# More Than Meets the Eye: A Unified Image Fusion Framework via Semantic-Pixel Entropy Trade-off for Zero-Shot Generalization

Xiaowen Liu[1,*], Jing Li[2,*], Hongtao Huo[1,†], Haozhe Cao[1], Renhua Wang[1], Xu Dong[1]
[1]People's Public Security University of China, [2]East China Normal University
huohongtao@ppsuc.edu.cn, 2024111025@stu.ppsuc.edu.cn

## Abstract

*Existing image fusion methods face difficulties in adapting to unseen fusion tasks and have limitations in balancing semantic information with pixel-level details. This limitation can be attributed to three key challenges: (1) the lack of a unified, task-agnostic optimization objective; (2) the inherent difficulty in balancing semantic fidelity and pixel-level richness; and (3) an over-reliance on supervised learning, which limits transferability across tasks. To overcome these issues, this work proposes a unified fusion framework that generalizes to diverse fusion tasks even when trained solely on infrared–visible image pairs. Specifically, inspired by the free-energy principle, we introduce a fusion paradigm that combines high pixel-entropy expectation with low semantic-entropy expectation, and we design a frequency-aware feature decoupling mechanism to balance semantic content and pixel detail. Furthermore, an unsupervised dual-path trade-off strategy provides collaborative constraints at both semantic and pixel levels. Experiments show that our method significantly outperforms existing state-of-the-art methods in visual quality and downstream-task performance. It not only handles trained tasks efficiently but also generalizes well to unseen fusion tasks, while featuring lightweight model parameters and strong practical applicability. The code is available at* https://github.com/XiaoW-Liu/DECC.

## 1. Introduction

Multi-source image fusion, which integrates complementary information from multiple inputs to enhance scene representation and interpretation, has been widely adopted in applications such as surveillance and medical imaging [13, 29, 61, 72]. Based on the characteristics of the input data, fusion tasks can be broadly categorized into two types: cross-modal fusion [14, 17, 33, 65, 70], which exploits complementary information from different sensor modalities to overcome individual limitations and improve scene understanding under challenging conditions [16, 42, 43, 53, 54, 56, 59], and digital photography fusion, where sequences of images captured by a single sensor under varying parameters are combined to produce high-quality outputs.

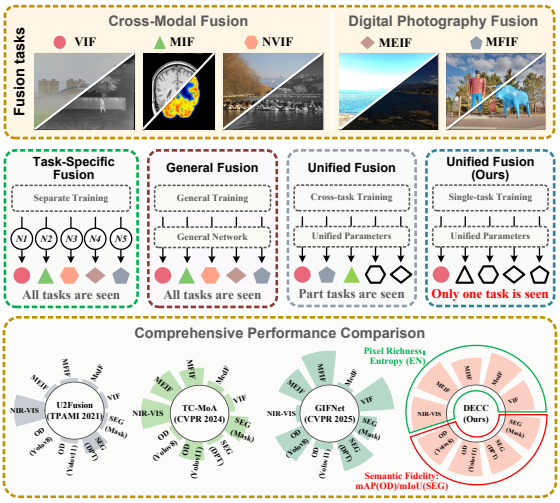


Figure 1. Comparison of task adaptability, training strategies, and performance metrics across image fusion methods. Metrics include pixel entropy for visual quality, and object detection (OD) and semantic segmentation (SEG) for task-oriented evaluation.

Traditionally, image fusion has relied on task-specific specialized networks. As illustrated in Figure 1, these methods require task-specific training data and architectures, limiting their adaptability and generalization across different tasks [12, 63]. To address this, recent studies have developed general-purpose networks for unifying multiple fusion tasks [5, 35, 66, 74]. While enabling flexible feature processing, these methods need explicit task identification during inference, hindering generalization to unseen fusion tasks. Unified fusion methods aim to solve this by integrating diverse tasks into a shared network with universal ob-

*: Equal contribution
†: Corresponding author

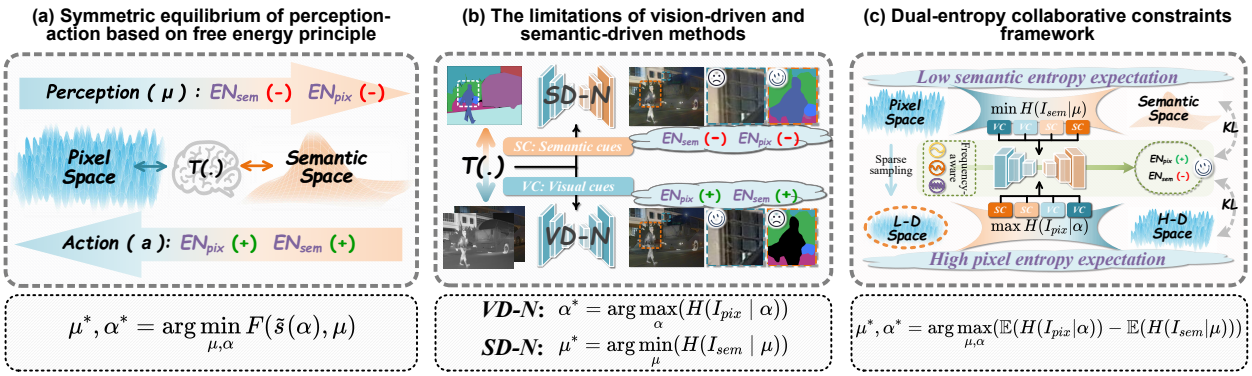


Figure 2. Visualization of perception-reconstruction mechanisms and fusion method limitations. (a) perception action equilibrium based on the free energy principle, showing perception/action (reconstruction) interaction as well as the trends of pixel entropy ($EN_{pix}$) and semantic entropy ($EN_{sem}$); (b) Limitations of vision-driven (VD-N) and semantics-driven (SD-N) methods; (c) Dual-Entropy Collaborative Constraints (DECC) framework, illustrating dual-entropy constraint implementation via dual-path training scheme, frequency-aware feature decoupling and KL-divergence constraint. $T(.)$ denotes the "thinking agent" that realizes information transformation.

jectives, facilitating task-agnostic knowledge transfer. Xu *et al.* pioneered such methods using continual learning for parameter optimization, but these strategies are data-intensive, susceptible to task bias, and still suffer from catastrophic forgetting [52]. Additionally, focusing on pixel-level commonalities in multi-task learning sacrifices semantic fidelity, often requiring auxiliary task-specific branches to capture discriminative cross-task information [4, 6]. This raises a fundamental question: **Can we design a concise fusion framework that achieves zero-shot generalization across diverse tasks while maintaining robust visual-semantic integrity, even when trained on only a single type of fusion task?**

In addressing this question, we identify three key challenges: (1) **Unified Optimization Objective:** Achieving zero-shot generalization requires the optimization space of the anchor task to be inherently aligned with other fusion tasks. This necessitates an objective function that transcends data-specific properties and instead captures the fundamental principles underlying fusion tasks. (2) **Trade-off between Semantic and Pixel-level Features:** While semantic fidelity is crucial for many fusion applications, a significant domain gap exists between semantic and pixel-level features. Methods heavily driven by downstream tasks often sacrifice fine-grained pixel details, making it challenging to strike an optimal balance between high-level semantics and low-level visual richness. (3) **Limitations of Supervised Learning:** Cross-modal fusion inherently suffers from the absence of ground-truth fused images. Over-reliance on task-specific supervision can hinder the learning of shared, transferable representations across different tasks, underscoring the need for robust unsupervised learning techniques.

In computational neuroscience, Friston *et al.* introduced the **free energy principle** to describe how intelligent agents minimize prediction errors through perception inference and active inference [9]. Drawing inspiration from this, we conceptualize image fusion as a unified process of variational free energy minimization, comprising two complementary mechanisms: perception inference, which maps pixel-level cues into semantic space to reduce uncertainty, and generative reconstruction, which reconstructs pixel-level details from semantic representations in a low-dimensional space to maximize information content. Concretely, perception inference seeks to minimize semantic entropy to preserve semantic fidelity, whereas generative reconstruction aims to maximize pixel entropy to enhance visual richness. As illustrated in Figure 2, this symmetric perception–reconstruction cycle underpins our dual-entropy optimization strategy, which effectively balances semantic certainty and pixel diversity in a unified manner.

To address **Challenge (1)**, we propose the **D**ual-**E**ntropy **C**ollaborative **C**onstraints (DECC) framework. Guided by a unified free-energy objective, DECC minimizes semantic entropy while maximizing pixel entropy, thereby enabling zero-shot generalization. This is realized through a dual-path training scheme: a perception path that preserves semantic fidelity by integrating fusion with object detection tasks, and a reconstruction path that enhances visual richness via joint fusion and generative reconstruction tasks. Shared parameters between the two paths align their optimization spaces and facilitate mutual regularization. For **Challenge (2)**, we introduce a dynamic frequency-aware feature decoupling mechanism, built upon a frequency-dynamic encoder. At its core lies a grouped frequency dynamic convolution [2], which decomposes representations into distinct frequency-aware feature groups. During iterative training, parameters corresponding to semantic-

oriented and pixel-oriented groups are alternately frozen, thereby achieving cross-domain decoupling. The cross-modal and cross-frequency fusion mechanism further enables dynamic feature interaction across modalities and domains. To tackle **Challenge (3)**, we design an unsupervised dual-path trade-off loss. In the perception path, a Kullback-Leibler (KL) divergence constraint applied between the fused image and semantic representations reduces semantic entropy. In the reconstruction path, a KL-divergence constraint between the fused result and pixel-level representations increases pixel entropy. Together, these form a complete and unified fusion method grounded in a semantic–pixel entropy trade-off.

Our contributions are detailed in the following aspects:

- We propose a unified fusion framework that minimizes semantic entropy while maximizing pixel entropy under a free-energy objective, enabling zero-shot generalization with training on only the visible-infrared fusion task.
- A symmetric training paradigm is proposed to jointly optimize perception inference for semantic fidelity and generative reconstruction for visual richness, with shared parameters enabling mutual regularization.
- We introduce a dynamic frequency-aware decoupling strategy with grouped frequency dynamic convolution that disentangles semantic and pixel features in the frequency domain, balancing their contributions through cross-modal and cross-frequency fusion.
- An unsupervised dual-path trade-off loss is proposed to balance semantic and pixel information via KL-divergence constraints, realizing the semantic-pixel distribution trade-off.

## 2. Related Work

### 2.1. Vision-driven fusion methods

Traditional image fusion methods rely on handcrafted algorithms that decompose source images into visual features for reconstruction [32]. With the rise of deep learning, these manual designs are replaced by end-to-end models that incorporate pixel-level or structural constraints to enhance the consistency between fused and source images [11, 18, 24, 57]. CNN-based models excel at capturing local details [21, 30, 34, 37, 55, 58], while transformer-based models focus on global context [12, 15, 19, 20, 23, 46, 47], aiding large-scale structure reconstruction. However, the lack of high-level perceptual constraints can introduce artifacts during pixel reconstruction, increasing semantic uncertainty.

### 2.2. Semantic-driven fusion methods

To preserve semantic information, recent methods incorporate downstream task supervision [1, 22, 63]. This includes cascaded frameworks that enforce semantic consistency [25, 40, 41, 48, 63], and cross-domain feature coupling or semantic feature injection to embed high-level semantic information [7, 28, 31, 44, 68]. Nonetheless, the heterogeneity between pixel-level and semantic-level features may disrupt microstructural fidelity, limiting generalization, especially on unseen fusion tasks.

### 2.3. Unified fusion methods

Unified methods aim to improve cross-task generalization across multiple fusion scenarios [45]. Xu *et al.* first proposed a continual learning-based unified fusion framework that accumulates and transfers knowledge across tasks. Subsequently, multi-task, multi-modal paradigms have been developed to learn shared features for better performance. Despite emphasizing generalization, these methods often focus on seen tasks. Cheng *et al.* introduced a cross-fusion gating mechanism, which achieves strong performance on unseen fusion tasks even when trained on a limited set of tasks [6]. However, this mechanism still requires task-specific selections during training and auxiliary networks to handle feature differences, which increases complexity.

## 3. Methodology

### 3.1. Motivation

A core challenge in existing image fusion methods is balancing visual richness with semantic fidelity. The lack of effective trade-off mechanisms hinders their ability to leverage both advantages simultaneously. Moreover, the absence of a unified optimization objective forces generalized fusion models to rely on extensive data fitting and task-specific auxiliary branches, which limits their generalization across diverse tasks.

Inspired by the free energy principle [9], we propose a unified dual-entropy collaborative constraints framework grounded in perception-reconstruction symmetry. This framework explicitly addresses the intrinsic trade-off between minimizing semantic uncertainty and maximizing pixel-level diversity. The free energy principle posits that intelligent agents minimize variational free energy through perception inference and active reconstruction. Mathematically, this is expressed as:

$$\mu^*, \alpha^* = \arg\min_{\mu,\alpha} F(\tilde{s}(\alpha), \mu), \tag{1}$$

where $F(\cdot,\cdot)$ represents the free energy, quantifying the balance between the "surprise" of sensory inputs $\tilde{s}(\alpha)$, governed by action $\alpha$, and the complexity of internal perception states $\mu$. We decompose the free energy as:

$$F(\tilde{s}(\alpha), \mu) = \underbrace{\mathbb{E}_{q(\mu)}[-\log p(\tilde{s}(\alpha)|\mu)]}_{\text{prediction error}} + \underbrace{\mathrm{KL}[q(\mu)\|p(\mu)]}_{\text{complexity}}, \tag{2}$$

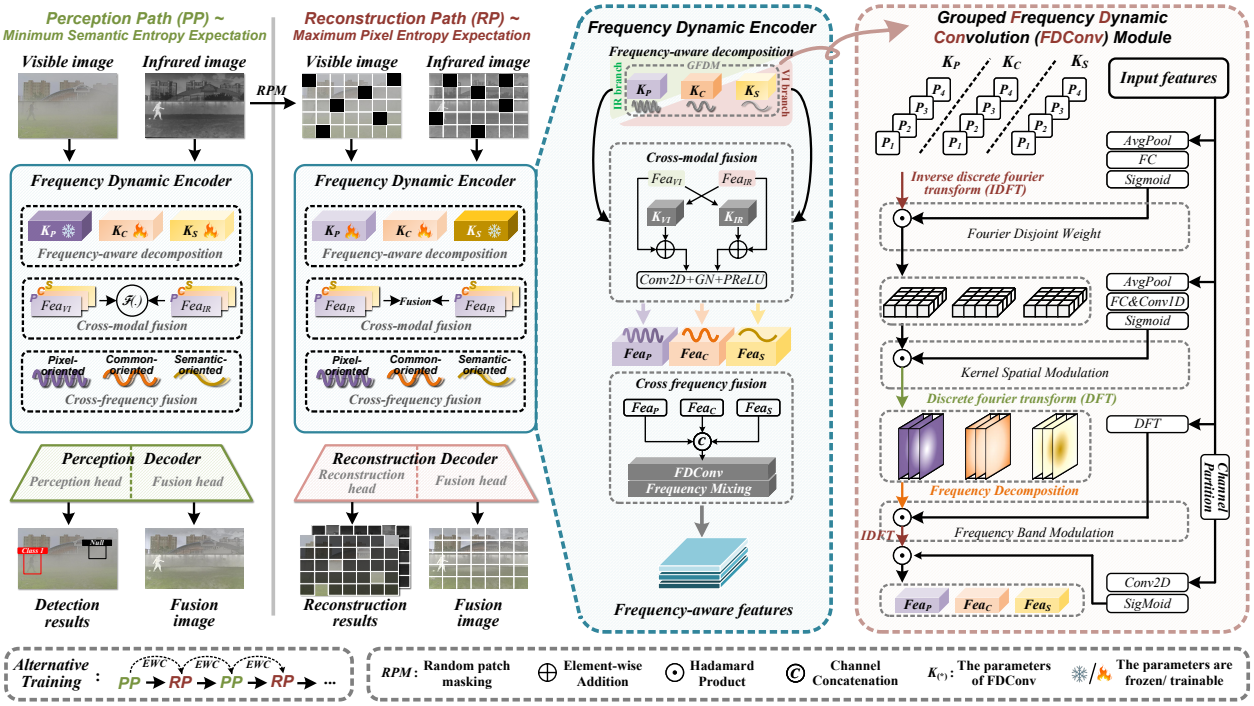


Figure 3. The network architecture of the **Dual-Entropy Collaborative Constraints (DECC)** framework, designed for the infrared/visible image semantic perception and pixel reconstruction workflow, includes the frequency dynamic encoder, perception decoder, reconstruction decoder, and a dual-path alternating training mechanism. The detailed encoder architecture is provided in the supplementary material.

where the first term denotes the prediction error, the second term regularizes the deviation from the prior, $q(\mu)$ is the variational posterior distribution over internal states $\mu$, and $p(\mu)$ is the prior distribution over internal states.

Based on the close relationship between the prediction error term in free energy and conditional entropy, we approximate the variational free energy minimization as:

$$F(\tilde{s}(\alpha),\mu) \propto \mathbb{E}[H(I_{\text{sem}}|\mu)] - \mathbb{E}[H(I_{\text{pix}}|\alpha)] + C \quad (3)$$

where $C$ represents constant terms irrelevant to the optimization (Detailed derivations of the formulations are provided in the supplement). Parallel to this, we posit that an image fusion model must simultaneously optimize semantic certainty during perception and pixel diversity during reconstruction. Accordingly, we conceptualize the fusion process as comprising two complementary paths: a perception path that maps pixels to semantics by minimizing semantic entropy $\mathbb{E}[H(I_{\text{sem}}|\mu)]$, and a reconstruction path that maps semantics back to pixels by maximizing pixel entropy $\mathbb{E}[H(I_{\text{pix}}|\alpha)]$. This dual-process is formally defined as:

$$\mu^*, \alpha^* = \arg\max_{\mu,\alpha} \mathbb{E}[H(I_{\text{pix}}|\alpha)] - \mathbb{E}[H(I_{\text{sem}}|\mu)], \quad (4)$$

By leveraging Eq. (4), we bridge the gap between perception inference and generative reconstruction. The proposed dual-entropy constraint effectively balances semantic stability and pixel diversity in image fusion. This unified objective is task-agnostic, thereby enabling zero-shot generalization across diverse fusion scenarios.

### 3.2. Network Architecture

Within the proposed dual-entropy collaborative constraints framework, we design two symmetric paths: a Perception Path (PP) and a Reconstruction Path (RP). The PP learns a pixel-to-semantic mapping by integrating image fusion with object detection, thereby minimizing semantic entropy and enhancing semantic representation. In contrast, the RP is trained to reconstruct occluded regions—with 50% of pixels randomly masked from each input image—thereby modeling a low-to-high-dimensional pixel mapping that maximizes pixel entropy and promotes visual diversity.

The optimization objective of the perception path can be formulated as:

$$\min_{\phi,\theta} \mathbb{E}_{x\sim\mathscr{X}} \left[ -\sum_{y} s_\phi(y|x;\theta) \log s_\phi(y|x;\theta) \right], \quad (5)$$

where $s_\phi(y|x;\theta)$ is the conditional probability of semantic variable $y$ given input $x$, parameterized by the shared encoder weights $\theta$ and perception head weights $\phi$.

The objective of the reconstruction path is to maximize pixel entropy, formulated as:

$$\max_{\psi,\theta} \mathbb{E}_{x\sim\mathscr{X}} \left[ -\sum_{f} p_{\psi}(f|x;\theta) \log p_{\psi}(f|x;\theta) \right], \quad (6)$$

where $p_{\psi}(f|x;\theta)$ is the conditional probability of the reconstructed image $f$ given input $x$, parameterized by shared encoder weights $\theta$ and reconstruction head weights $\psi$.

To decouple semantic and pixel-level features, we introduce a shared encoder based on frequency-domain dynamic perception. As shown in Figure 3, the encoder processes input images through a frequency-dynamic convolution module (GFDM) with grouped dynamic parameters. To mitigate cross-domain interference, we freeze groups irrelevant to each path during training. The resulting frequency-domain representations are then fused via cross-modal and cross-frequency integration, both implemented through dedicated frequency-domain perception convolutions.

In the perception path, encoded features are fed into dedicated perception and fusion heads to yield detection and fusion outputs. A KL-divergence constraint is applied to enforce distributional consistency between the fused result and the perception features. Similarly, in the reconstruction path, the features are processed by reconstruction and fusion heads to produce reconstructed and fused images, with a KL loss aligning the fusion output with reconstruction-oriented features.

### 3.3. Loss Function

During training, the total loss for each path integrates four key components: a supervised base loss ($\mathscr{L}_{\text{base}}$), an unsupervised information expectation loss, a trade-off loss based on KL-divergence, and an Elastic Weight Consolidation (EWC) term to mitigate catastrophic forgetting during iterative optimization (Further details of the loss functions are provided in the supplement).

The overall loss for each path is defined as:

$$\mathscr{L}_{\text{path}} = \mathscr{L}_{\text{base}} + \alpha \mathscr{L}_{\text{info}}^{\text{path}} + \beta \mathscr{L}_{\text{trade}}^{\text{path}} + \mathscr{L}_{\text{ewc}}, \quad (7)$$

where path $\in \{\text{PP}, \text{RP}\}$. $\alpha$ and $\beta$ are adjustable parameters.

The base loss combines intensity and gradient terms:

$$\mathscr{L}_{\text{base}} = \mathscr{L}_{\text{int}} + \gamma \mathscr{L}_{\text{grad}}, \quad (8)$$

where $\mathscr{L}_{\text{int}}$ enforces content consistency between the fused and source images, $\mathscr{L}_{\text{grad}}$ constrains gradient coherence, and $\gamma$ is adjustable parameter.

**Perception Path:** The information expectation loss $\mathscr{L}_{\text{info}}^{\text{PP}}$ minimizes semantic entropy via cross-entropy between detection results and ground-truth labels. The trade-off loss $\mathscr{L}_{\text{trade}}^{\text{PP}}$ enforces distributional consistency between the fused image and perception features:

$$\mathscr{L}_{\text{trade}}^{\text{PP}} = \mathscr{L}_{\text{KL}} \left( x_f, \text{Fea}_{\text{PP}} \right), \quad (9)$$

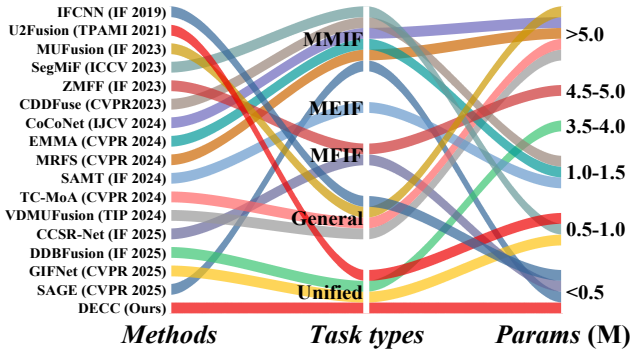


Figure 4. Comparison of task types and parameter scales between our method and the SOTA methods compared in this study. MMIF denotes Multi-Modal Image Fusion, encompassing three categories of tasks: Visible-Infrared Image Fusion, Near-Infrared and Visible Image Fusion, and Medical Image Fusion. Note that methods within each parameter group are not sorted by size.

where $Fea_{\text{PP}}$ denotes features input to the perception head.

**Reconstruction Path:** The information expectation loss $\mathscr{L}_{\text{info}}^{\text{RP}}$ promotes accurate pixel-level reconstruction. The corresponding trade-off loss $\mathscr{L}_{\text{trade}}^{\text{RP}}$ aligns the fused output with high-dimensional pixel distributions:

$$\mathscr{L}_{\text{trade}}^{\text{RP}} = \mathscr{L}_{\text{KL}} \left( x_f, \text{Fea}_{\text{RP}} \right), \quad (10)$$

where $Fea_{\text{RP}}$ represents features passed to the reconstruction head.

**EWC Regularization:** The EWC loss preserves crucial parameters from prior training phases, enhancing stability and alleviating catastrophic forgetting throughout dual-path alternating learning.

## 4. Experimental Results

We employ the M3FD dataset as the training data. In the perception path training, we incorporate semantic entropy constraints by leveraging object detection labels. For the reconstruction path, we apply random masking to infrared and visible images, occluding 50% of pixels in non-overlapping regional ranges for each modality, while additionally implementing degradation simulation on visible images to comprehensively enhance model robustness and generalization capability.

To evaluate the multi-task generalization capability of our method, we conduct extensive experiments on multiple datasets not seen during training. These include specialized fusion benchmarks: Harvard [5] (for medical image fusion, MIF), VIS-NIR Scene [53] (for near-infrared and visible image fusion, NVIF), MEFB [64] (for multi-exposure image fusion, MEIF), and MFI-WHU [60] (for multi-focus image fusion, MFIF). Furthermore, we validate downstream application performance by evaluating object detection on

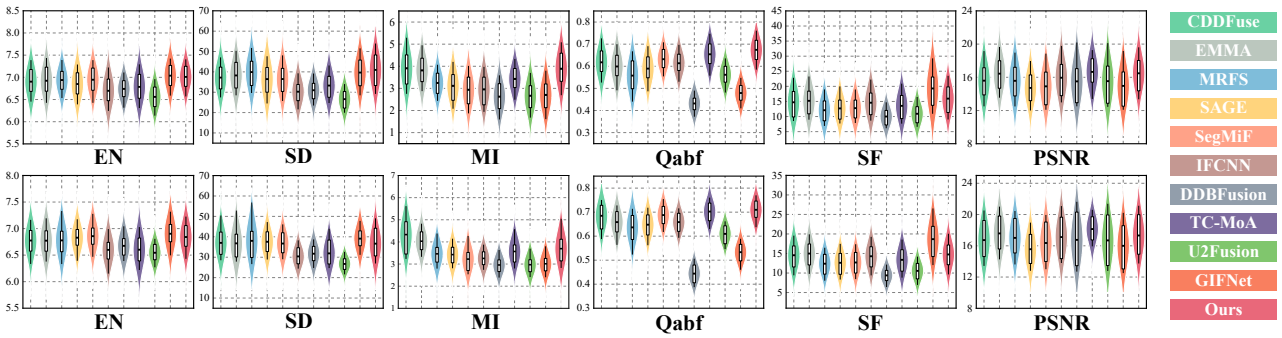


Figure 5. Visualization of quantitative metrics across different fusion methods for infrared and visible images.

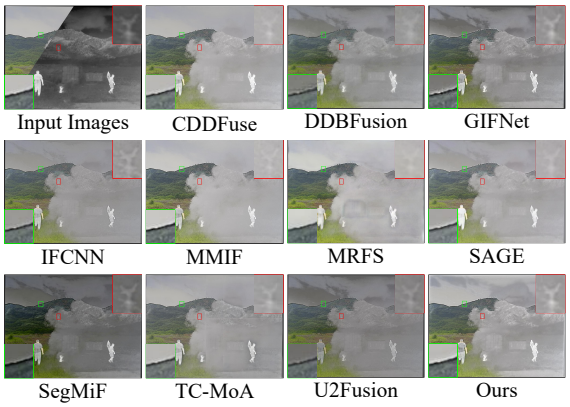


Figure 6. Qualitative comparison of different fusion methods for infrared and visible images. Key regions are zoomed in for detailed inspection.

Table 1. Performance comparison of fusion methods on multiple tasks. For each column, the optimal value is displayed in red, the second optimal in **bold**, and the third optimal underlined.

| MIF (medical image fusion) | | | | | NVIF (near-infrared and visible image fusion) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | EN | SD | MI | Qabf | Method | EN | SD | MI | Qabf |
| CoCo[26] | **6.43** | 83.98 | 2.76 | 0.57 | DDBF[67] | 7.20 | 51.39 | **4.41** | 0.60 |
| DDBF[67] | 5.58 | 76.08 | 3.07 | 0.49 | IFCN[66] | 7.23 | 55.53 | 4.34 | 0.59 |
| EMMA[71] | 5.35 | 92.57 | **2.85** | 0.58 | GIF[6] | 7.22 | **61.68** | 3.80 | 0.39 |
| GIF[6] | 5.69 | 84.08 | 2.61 | 0.43 | MUF[5] | 7.36 | 59.04 | 3.18 | 0.42 |
| MUF[5] | 6.21 | 79.60 | 2.84 | **0.65** | TCM[74] | 4.39 | 48.56 | 4.39 | 0.67 |
| TCM[74] | 5.78 | 84.29 | 2.83 | 0.66 | U2F[52] | 7.20 | 51.53 | 4.06 | 0.56 |
| U2F[52] | 5.55 | 58.49 | 2.66 | 0.41 | SegM[27] | 7.17 | 61.71 | 4.08 | 0.56 |
| DECC (Ours) | 7.10 | **85.52** | 2.74 | 0.66 | DECC (Ours) | **7.26** | 56.41 | 4.63 | **0.62** |
| MFIF (multi-focus image fusion) | | | | | MEIF (multi-exposure image fusion) | | | | |
| Method | EN | SD | VIFF | SCD | Method | EN | SD | VIFF | SCD |
| DDBF[67] | 7.30 | **52.47** | 0.67 | 0.89 | VDMU[39] | 6.93 | 47.26 | 0.49 | 0.50 |
| IFCN[66] | 7.31 | 51.89 | 0.95 | 0.41 | U2F[52] | 6.66 | 45.43 | 0.62 | 0.46 |
| CCSR[73] | 7.31 | 52.37 | 0.95 | 0.53 | TCM[74] | 7.04 | 47.18 | 0.78 | 0.33 |
| MDLS[49] | **7.32** | 52.34 | 0.95 | 0.53 | SAMT[51] | 7.05 | 50.96 | 0.75 | 0.64 |
| TCM[74] | 7.28 | 51.11 | 0.80 | 0.42 | MUF[5] | 7.02 | 54.96 | 0.77 | 0.91 |
| U2F[52] | 7.12 | 47.94 | 0.82 | 0.14 | IFCN[66] | **7.13** | 59.42 | 1.00 | **0.99** |
| ZMFF[10] | 7.29 | 51.74 | 0.80 | 0.32 | GIF[6] | 7.06 | 67.33 | **0.86** | **1.23** |
| DECC (Ours) | 7.37 | 57.51 | 0.84 | **0.88** | DECC (Ours) | 7.20 | **64.26** | 0.78 | 1.25 |

the M3FD dataset and semantic segmentation on the FMB dataset, using officially released pretrained models for both tasks. The parameter sizes and supported task types of all compared methods are summarized in the Figure 4.

### 4.1. Infrared and visible image fusion

In the VIF task, we compare our method to ten state-of-the-art fusion algorithms on both the M3FD and FMB datasets. Quantitative evaluation employs no-reference metrics, including entropy (EN) [38], standard deviation (SD) [8], and spatial frequency (SF) [8], as well as reference-based metrics such as mutual information (MI) [61], fusion quality ($Q_{abf}$) [61], and peak signal-to-noise ratio (PSNR) [50]. As illustrated in Figure 5, our method achieves significantly higher EN than other approaches, reflecting its ability to preserve richer visual information. Superior performance across other metrics further confirms its advantages in retaining structural details and improving signal-to-noise ratio. Qualitative results in Figure 6 show that our method maintains high visual fidelity, with consistent texture rendering in regions such as sky and grass, while preserving more complete infrared information in foggy conditions compared to other methods.

### 4.2. Unseen task generalization

To assess cross-modality generalization, we further evaluate our approach on medical image fusion (MIF) and visible-near infrared image fusion (NVIF) tasks. As summarized in Table 1, our method achieves notable advantages in both visual richness and structural fidelity on the MIF task. On the NVIF task, DECC effectively preserves complementary information while maintaining balanced modality representation, demonstrating robust cross-modal generalization capability.

Furthermore, our method demonstrates strong performance in digital photography fusion tasks. As shown in Table 1, for both multi-focus and multi-exposure fusion scenarios, DECC achieves superior performance in pixel-level information retention, contrast preservation, and structural consistency with source images. Qualitative results in Figure 7 illustrate the method's robustness under varying exposure and focus conditions, where it effectively preserves fine textures, mitigates undesirable brightness variations, and produces naturally rendered depth-of-field effects.

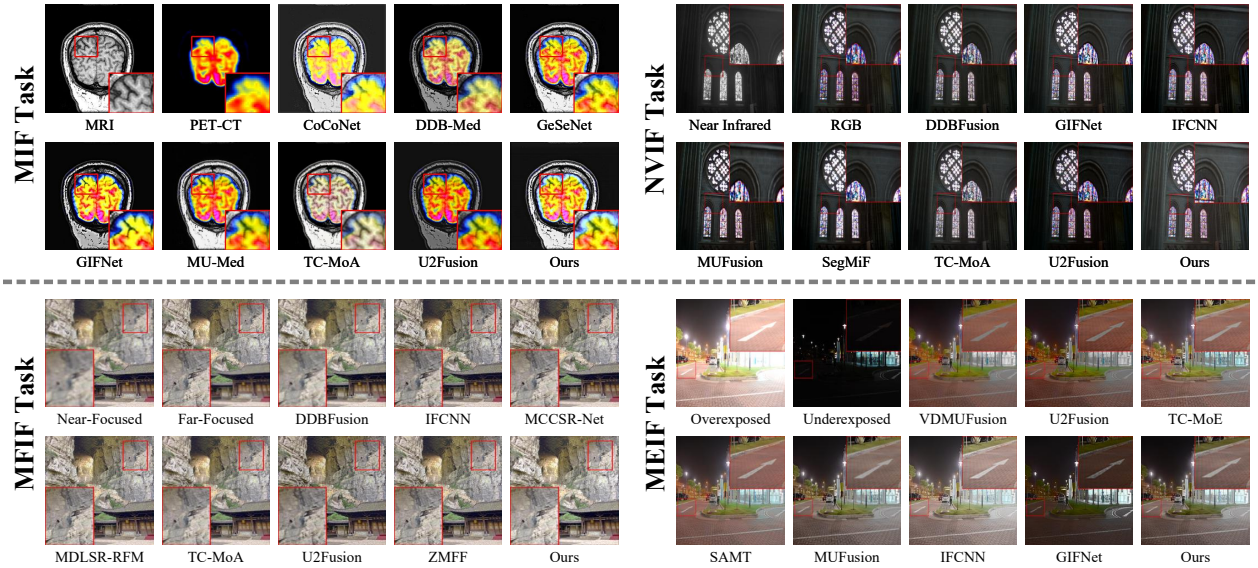


Figure 7. Qualitative comparison of different fusion methods across various fusion tasks.

Table 2. Object detection performance comparison of different fusion methods on the M3FD dataset. The models (YOLOv8 and YOLOv11) are evaluated by Average Precision (AP) for Person ($AP_P$), Car ($AP_C$), and the mean Average Precision (mAP) at different IoU thresholds. For each column, the optimal value is displayed in red, the second optimal in **bold**, and the third optimal underlined.

| Methods | Detector: YOLOv8 with officially pretrained weights | | | | | | | | | | Detector: YOLOv11 with officially pretrained weights | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.5 | | | 0.65 | | | 0.8 | | | *mAP@* | 0.5 | | | 0.65 | | | 0.8 | | | *mAP@* |
| | $AP_P$ | $AP_C$ | *mAP* | $AP_P$ | $AP_C$ | *mAP* | $AP_P$ | $AP_C$ | *mAP* | [0.5 : 0.95] | $AP_P$ | $AP_C$ | *mAP* | $AP_P$ | $AP_C$ | *mAP* | $AP_P$ | $AP_C$ | *mAP* | [0.5 : 0.95] |
| CDDFuse[69] | 49.42 | 64.61 | 57.02 | 37.27 | 52.18 | 44.72 | 20.05 | 32.81 | 26.43 | 33.53 | 49.80 | 68.31 | 59.05 | 38.71 | 54.58 | 46.64 | 20.96 | 35.84 | 28.40 | 34.87 |
| EMMA[71] | 47.16 | 64.11 | 55.63 | 35.61 | 51.72 | 43.67 | 18.81 | 32.50 | 25.66 | 32.81 | 50.31 | 66.02 | 58.17 | 37.68 | 53.12 | 45.40 | 20.42 | 35.56 | 27.99 | 34.15 |
| MRFS[62] | 43.78 | 64.67 | 54.22 | 34.09 | 51.37 | 42.73 | 18.02 | 32.56 | 25.29 | 32.12 | 44.76 | 67.39 | 56.08 | 34.37 | 54.47 | 44.42 | 17.68 | 34.02 | 25.85 | 32.84 |
| SAGE[51] | 46.82 | 64.82 | 55.82 | 36.18 | 52.09 | 44.13 | 19.80 | 33.10 | 26.45 | 32.98 | 50.16 | 67.39 | 58.78 | 38.11 | 54.14 | 46.12 | **21.02** | 34.94 | 27.98 | 34.76 |
| SegMiF[27] | **50.65** | 66.40 | 58.52 | **38.83** | 53.43 | 46.13 | 20.47 | 34.05 | 27.26 | 34.53 | 51.85 | 69.13 | 60.49 | **40.19** | 54.48 | 47.33 | 21.57 | 37.57 | **29.57** | 36.00 |
| IFCNN[66] | 49.57 | 62.87 | 56.22 | 37.75 | 50.74 | 44.25 | **20.86** | 32.91 | 26.89 | 33.28 | 50.82 | 66.49 | 58.65 | 39.87 | 53.42 | 46.64 | 20.69 | 36.78 | 28.73 | 35.02 |
| DDBFusion[67] | 50.61 | 64.20 | 57.41 | 38.72 | 51.52 | 45.12 | 19.94 | 32.31 | 26.13 | 33.70 | 51.42 | 67.27 | 59.35 | 39.87 | 54.29 | 47.08 | 19.96 | 37.09 | 28.52 | 35.30 |
| TC-MoA[74] | 50.55 | 68.17 | **59.36** | 38.44 | **54.33** | **46.38** | 20.39 | 34.32 | 27.35 | **34.76** | 52.29 | **70.24** | **61.27** | 39.58 | **56.69** | **48.14** | 20.71 | 38.09 | 29.40 | **36.25** |
| U2Fusion[52] | 48.99 | 66.09 | 57.54 | 36.84 | 53.20 | 45.02 | 19.96 | **34.90** | **27.43** | 33.98 | 51.03 | 68.52 | 59.77 | 38.77 | 55.42 | 47.09 | 20.49 | **38.25** | 29.37 | 35.57 |
| GIFNet[6] | 50.49 | 65.33 | 57.91 | 37.97 | 52.65 | 45.31 | 20.21 | 34.47 | 27.34 | 34.06 | 53.97 | 68.01 | 60.99 | 40.95 | 54.72 | 47.84 | 20.96 | 36.61 | 28.79 | 35.82 |
| DECC (Ours) | 52.52 | **67.83** | 60.17 | 39.64 | 55.34 | 47.49 | 21.05 | 36.34 | 28.70 | 35.34 | **53.82** | 70.63 | 62.22 | 39.73 | 57.93 | 48.83 | 20.89 | 38.72 | 29.81 | 36.85 |

### 4.3. Task-driven downstream applications

We assess the object detection performance on fused images from the M3FD dataset with publicly available pretrained models of YOLOv8 and YOLOv11. As shown in Table 2, our method achieves the best detection precision for both Person and Car categories across different IoU thresholds.

For semantic segmentation evaluation, we employ officially released pretrained MaskFormer and DPT models on fusion outputs from the FMB dataset. Table 3 demonstrates that our method yields significant improvements across most semantic categories, achieving the highest mean Intersection over Union (mIoU) scores. Visualizations in Figure 8 show this consistent quality of segmentation for all day-and-night scenarios, therefore demonstrating that our fused images retain rich visual information and significantly enhance subsequent perceptual performance. These extensive tests provide extra evidence of the robustness and real-world generalization of our methods.

### 4.4. Ablation experiments

We perform comprehensive ablation studies on key components, including the Perception Path, the Reconstruction Path, the trade-off loss, base loss, and the Grouped Frequency Dynamic Convolution Module (GFDM). As shown in Figure 9, when trained solely with the Perception Path objective ("only PP"), the model maintains clear pedestrian contours. Even when further removing the base loss ("only $\mathscr{L}_{info}^{PP} + \mathscr{L}_{trade}^{PP}$"), essential semantic information remains effectively preserved. Conversely, models trained exclusively with the Reconstruction Path objective ("only RP") produce results with enhanced detail preservation. The variant "only $\mathscr{L}_{info}^{RP} + \mathscr{L}_{trade}^{RP}$" retains rich structural information despite the absence of base losses. Importantly, the absence of trade-off loss ("w/o $\mathscr{L}_{trade}$") leads to simultaneous degradation in both contour precision and background details. Furthermore, replacing the frequency dynamic convolution with conventional convolution ("w/o GFDM") causes sub-

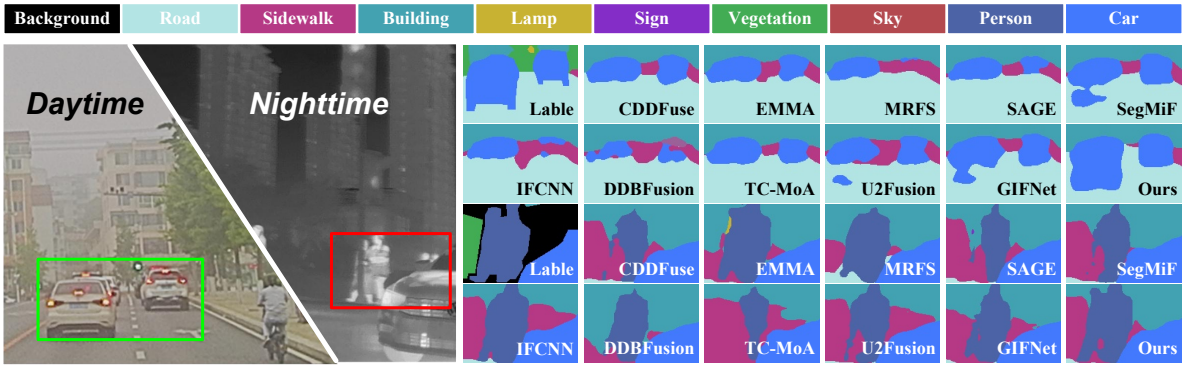


Figure 8. Qualitative comparison of different fusion methods for semantic segmentation on the FMB dataset. Results are generated using MaskFormer, showcasing the segmentation performance under both daytime and nighttime conditions.

Table 3. Semantic segmentation performance comparison of different fusion methods on the FMB dataset. The models (Mask2Former [3] and DPT [36]) are evaluated by mean Intersection over Union (mIoU) and class-wise IoU. For each column, the optimal value is displayed in red, the second optimal in **bold**, and the third optimal underlined.

| Detector: Mask2Former with officially pretrained weights | | | | | | | |
|---|---|---|---|---|---|---|---|
| Methods | Road | Sidewalk | Building | Sky | Person | Car | mIoU |
| CDDFuse[69] | 71.77 | 26.91 | 55.89 | 89.10 | 19.00 | 62.25 | 54.15 |
| EMMA[71] | 71.36 | 27.48 | 56.28 | 88.54 | 19.35 | 61.80 | 54.13 |
| MRFS[62] | 70.28 | 27.24 | 55.63 | 88.35 | 17.19 | 60.43 | 53.19 |
| SAGE[51] | 71.50 | 27.79 | **57.35** | 89.09 | 19.09 | 59.57 | 54.06 |
| SegMiF[27] | **72.33** | 28.35 | 57.92 | 89.03 | 19.37 | 62.68 | 54.95 |
| IFCNN[66] | 71.47 | 27.28 | 57.16 | 89.67 | 19.34 | 61.61 | 54.42 |
| DDBFusion[67] | 70.71 | 26.90 | 56.16 | 88.80 | 17.52 | 60.68 | 53.46 |
| TC-MoA[74] | 72.15 | 27.62 | 56.62 | **89.47** | 17.50 | 62.78 | 54.36 |
| U2Fusion[52] | 72.22 | 27.46 | 57.15 | 89.36 | 19.80 | 63.00 | 54.83 |
| GIFNet[6] | 72.07 | 27.41 | 56.86 | 88.82 | 22.31 | **64.21** | **55.28** |
| DECC (Ours) | 72.56 | **27.79** | 57.09 | 89.33 | **20.81** | 64.75 | 55.39 |
| Detector: DPT with officially pretrained weights | | | | | | | |
| Methods | Road | Sidewalk | Building | Sky | Person | Car | mIoU |
| CDDFuse[69] | **66.42** | 23.09 | 48.82 | 86.24 | 10.94 | 55.81 | 48.56 |
| EMMA[71] | 65.32 | **23.11** | 49.84 | 85.89 | 12.04 | 54.95 | 48.52 |
| MRFS[62] | 61.56 | 22.15 | **51.78** | 86.14 | 11.42 | 51.81 | 47.48 |
| SAGE[51] | 64.22 | 21.89 | 50.23 | 86.20 | 10.06 | 52.34 | 47.49 |
| SegMiF[27] | 65.25 | 22.94 | 48.74 | 85.41 | 12.11 | 55.07 | 48.25 |
| IFCNN[66] | 62.85 | 22.08 | 47.85 | 85.24 | 10.45 | 50.88 | 46.56 |
| DDBFusion[67] | 63.88 | 21.96 | 48.56 | 86.24 | 11.70 | 52.16 | 47.42 |
| TC-MoA[74] | 65.73 | 23.05 | 50.87 | 85.97 | 10.20 | **56.96** | **48.80** |
| U2Fusion[52] | 64.02 | 22.93 | 51.73 | 86.79 | 11.37 | 51.49 | 48.05 |
| GIFNet[6] | 65.26 | 21.40 | 52.00 | 86.04 | **12.74** | 54.87 | 48.72 |
| DECC (Ours) | 66.68 | 23.19 | 49.86 | **86.29** | 12.81 | 57.47 | 49.38 |

Table 4. Quantitative results of ablation studies on the M3FD dataset, including vision-related metrics and object detection metrics. For each column, the optimal value is displayed in red, the second optimal in **bold**, and the third optimal underlined.

| Methods | MI | Qabf | EN | SD | SF | PSNR | $mAP_{v8}$ | $mAP_{v11}$ |
|---|---|---|---|---|---|---|---|---|
| only PP | 4.25 | 0.65 | 6.96 | 39.62 | 15.47 | 15.89 | 34.28 | 35.66 |
| only RP | 3.20 | 0.45 | 7.07 | 56.92 | 17.67 | 11.39 | 28.93 | 31.36 |
| w/o GFDM | 3.10 | 0.61 | 6.49 | 27.67 | 11.26 | 13.28 | 32.76 | 34.17 |
| w/o $L_{trade}$ | **4.19** | **0.65** | **7.06** | 40.21 | 15.09 | **16.02** | **34.61** | **36.01** |
| DECC (Ours) | 3.89 | 0.67 | 7.01 | **40.97** | **15.94** | 16.48 | 36.34 | 36.85 |

stantial deterioration in visual quality. Quantitative results

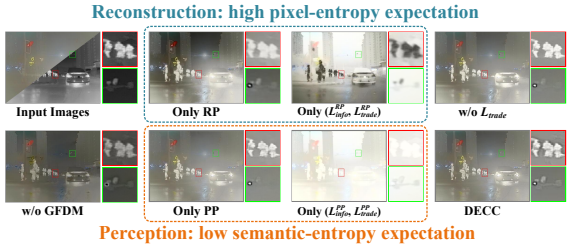


Figure 9. Qualitative comparison of ablation experiments.

in Table 4 further corroborate the indispensable role of each component in achieving optimal fusion performance.

## 5. Conclusion

This work tackles two core challenges in image fusion: poor generalization and the semantics-pixels trade-off. We propose a unified framework based on the free-energy principle, combining semantic perception and detail reconstruction paths with frequency-domain feature decoupling. Trained only on infrared-visible pairs without task-specific design, our method achieves excellent semantics-pixels balance, strong zero-shot generalization, high parameter efficiency, and significant downstream task improvement. This framework provides an effective solution to fundamental fusion problems and offers insights for future research.

## Acknowledgement

This work is supported by the National Key Research and Development Program of China under Grant 31400, the Program of High Resolution Earth Observation, China under Grant 01-Y30F05-9001-20/22 and GFZX0404130307, the Key Research and Transformation project of Qinghai Province, China under Grant 2022-SF-150.

## References


[1] Jun Chen, Liling Yang, Wei Yu, Wenping Gong, Zhanchuan Cai, and Jiayi Ma. Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing*, 34:3139–3153, 2025. 3

[2] Linwei Chen, Lin Gu, Liang Li, Chenggang Yan, and Ying Fu. Frequency dynamic convolution for dense image prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30178–30188, 2025. 2

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 8

[4] Chunyang Cheng, Xiao-Jun Wu, Tianyang Xu, and Guoyang Chen. Unifusion: A lightweight unified image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 2

[5] 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. 1, 5, 6

[6] 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 (CVPR)*, pages 28102–28112, 2025. 2, 3, 6, 7, 8

[7] Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Hui Li, Xi Li, Zhangyong Tang, and Josef Kittler. Textfusion: Unveiling the power of textual semantics for controllable image fusion. *Information Fusion*, 117:102790, 2025. 3

[8] A.M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 6

[9] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. 2, 3

[10] Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. Zmff: Zero-shot multi-focus image fusion. *Information Fusion*, 92:127–138, 2023. 6

[11] Huan Kang, Hui Li, Tianyang Xu, Xiao-Jun Wu, Rui Wang, Chunyang Cheng, and Josef Kittler. Smlnet: A spd manifold learning network for infrared and visible image fusion. *International Journal of Computer Vision*, pages 1–22, 2025. 3

[12] Huan Kang, Hui Li, Xiao-Jun Wu, Tianyang Xu, Rui Wang, Chunyang Cheng, and Josef Kittler. Grformer: A novel transformer on grassmann manifold for infrared and visible image fusion. *Information Fusion*, 125:103402, 2026. 1, 3

[13] Shahid Karim, Geng Tong, Jinyang Li, Akeel Qadir, Umar Farooq, and Yiting Yu. Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, 90:185–217, 2023. 1

[14] Hui Li and Xiao-Jun Wu. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 1

[15] Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024. 3

[16] Huafeng Li, Yueliang Cen, Yu Liu, Xun Chen, and Zhengtao Yu. Different Input Resolutions and Arbitrary Output Resolution: A Meta Learning-Based Deep Framework for Infrared and Visible Image Fusion. *IEEE Transactions on Image Processing*, 30:4070–4083, 2021. 1

[17] 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. 1

[18] Huafeng Li, Junzhi Zhao, Jinxing Li, Zhengtao Yu, and Guangming Lu. Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion. *Information Fusion*, 95:26–41, 2023. 3

[19] Jing Li, Jianming Zhu, Chang Li, Xun Chen, and Bin Yang. CGTF: Convolution-Guided Transformer for Infrared and Visible Image Fusion. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 3

[20] Jing Li, Bin Yang, Lu Bai, Hao Dou, Chang Li, and Lingfei Ma. Tfiv: Multigrained token fusion for infrared and visible image via transformer. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023. 3

[21] Jing Li, Lu Bai, Bin Yang, Chang Li, and Lingfei Ma. Graph representation learning for infrared and visible image fusion. *IEEE Transactions on Automation Science and Engineering*, 22:13801–13813, 2025. 3

[22] Jing Li, Lu Bai, Bin Yang, Chang Li, Lingfei Ma, Lixin Cui, and Edwin R. Hancock. Dual-modal prior semantic guided infrared and visible image fusion for intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems*, 26(7):9767–9780, 2025. 3

[23] Jiayang Li, Junjun Jiang, Pengwei Liang, Jiayi Ma, and Liqiang Nie. Maefuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training. *IEEE Transactions on Image Processing*, 34:1340–1353, 2025. 3

[24] Jing Li, Yifan Wang, Jiafeng Yan, Renlong Zhang, and Bin Yang. Mdaif: Robust one-stop multi-degradation-aware image fusion with language-driven semantics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6226–6234, 2026. 3

[25] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5792–5801, New Orleans, LA, USA, 2022. IEEE. 3

[26] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, pages 1–28, 2023. 6

[27] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023. 6, 7, 8

[28] 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*, 12(3):502–515, 2025. 3

[29] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025. 1

[30] 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 (CVPR)*, pages 2226–2235, 2025. 3

[31] Xiaowen Liu, Hongtao Huo, Jing Li, Shan Pang, and Bowen Zheng. A semantic-driven coupled network for infrared and visible image fusion. *Information Fusion*, 108:102352, 2024. 3

[32] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016. 3

[33] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48: 11–26, 2019. 1

[34] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 3

[35] 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

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168, 2021. 8

[37] Yujing Rao, Dan Wu, Mina Han, Ting Wang, Yang Yang, Tao Lei, Chengjiang Zhou, Haicheng Bai, and Lin Xing. AT-GAN: A generative adversarial network with attention and transition for infrared and visible image fusion. *Information Fusion*, 92:336–349, 2023. 3

[38] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 6

[39] Yu Shi, Yu Liu, Juan Cheng, Z. Jane Wang, and Xun Chen. Vdmufusion: A versatile diffusion model-based unsupervised framework for image fusion. *IEEE Transactions on Image Processing*, 34:441–454, 2025. 6

[40] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022. 3

[41] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 3

[42] 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*, 83-84:79–92, 2022. 1

[43] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023. 1

[44] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99:101870, 2023. 3

[45] 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*, pages 1–18, 2025. 3

[46] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. DATFuse: Infrared and Visible Image Fusion via Dual Attention Transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023. 3

[47] Vibashan Vs, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M. Patel. Image Fusion Transformer. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3566–3570, Bordeaux, France, 2022. IEEE. 3

[48] Di Wang, Jinyuan Liu, Risheng Liu, and Xin Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98: 101828, 2023. 3

[49] Jiwei Wang, Huaijing Qu, Zhisheng Zhang, and Ming Xie. New insights into multi-focus image fusion: A fusion method based on multi-dictionary linear sparse representation and region fusion model. *Information Fusion*, 105: 102230, 2024. 6

[50] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[51] 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. 6, 7, 8

[52] 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*, 2020. 2, 6, 7, 8

[53] Han Xu, Jiteng Yuan, and Jiayi Ma. MURF: Mutually Reinforcing Multi-Modal Image Registration and Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. 1, 5

[54] Han Xu, Xunpeng Yi, Chen Lu, Guangcan Liu, and Jiayi Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 34:5803–5818, 2025. 1

[55] Bin Yang, Yuxuan Hu, Xiaowen Liu, and Jing Li. Cefusion: An infrared and visible image fusion network based on cross-modal multi-granularity information interaction and edge guidance. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):17794–17809, 2024. 3

[56] 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 (CVPR)*, pages 27026–27035, 2024. 1

[57] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex++: Retinex-driven reinforced diffusion model for low-light image enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6823–6841, 2025. 3

[58] 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*, 2025. 3

[59] Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi Ma. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32:5705–5720, 2023. 1

[60] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021. 5

[61] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1, 6

[62] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26974–26983, 2024. 7, 8

[63] Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025. 1, 3

[64] Xingchen Zhang. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, pages 111–131, 2021. 5

[65] Xingchen Zhang and Yiannis Demiris. Visible and Infrared Image Fusion Using Deep Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023. 1

[66] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118, 2020. 1, 6, 7, 8

[67] Zeyang Zhang, Hui Li, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler. DDBFusion: An unified image decomposition and fusion framework based on dual decomposition and Bézier curves. *Information Fusion*, 114:102655, 2025. 6, 7, 8

[68] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding from Object Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13955–13965, Vancouver, BC, Canada, 2023. IEEE. 3

[69] Zixiang Zhao, Haowen Bai, Jiangshe 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5906–5916, Vancouver, BC, Canada, 2023. IEEE. 7, 8

[70] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe 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 (ICCV)*, pages 8082–8093, 2023. 1

[71] Zixiang Zhao, Haowen Bai, Jiangshe 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 (CVPR)*, 2024. 6, 7, 8

[72] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1

[73] Kecheng Zheng, Juan Cheng, and Yu Liu. Unfolding coupled convolutional sparse representation for multi-focus image fusion. *Information Fusion*, 118:102974, 2025. 6

[74] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7099–7108, 2024. 1, 6, 7, 8