason about what each modality uniquely contributes.

Next, we aggregate complementary and local features:

$$\Theta_k^c = \Theta_k^{v \rightarrow i} + \Theta_k^{i \rightarrow v}, \quad \Theta_k^l = \Theta_k^i + \Theta_k^v, \quad (7)$$

where  $\Theta_k^c$  represents cross-modal complementary features and  $\Theta_k^l$  contains local modal information.

To emphasize intervention-stable information, a learnable invariance gate  $\mathcal{G}_k$  is employed to blend the two feature types:

$$\Theta_k^{\text{CFI}} = \mathcal{G}_k \odot \Theta_k^c + (1 - \mathcal{G}_k) \odot \Theta_k^l, \quad (8)$$

where  $\mathcal{G}_k = \sigma(\text{Conv}_{3 \times 3}(\Theta_k^c))$  with sigmoid activation  $\sigma(\cdot)$ . High gate values indicate regions that remain informative across interventions (stable features), while low values mark intervention-sensitive areas (potentially spurious). This gating mechanism explicitly models feature stability, enabling the network to prioritize robust dependencies over spurious correlations.

**Feature Reconstruction.** The decoder progressively refines features through upsampling and skip connections, integrating CFI outputs at each scale. This hierarchical design enables both local detail preservation and global semantic consistency, with intervention stability propagating across scales. Figure 3 visualizes how CFI progressively focuses on genuine intervention-stable features across scales, achieving interpretable fusion decisions based on intervention stability rather than statistical saliency.

### 3.4. Training Objective

During the training phase, the model simultaneously performs principled intervention, outputting four intervention results  $(I_f^c, I_f^r, I_f^i, I_f^v)$ , which are used to impose objective loss constraints. Our training objective combines three components to optimize fusion quality and learn intervention-stable patterns: **I.** Fusion fidelity loss  $\mathcal{L}_f$  for perceptual fidelity to source modalities; **II.** Intervention consistency loss  $\mathcal{L}_{\text{inv}}$  for feature stability under perturbations; **III.** Modal necessity loss  $\mathcal{L}_{\text{nec}}$  to avoid over-dependence on single modalities:

$$\mathcal{L} = \mathcal{L}_f + \alpha \mathcal{L}_{\text{inv}} + \beta \mathcal{L}_{\text{nec}}, \quad (9)$$

where  $\alpha$  and  $\beta$  balance stability constraints against fusion quality.

**Fusion Fidelity Loss.** We use a composite objective to preserve intensity and structural details [44, 56]:

$$\begin{aligned} \mathcal{L}_f = & \|I_f - I_{vi}\|_1 + \|I_f - I_{ir}\|_1 \\ & + \lambda_1 \|\nabla I_f - \max(\nabla I_{vi}, \nabla I_{ir})\|_1, \end{aligned} \quad (10)$$

where  $\nabla$  is the Laplacian operator and  $\lambda_1$  is a trade-off coefficient. This ensures the fused image retains features from both modalities.

**Intervention Consistency Loss.** To enforce stability in key regions across perturbations, we apply:

$$\mathcal{L}_{\text{inv}} = \sum_{I_j \in \{I_f^c, I_f^r\}} \|(I_j - I_f) \odot \bar{\mathcal{G}}\|_1 + \mathcal{R}(\bar{\mathcal{G}}), \quad (11)$$

where  $I_f^c$  and  $I_f^r$  are outputs under complementary and random masking, and  $\bar{\mathcal{G}} = (\mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3)/3$  aggregates multi-scale gates. The first term penalizes differences in gate-selected stable regions.

The regularization  $\mathcal{R}(\bar{\mathcal{G}})$  avoids degenerate solutions:

$$\mathcal{R}(\bar{\mathcal{G}}) = \|\mu(\bar{\mathcal{G}} - \eta)\|_1 - \mathbf{H}(\bar{\mathcal{G}}), \quad (12)$$

where  $\mu(\cdot)$  is spatial mean,  $\eta$  targets moderate activation, and  $\mathbf{H}(\cdot)$  is spatial entropy, promoting binary-like decisions for interpretability. This loss trains gates to identify invariant features, focusing on robust cross-modal dependencies rather than spurious correlations.

**Modal Necessity Loss.** To ensure balanced modality use:

$$\mathcal{L}_{\text{nec}} = \|I_f - I_f^i\|_1 + \|I_f - I_f^v\|_1, \quad (13)$$

where  $I_f^i$  and  $I_f^v$  are infrared-only and visible-only outputs. This maximizes differences from single-modal fusions, preventing over-reliance on one modality, encouraging complementary feature discovery, and regularizing against biases. The combined objective drives the model toward high-quality fusion with stable, complementary cross-modal patterns.

## 4. Experiment

### 4.1. Implementation Details

Our experiments were conducted using PyTorch on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU. We employed the training data provided by MSRS [27] as the training dataset and RoadScene [41] as the validation dataset. Training image pairs were cropped into  $256 \times 256$  patches with random augmentation and fed into the network with a batch size of 16. The model was trained for 50 epochs using the Adam optimizer with a learning rate of  $1e-4$ . We empirically set the hyperparameters as  $\alpha = 0.1$ ,  $\beta = 0.05$ , and  $\lambda_1 = 1.0$  to ensure comparable magnitudes among the loss terms while prioritizing fusion quality as the primary objective. Following parameter analysis (*reported in the Supplementary Materials due to space constraints*), we set both the size of Complementary and Random masks to  $16 \times 16$ ; the number of masks was randomly sampled from 1 to 6 with the total masked area constrained not to exceed the training patch, and we set  $\eta = 0.3$  and  $r = 8$ . Five quantitative metrics, including PSNR, SF, AG, CC, and  $\mathcal{Q}_{abf}$ , were employed to objectively evaluate the fusion performance. Higher values indicate superior fusion results, with computational details provided in [18].

### 4.2. Infrared and visible image fusion

We evaluate our proposed method on three widely-used IVIF benchmarks: MSRS, TNO [36], and M<sup>3</sup>FD [16]. Our method was compared with 9 SOTA IVIF methods, including TIMFusion [20], Conti [10], SAGE [39], MUFusion [3], LUT-Fuse [49], LRRNet [9], IGRNet [12], DCEvo [19], and A<sup>2</sup>RNet [13].

#### 4.2.1. Comparison with SOTA methods

**Qualitative comparison.** Figure 4 presents qualitative comparisons across three benchmarks. Through systematic intervention-based training, our method robustly integrates



Figure 4. Qualitative comparison with SOTA methods on the IVIF benchmarks.

Table 1. Quantitative results of our proposed method vs. SOTA methods on the IVIF benchmarks. The best value is highlighted with **Bold**.

Method	TNO					MSRS					M <sup>3</sup> FD				
	AG	SF	PSNR	CC	$Q_{abf}$	AG	SF	PSNR	CC	$Q_{abf}$	AG	SF	PSNR	CC	$Q_{abf}$
TIMFusion	4.019	3.901	61.41	0.379	0.397	3.747	4.378	59.48	0.525	0.424	4.272	5.176	61.85	0.425	0.489
Conti	3.860	3.987	61.12	0.446	0.437	3.737	4.504	64.26	0.603	0.570	4.476	5.545	61.11	0.479	0.499
SAGE	3.817	3.745	59.81	0.317	0.444	3.431	4.373	65.59	0.594	0.529	4.590	5.443	61.88	0.557	0.510
MUFusion	4.713	3.867	61.73	0.448	0.319	3.160	3.482	63.85	0.589	0.422	4.220	4.155	61.72	0.451	0.372
LUT-Fuse	3.810	3.830	61.23	0.436	0.403	3.801	4.494	64.24	0.602	<b>0.579</b>	4.080	5.019	61.31	0.470	0.446
LRRNet	3.855	3.806	61.72	0.436	0.346	2.672	3.309	64.68	0.515	0.447	3.613	4.212	<b>62.95</b>	0.541	0.484
IGNet	4.198	3.857	60.48	0.459	0.368	3.318	3.889	65.33	<b>0.655</b>	0.455	4.716	5.137	60.55	0.525	0.400
DCEvo	3.942	4.015	61.24	0.460	<b>0.447</b>	3.807	4.512	64.49	0.605	0.620	4.575	5.517	61.33	0.500	0.504
A <sup>2</sup> RNet	3.292	3.429	61.38	0.462	0.289	2.924	3.493	63.37	0.603	0.416	3.005	3.470	62.12	0.519	0.302
<b>Ours</b>	<b>5.128</b>	<b>5.132</b>	<b>62.06</b>	<b>0.502</b>	0.423	<b>4.129</b>	<b>4.743</b>	<b>66.02</b>	0.646	0.545	<b>5.276</b>	<b>6.105</b>	62.13	<b>0.565</b>	<b>0.512</b>

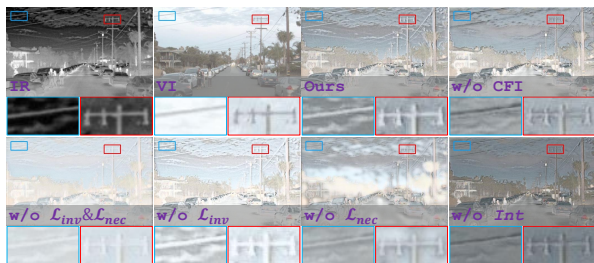


Figure 5. Visualization comparison of ablation study

complementary features under extreme conditions such as heavy fog and overexposure, demonstrating significant advantages. In contrast, LRRNet’s lightweight design and hand-crafted parameters limit its representation capacity, while semantically-guided methods like SAGE and DCEvo often produce blurred details or insufficient contrast in thermal regions due to their reliance on statistical correlations. By learning intervention-stable features through our principled perturbation strategies, our method achieves clear and

Table 2. Ablation experiment results. The best value is highlighted with **Bold**.

Case	AG	SF	PSNR	CC	$Q_{abf}$
w/o CFI	5.764	5.972	60.21	0.544	0.428
w/o $\mathcal{L}_{inv}$	5.179	5.728	58.08	0.573	0.331
w/o $\mathcal{L}_{nec}$	4.016	4.018	61.39	0.393	0.368
w/o $\mathcal{L}_{nec}&\mathcal{L}_{inv}$	3.361	3.478	59.85	0.524	0.312
w/o $\mathcal{L}_{int}$	5.332	5.348	<b>63.95</b>	0.598	<b>0.524</b>
<b>Ours</b>	<b>6.136</b>	<b>6.244</b>	63.62	<b>0.605</b>	0.467

accurate representation of thermal objects, effectively preserving critical semantic information.

**Quantitative Comparisons.** Table 1 validates the systematic advantages of our intervention framework, achieving optimal overall performance across three benchmarks. The intervention consistency constraints enhance edge preservation, cross-modal compensation through complementary masking ensures high-fidelity reconstruction in complex scenes, and the invariance gating mechanism

achieves stable cross-scenario generalization. Compared to correlation-based methods, our intervention-guided approach delivers stronger robustness and generalization.

#### 4.2.2. Ablation Study

The ablation study in Table 2 and Figure 5 systematically validates each component’s contribution. Removing CFI maintains edge metrics but introduces visible noise and structural distortions, confirming its role in identifying intervention-stable features. Without  $\mathcal{L}_{inv}$ , PSNR degrades substantially, demonstrating that intervention consistency constraints are essential for robust feature learning. Removing  $\mathcal{L}_{nec}$  significantly impairs AG and SF with evident image distortion, validating its role in ensuring balanced modal utilization. Notably, completely eliminate intervention mechanisms and only use  $\mathcal{L}_f$  and backbone network (w/o *Int*) setting attains higher PSNR and  $Q_{abf}$  but markedly lower AG and SF, revealing a core trade-off in fusion goals. Correlation-driven optimization favors pixel fidelity and local gradients, whereas AG and SF capture structural coherence and texture richness crucial for semantics. As shown in Figure 5, w/o *Int* yields smoother outputs but loses fine thermal details and edge definition. Our intervention framework instead prioritizes structural integrity and feature stability, achieving more balanced performance across metrics.

#### 4.2.3. Intervention Impact Analysis

We quantify the impact of our interventions through Average Treatment Effect (ATE) [37] analysis. For intervention  $T \in \{0, 1, 2, 3, 4\}$  (baseline, complementary/random masking, IR/VI dropout) and outcome  $Y_i(t)$  representing fusion quality, ATE is defined as  $\mathbb{E}[Y_i(0) - Y_i(t)]$ . We estimate ATE via sample average:  $\widehat{ATE}(t) = N^{-1} \sum_{i=1}^N [Q(f(I_i)) - Q(f(M_t(I_i)))]$ , where  $Q$  measures PSNR/CC and  $M_t$  applies intervention  $t$ . Figure 6 reveals distinct intervention impacts. Modality dropout induces the largest degradation, confirming both modalities provide irreplaceable information. Random masking produces minimal effects, indicating successful learning of locally sufficient features. Complementary masking shows moderate impact, validating cross-modal compensation capabilities. This confirms our framework learns intervention-stable features from modal-specific information to local patterns to cross-modal relationships.

### 4.3. Performance on High-Level Vision Tasks

To further validate the generalization capability of our intervention framework, we evaluate its performance on two representative high-level vision tasks: object detection and semantic segmentation.

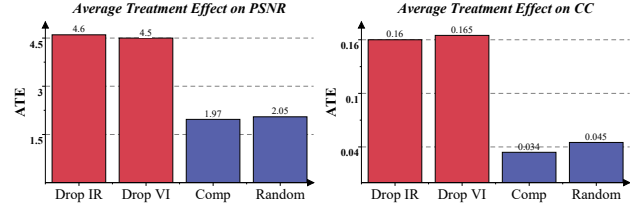


Figure 6. Quantitative results of ATE.

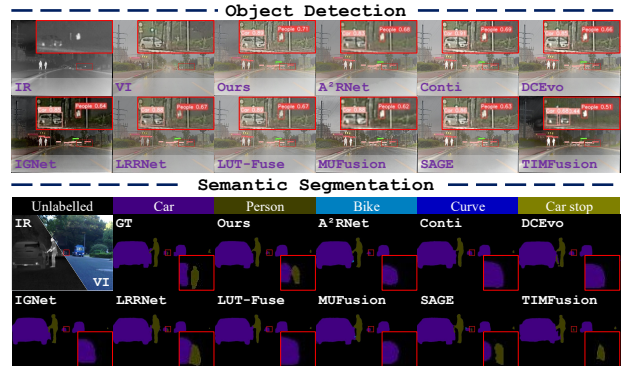


Figure 7. Qualitative comparison with SOTA IVIF methods in object detection and semantic segmentation tasks.

#### 4.3.1. Object Detection

We follow [38] in splitting the M<sup>3</sup>FD object detection dataset into training, validation, and testing sets with a ratio of 6:2:2. YOLOv5 [8] is adopted as the baseline detector to evaluate detection performance of the proposed method. As summarized in Table 3 and the upper panel of Figure 7, our method attains the highest mAP and the best class-wise AP on *Car* and *Bus*. While methods such as TIMFusion achieve leading AP on specific categories, their substantially lower overall mAP suggests limited cross-category generalization. We attribute our gains to CFI’s bidirectional cross-modal attention combined with invariance gating, which identifies features that remain informative across different intervention patterns. This yields more complete and robust fused representations. In contrast, correlation-driven methods (e.g., LRRNet, DCEvo) often produce blurred thermal structures or insufficient local contrast, ultimately reducing detection accuracy.

#### 4.3.2. Semantic Segmentation

Following [38], we adopt the MSRS dataset and employ the official segmentation network [2] to perform the semantic segmentation task, thereby evaluating the performance of the proposed method in segmentation. Table 4 presents quantitative results across six categories. Our method attains the best or near-best performance on multiple key classes, demonstrating balanced generalization across semantic categories. The lower panel of Figure 7 shows

Table 3. Quantitative results of our proposed method vs. SOTA methods on the object detection. The best value is highlighted with **Bold**.

Method	mAP	Peo	Car	Bus	Mot	Tru	Lam
TIMFusion	0.796	<b>0.802</b>	0.900	0.846	0.653	0.816	0.756
Conti	0.809	0.789	0.910	0.875	0.650	0.834	0.797
SAGE	0.815	0.783	0.908	0.870	<b>0.704</b>	0.813	0.814
MUFusion	0.804	0.798	0.906	0.881	0.660	0.792	0.788
LUT-Fuse	0.804	0.772	0.905	0.881	0.654	<b>0.836</b>	0.775
LRRNet	0.811	0.780	0.911	0.878	0.694	0.804	0.798
IGNet	0.673	0.767	0.855	0.739	0.452	0.662	0.565
DCEvo	0.809	0.780	0.907	0.891	0.675	0.788	<b>0.815</b>
A <sup>2</sup> RNet	0.809	0.795	0.906	0.886	0.659	0.810	0.798
<b>Ours</b>	<b>0.821</b>	0.800	<b>0.916</b>	<b>0.894</b>	0.693	0.812	0.809

Table 4. Quantitative results of our proposed method vs. SOTA methods on the semantic segmentation. The best value is highlighted with **Bold**.

Method	Unl	Car	Per	Bik	Cur	Roa	mIOU
TIMFusion	0.975	0.826	0.597	0.597	0.389	0.450	0.639
Conti	0.982	0.878	0.699	0.671	0.540	0.646	0.736
SAGE	0.982	0.871	0.678	0.680	0.548	0.621	0.730
MUFusion	0.981	0.866	0.667	0.679	0.507	0.634	0.722
LUT-Fuse	0.982	0.878	0.687	0.681	0.564	0.623	0.736
LRRNet	0.982	0.880	0.676	0.679	0.548	0.642	0.734
IGNet	0.982	0.872	0.685	0.688	0.541	<b>0.648</b>	0.736
DCEvo	0.982	0.878	0.687	0.679	0.524	0.638	0.731
A <sup>2</sup> RNet	0.982	0.881	0.687	<b>0.688</b>	0.556	0.644	0.740
<b>Ours</b>	<b>0.983</b>	<b>0.883</b>	<b>0.707</b>	0.686	<b>0.584</b>	0.642	<b>0.747</b>

that our method produces finer delineation of vehicles and pedestrians than SOTA methods. This advantage stems from our random masking intervention during training: by corrupting identical regions across modalities, the model learns to reconstruct complete semantics from partial evidence, thereby achieving higher detail fidelity in regions with complex boundaries or delicate textures.

#### 4.4. Medical Image Fusion

To evaluate cross-domain generalization, we conduct medical image fusion (MIF) experiments on the Harvard medical dataset [1] containing 20 MRI-PET/SPECT image pairs. Notably, we directly deploy the IVIF-trained model to MIF without fine-tuning, constituting a challenging cross-sensor transfer. Table 5 shows our method achieves the highest AG and SF with competitive PSNR and CC. Figure 8 further demonstrates that our method produces coherent contours and stable contrast for fine anatomical structures, while correlation-driven methods like LRRNet and DCEvo yield blurred details, and semantically-guided methods like TIMFusion and SAGE exhibit over-sharpening artifacts

Table 5. Quantitative results of our proposed method vs. SOTA methods on the MIF. The best value is highlighted with **Bold**.

Method	AG	SF	PSNR	CC	$Q_{abf}$
TIMFusion	5.122	5.840	61.37	0.800	0.467
Conti	7.205	9.473	63.98	0.857	0.637
SAGE	5.406	6.860	63.91	0.861	0.376
MUFusion	4.777	5.234	62.21	0.861	0.355
LUT-Fuse	6.549	8.467	64.34	0.852	0.605
LRRNet	3.753	4.564	63.92	0.834	0.211
IGNet	5.623	6.035	<b>65.13</b>	0.860	0.499
DCEVO	7.391	9.585	63.82	0.850	<b>0.688</b>
A <sup>2</sup> RNet	5.131	5.317	63.18	0.860	0.386
<b>Ours</b>	<b>7.681</b>	<b>10.06</b>	64.33	<b>0.862</b>	0.599

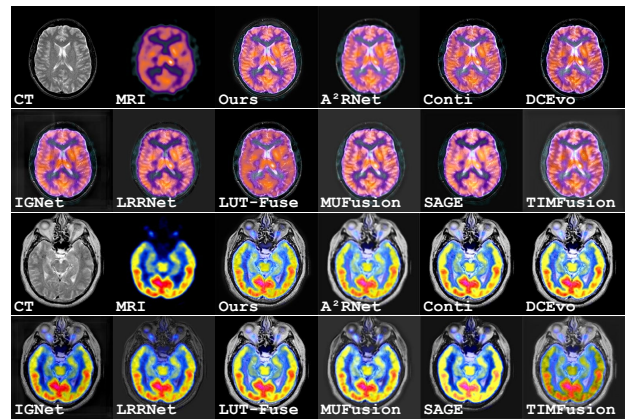


Figure 8. Qualitative comparison with SOTA methods on the MIF.

or texture collapse. This cross-sensor transfer success validates that intervention-based training captures fundamental fusion principles rather than domain-specific patterns. The invariance gating mechanism successfully identifies features that remain informative across vastly different imaging modalities, from thermal-visible to anatomical-functional medical imaging, confirming the generalizability of intervention-stable features.

#### 5. Conclusion

We propose an intervention-based MMIF framework that actively probes modal dependencies through three complementary masking strategies, testing cross-modal compensation, local feature robustness, and balanced modal utilization respectively. Our CFI employs learnable invariance gating to prioritize perturbation-robust features, steering the network toward genuine dependencies rather than spurious correlations. Experiments demonstrate SOTA performance, cross-sensor robustness, and zero-shot IVIF-to-medical transferability, confirming that our framework captures fundamental fusion principles.

## Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants No.62162065 and 61761045; Yunnan Province Visual and Cultural Innovation Team No.202505AS350009; Yunnan Fundamental Research Projects No.202201AT070167; Joint Special Project Research Foundation of Yunnan Province No.202401BF070001023.

## References

- [1] Harvard medical website. Available at: <http://www.med.harvard.edu/AANLIB/home.html>. 8
- [2] Wenzhi Cao, Minghui Zheng, and Qing Liao. Semantic region adaptive fusion of infrared and visible images via dual-deeplab guidance. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. 1, 7
- [3] Chunyang Cheng, Tianyang Xu, and Xiao-Jun Wu. Mufusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92:80–92, 2023. 5
- [4] Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. 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*, pages 28102–28112, 2025. 2
- [5] Zheng Guan, Xue Wang, Wenhua Qian, Peng Liu, and Runzhuo Ma. Residual prior-driven frequency-aware network for image fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1082–1091, 2025. 1
- [6] Lin Guo, Xiaoqing Luo, Wei Xie, Zhancheng Zhang, Hui Li, Rui Wang, Zhenhua Feng, and Xiaoning Song. Revisiting generative infrared and visible image fusion based on human cognitive laws. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1
- [7] Menghua Jiang, Yuxia Lin, Baoliang Chen, Haifeng Hu, Yuncheng Jiang, and Sijie Mai. Disentangling bias by modeling intra-and inter-modal causal attention for multimodal sentiment analysis. *arXiv preprint arXiv:2508.04999*, 2025. 3
- [8] Glenn Jocher. YOLOv5. <https://github.com/ultralytics/yolov5>, 2020. 7
- [9] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):11040–11052, 2023. 5
- [10] Hui Li, Haolong Ma, Chunyang Cheng, Zhongwei Shen, Xiaoning Song, and Xiao-Jun Wu. Conti-fuse: A novel continuous decomposition-based fusion framework for infrared and visible images. *Information Fusion*, 117:102839, 2025. 5
- [11] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1
- [12] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4471–4479, 2023. 5
- [13] Jiawei Li, Hongwei Yu, Jiansheng Chen, Xinlong Ding, Jinlong Wang, Jinyuan Liu, Bochao Zou, and Huimin Ma. A<sup>2</sup>rnet: Adversarial attack resilient network for robust infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4770–4778, 2025. 1, 5
- [14] Xilai Li, Xiaosong Li, Tianshu Tan, Huafeng Li, and Tao Ye. Umcfuse: A unified multiple complex scenes infrared and visible image fusion framework. *IEEE Transactions on Image Processing*, 2025. 1
- [15] Xilai Li, Wuyang Liu, Xiaosong Li, Fuqiang Zhou, Huafeng Li, and Feiping Nie. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *Information Fusion*, page 104130, 2026. 1
- [16] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 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*, pages 5802–5811, 2022. 1, 2, 5
- [17] Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 2024. 2
- [18] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 5
- [19] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025. 1, 5
- [20] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. 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, 2024. 5
- [21] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023. 3
- [22] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8046–8056, 2022. 3
- [23] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range

- learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 1, 2
- [24] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010. 2, 3
- [25] Teng Sun, Wenjie Wang, Liqing Jing, Yiran Cui, Xueming Song, and Liqiang Nie. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 15–23, 2022. 3
- [26] Teng Sun, Juntong Ni, Wenjie Wang, Liqing Jing, Yinwei Wei, and Liqiang Nie. General debiasing for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5861–5869, 2023. 3
- [27] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 5
- [28] Linfeng Tang, Ziang Chen, Jun Huang, and Jiayi Ma. Camf: An interpretable infrared and visible image fusion network based on class activation mapping. *IEEE Transactions on Multimedia*, 26:4776–4791, 2023. 2
- [29] Linfeng Tang, Yuxin Deng, Xunpeng Yi, Qinglong Yan, Yixuan Yuan, and Jiayi Ma. Drmf: Degradation-robust multimodal image fusion via composable diffusion prior. In *Proceedings of the ACM International Conference on Multimedia*, pages 8546–8555, 2024. 1
- [30] Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2
- [31] Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion framework with language-vision degradation prompts. *Advances in Neural Information Processing Systems*, 2025. 1, 3
- [32] Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: A spatio-temporal collaborative network for multi-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*, 2025. 2
- [33] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, 133:5262–5280, 2025. 1
- [34] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25:5413–5428, 2022. 2
- [35] Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022. 2
- [36] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249, 2017. 5
- [37] Tyler J VanderWeele and Miguel A Hernan. Causal inference under multiple versions of treatment. *Journal of causal inference*, 1(1):1–20, 2013. 7
- [38] Xue Wang, Zheng Guan, Wenhua Qian, Jinde Cao, Runzhuo Ma, and Cong Bi. A degradation-aware guided fusion network for infrared and visible image. *Information Fusion*, 118:102931, 2025. 2, 7
- [39] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025. 5
- [40] Jingwei Xin, Boneng Shi, Nannan Wang, Jie Li, and Xinbo Gao. Mvffusion: Generative representation learning with masked variational autoencoders for multi-modality image fusion. *IEEE Transactions on Image Processing*, 2025. 1
- [41] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):502–518, 2020. 5
- [42] Han Xu, Xinya Wang, and Jiayi Ma. Drf: Disentangled representation for visible and infrared image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. 2
- [43] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021. 2
- [44] Han Xu, Meiqi Gong, Xin Tian, Jun Huang, and Jiayi Ma. Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. *Computer Vision and Image Understanding*, 218:103407, 2022. 5
- [45] Zhi Xu, Dingkan Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. Debaised multimodal understanding for human language sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14450–14458, 2025. 3
- [46] Jiexi Yan, Cheng Deng, Heng Huang, and Wei Liu. Causality-invariant interactive mining for cross-modal similarity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6216–6230, 2024. 3
- [47] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024. 2
- [48] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. 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*, pages 27026–27035, 2024. 1
- [49] Xunpeng Yi, Yibing Zhang, Xinyu Xiang, Qinglong Yan, Han Xu, and Jiayi Ma. Lut-fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14559–14568, 2025. 1, 5
- [50] Tongshun Zhang, Pingping Liu, Yubing Lu, Menggen Cai, Zijian Zhang, Zhe Zhang, and Qiuzhan Zhou. Cwnet: Causal wavelet network for low-light image enhancement.

- In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8789–8799, 2025. [3](#)
- [51] Tongshun Zhang, Pingping Liu, Zhe Zhang, and Qiuzhan Zhou. Civqlie: Causal intervention with vector quantization for low-light image enhancement. *arXiv preprint arXiv:2508.03338*, 2025. [3](#)
- [52] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyang Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5372–5382, 2021. [3](#)
- [53] Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. Freefusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. [1](#)
- [54] Zixiang Zhao, Haowen Bai, Jianshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. 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*, pages 5906–5916, 2023. [1](#), [2](#)
- [55] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jianshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8082–8093, 2023. [2](#)
- [56] Zixiang Zhao, Haowen Bai, Jianshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024. [5](#)