

Task-driven Image Fusion with Learnable Fusion Loss

Haowen Bai¹ Jiangshe Zhang^{1*} Zixiang Zhao^{2*} Yichen Wu³
 Lilun Deng¹ Yukun Cui¹ Tao Feng⁴ Shuang Xu⁵

¹Xi'an Jiaotong University ²ETH Zürich ³City University of Hong Kong

⁴Tsinghua University ⁵Northwestern Polytechnical University

hwbaii@stu.xjtu.edu.cn

Abstract

*Multi-modal image fusion aggregates information from multiple sensor sources, achieving superior visual quality and perceptual features compared to single-source images, often improving downstream tasks. However, current fusion methods for downstream tasks still use predefined fusion objectives that potentially mismatch the downstream tasks, limiting adaptive guidance and reducing model flexibility. To address this, we propose Task-driven Image Fusion (**TD-Fusion**), a fusion framework incorporating a learnable fusion loss guided by task loss. Specifically, our fusion loss includes learnable parameters modeled by a neural network called the loss generation module. This module is supervised by the downstream task loss in a meta-learning manner. The learning objective is to minimize the task loss of fused images after optimizing the fusion module with the fusion loss. Iterative updates between the fusion module and the loss module ensure that the fusion network evolves toward minimizing task loss, guiding the fusion process toward the task objectives. TDFusion's training relies entirely on the downstream task loss, making it adaptable to any specific task. It can be applied to any architecture of fusion and task networks. Experiments demonstrate TDFusion's performance through fusion experiments conducted on four different datasets, in addition to evaluations on semantic segmentation and object detection tasks. The code is available at <https://github.com/HaowenBai/TDFusion>.*

1. Introduction

Multi-modal image fusion [25, 31, 51, 59, 72, 79, 82] combines information from multiple sensors to produce a more holistic and detailed representation. Infrared images capture thermal radiation regardless of lighting conditions, while visible images provide richer texture details. Fused im-

ages enhance downstream tasks through improved information density and robustness [18, 41, 45, 62, 63, 70, 71], outperforming single-modal inputs in semantic segmentation [13, 28, 34, 56], object detection [8, 32], and other related applications [1, 3, 20, 27]. Conventional methods typically treat fusion as image restoration using unsupervised loss [76, 83, 87–89] or perceptual loss [21, 23, 77]. These approaches prioritize visual-level fusion through predefined aggregation objectives, often neglecting semantic feature extraction. This limitation hinders scene interpretation and task performance [14, 24, 52, 64]. Recent advances explore the mutual enhancement between fusion and downstream tasks [37]. By cascading the fusion network with downstream task network [39, 49], the task loss constrains the fusion learning, ensuring the fused images meet the task requirements [32, 56]. Alternatively, some methods incorporate high-level visual task features [34, 76, 78] or focus on learning optimal initializations [40] to enhance fusion.

While integrating downstream tasks, existing frameworks still rely on *predefined* fusion loss terms lacking dynamic adaptation. The impact of downstream tasks remains limited due to specific combinations. Manually defined losses preserve predefined guidance, frequently overlooking task-specific requirements. This guidance imposes manually designed prior constraints on the fusion process, limiting the dynamic and adaptive influence of downstream tasks on specific image pairs. These approaches, whether incorporating task features [34, 78] or employing task losses [32, 56], still face the limitations of fixed fusion loss terms. Task-specialized networks [76] create fusion-task dependencies, restricting flexibility and limiting their applicability to various high-level vision tasks. We address these limitations through a task-driven framework with learnable loss. The fusion loss contains learnable parameters, generated by a loss generation module, and is designed to retain the intensity information of the source images for specific downstream tasks. The purpose of updating the fusion loss is to guide the fusion network in generating fused images that minimize downstream task loss,

*Corresponding authors.

thereby enhancing adaptability. Moreover, the fusion loss update relies on the downstream task loss, making it independent of any specific task or network architecture.

The loss generation module produces fusion losses for subsequent fusion module updates. This complexity presents challenges to standard end-to-end training. Fortunately, meta-learning techniques, which are strategies for learning how to learn, can effectively achieve the learning objectives of the loss generation module. This involves minimizing task loss for fused images through an optimized fusion loss. Meta-learning tackles common deep-learning challenges like limited data, high computational costs, and the need for better generalization. Core optimization areas include parameter tuning [11], optimization strategies [26], and network architectures search [30]. In this paper, we draw inspiration from the Model-Agnostic Meta-Learning (MAML) approach [11] to train our loss generation module. Specifically, training the loss generation module involves two stages: inner updates and outer updates. During inner updates, the output of the loss generation module updates a surrogate fusion module without altering the original parameters. During outer updates, the fused image from the surrogate module is fed into the task network. The resulting task loss then updates the loss generation module through backpropagation. This alternating training ensures that the loss generation module consistently produces fusion losses that minimize the downstream task loss of the fused images.

This paper introduces TDFusion, a task-oriented fusion framework driven by downstream tasks. It consists of a fusion module, a task module, and a loss generation module that learns to optimize the fusion loss. The fusion loss incorporates intensity preferences from source images and gradient preservation, guided by downstream task loss to refine intensity preferences. This model follows the general form of fusion loss used in advanced methods [76, 83, 88, 89], ensuring adaptability to various tasks. The updates of the loss generation module are performed using a meta-learning approach, optimizing the loss generation module parameters based on task loss from the fused images after each update of the fusion module. This process ensures that the loss function guides the fusion module continuously, optimizing feature aggregation and minimizing downstream task losses. The loss generation module dynamically adjusts through alternating updates with the fusion and task modules, generating optimal fusion losses at each model state. Our contributions can be summarized as follows:

- We propose TDFusion, a meta-learning-based fusion framework that leverages the loss functions of downstream tasks for training. This method promotes task-driven fusion and alleviates challenges caused by the absence of ground truth. Moreover, this framework is agnostic of specific downstream tasks or network architectures, which enhances its adaptability and flexibility.

- Our framework includes a dynamically updated, learnable fusion loss generation module. It selectively extracts source image information, minimizing the loss in downstream tasks. This ensures optimal fusion performance while maximizing adaptability to downstream tasks.
- We analyze the information preferences of downstream tasks such as semantic segmentation and object detection, providing deeper insights into multi-modal high-level vision tasks.
- TDFusion achieves outstanding performance in both fusion and high-level vision tasks, validated on four fusion datasets with semantic segmentation and object detection.

2. Related Work

2.1. Deep learning-based Image Fusion

Deep learning-based image fusion methods have revolutionized the field by exploiting the powerful feature extraction capabilities of neural networks [4, 6, 33, 35, 36, 42–44, 65, 77]. These methods are broadly categorized into discriminative and generative approaches. Discriminative methods [10, 81, 85, 88] leverage the strong reconstruction ability of neural networks to directly learn the mapping between source and fused images [21, 23, 80, 82]. Generative methods, on the other hand, model the image generation process using generative approaches, integrating source images from a distributional perspective. These include methods based on Generative Adversarial Networks [46–48, 87] and diffusion models [73, 84]. Unified fusion methods [66, 75, 89] bridge the gap between different fusion sub-tasks, incorporating strategies such as continual learning [67] and self-supervised decomposition techniques [29]. The introduction of registration modules helps to mitigate misalignment issues [16, 60, 68] in source images. Recent studies have further explored the synergy between fusion and high-level vision tasks. These include leveraging downstream task losses to optimize fusion networks [32, 56], embedding high-level task features [34, 76, 78], and employing initialization techniques [40].

2.2. Meta-Learning in Vision

Meta-learning develops algorithms to automatically fine-tune hyperparameters for specific tasks, showing its versatility and effectiveness across various domains. MAML [11] and its variations [12, 50, 53] focus on learning efficient initialization parameters to quickly adapt to new tasks using minimal data. Meta-SGD [26] extends MAML by learning optimal update directions and rates, beneficial in few-shot learning scenarios. Other approaches like MW-Net [55] and L2RW [54] emphasize selecting relevant sample weights to tackle noisy data using a compact validation set. Additionally, some studies focus on improving model adaptability through learning loss functions [2, 7, 15].

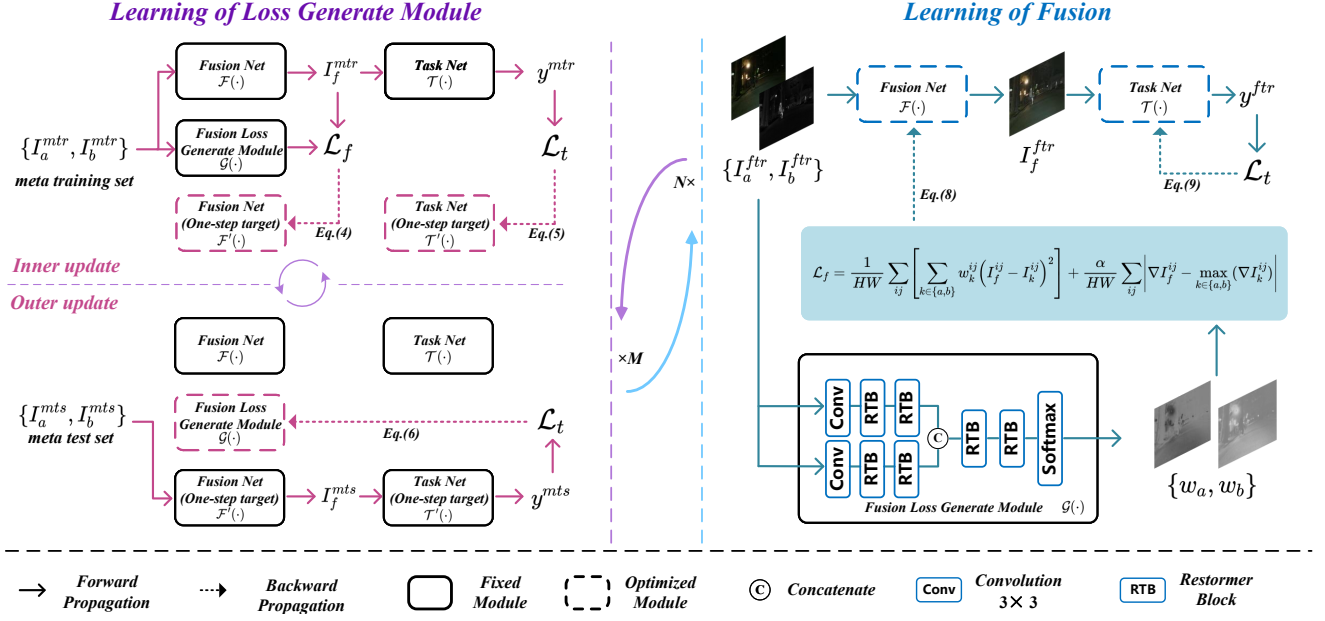


Figure 1. The TDFusion workflow alternates between training the loss generation module and the fusion module. Training of the loss generation module involves both inner and outer updates, learned through meta-learning.

In image fusion, learnable filters [22] enable the fusion of images at arbitrary resolutions. MetaFusion [78] introduces a mechanism that improves image fusion and object detection by aligning semantics with fusion-specific features. ReFusion [5] guides the learnable fusion loss for various fusion tasks by reconstructing the source images. Meta-learning also supports neural architecture search [38, 40] to identify optimal network architectures for image fusion and customizes network initialization for various tasks [40].

2.3. Comparison with Existing Approaches

We propose a novel image fusion method tailored for downstream tasks, leveraging a learnable fusion loss driven by task-specific objectives. Our method employs a meta-learning algorithm that alternates between inner and outer updates, allowing the downstream task loss to guide the optimization of learnable fusion parameters. This results in fused images that minimize the downstream task loss, enhancing their adaptability across various tasks. Unlike previous methods, our approach develops a task-specific fusion loss, shifting focus from traditional factors such as resolution and network structure, and avoiding reliance on predefined fusion loss terms. This renders our fusion framework more flexible and applicable to various scenarios.

3. Method

3.1. Overview

Our TDFusion framework, as shown in Fig. 1, consists of a fusion network $\mathcal{F}(\cdot)$, a downstream task network $\mathcal{T}(\cdot)$, and a fusion loss generation module $\mathcal{G}(\cdot)$, which produces

parameters for a learnable loss function. The parameters of these modules are denoted as $\theta_{\mathcal{F}}$, $\theta_{\mathcal{T}}$, and $\theta_{\mathcal{G}}$, respectively. The one-step updated clones of \mathcal{F} and \mathcal{T} are denoted as \mathcal{F}' and \mathcal{T}' , with parameters $\theta_{\mathcal{F}'}$ and $\theta_{\mathcal{T}'}$. During the updates, the fusion network and the loss generation module alternate in learning, as depicted by blue and purple in Fig. 1. The update of the loss generation module consists of inner and outer updates, detailed in the following subsections. \mathcal{L}_f and \mathcal{L}_t represent the learnable fusion loss and task-specific loss, with their formulations provided in the next section.

3.2. Loss Function

The learnable fusion loss \mathcal{L}_f consists of the intensity term and the gradient term. The intensity term is defined by the output of the loss generation module $\{w_a, w_b\} = \mathcal{G}(I_a, I_b)$, where w_a and w_b control the intensity preference in the fusion loss. The Softmax function in the loss generation module ensures $w_a^{ij} + w_b^{ij} = 1$ for each pixel, thereby selectively retaining the intensity information from the source images. The gradient term emphasizes higher gradient values from the input images [76, 83, 89], aiming to preserve maximal information from the source. The detailed formulation of the learnable fusion loss is as follows:

$$\mathcal{L}_f = \mathcal{L}_f^{int} + \alpha \mathcal{L}_f^{grad}, \quad (1)$$

$$\mathcal{L}_f^{int} = \frac{1}{HW} \sum_{ij} \left[\sum_{k \in \{a,b\}} w_k^{ij} (I_f^{ij} - I_k^{ij})^2 \right], \quad (2)$$

$$\mathcal{L}_f^{grad} = \frac{1}{HW} \sum_{ij} \left| \nabla I_f^{ij} - \max_{k \in \{a,b\}} (\nabla I_k^{ij}) \right|, \quad (3)$$

where ∇ denotes the Sobel operator, commonly employed for gradient extraction in image fusion [32, 57, 83]. The parameter α serves as a scaling factor, while \mathcal{L}_f^{int} and \mathcal{L}_f^{grad} represent the intensity loss and gradient loss, respectively. The weights $\{w_a, w_b\}$ control the emphasis of the loss function on the intensity information from each source image. These parameters enable the fusion process to selectively aggregate and incorporate relevant information from the source images. Variations in θ_G lead to different configurations of w_a and w_b , thereby influencing the characteristics of the fusion loss. θ_G undergoes updates driven by the high-level task loss during the in-step update process, as detailed in Sec. 3.4.

The loss function \mathcal{L}_t depends on the specific task. In this study, we adopt SegFormer [9] and YOLOv8 [19] for the downstream tasks, employing cross-entropy loss and YOLO loss [19] for each task, respectively.

3.3. Dataset Partitioning

In order to enhance the effectiveness of the loss generation module \mathcal{G} in guiding the fusion tasks, we create non-overlapping subsets of size M at each training epoch. The meta-training set $\{I_a^{mtr}, I_b^{mtr}\}$ and the meta-test set $\{I_a^{mts}, I_b^{mts}\}$ are randomly drawn from the fusion training set $\{I_a^{ftr}, I_b^{ftr}\}$. These subsets are fed into the model sequentially during the training process of the loss generation module, covering both the inner and outer updates. The entire fusion training set $\{I_a^{ftr}, I_b^{ftr}\}$ is fully utilized during the training of the fusion network.

3.4. Learning of loss generation module

Fig. 1 illustrates that the loss generation and fusion modules are trained alternately, ensuring the fusion loss is optimized at various stages of training and under different states of the fusion network. The process of training to optimize the fusion loss involves two key steps: the inner update and the outer update. In the inner update, clones of both the fusion network and the task network are generated, with each network undergoing a single training iteration using the fusion loss and task loss, respectively. This process is designed to obtain the state of the network guided by the fusion loss. During the outer update, the task loss of the fusion image produced by the updated clone is calculated. This loss is then backpropagated to the loss generation network. The goal of this step is to direct the fusion network to generate fusion images that result in lower downstream task loss, once guided by the fusion loss. The alternating updates between the inner and outer steps constitute the learning procedure for the loss generation module.

3.4.1. Inner Update

During the inner update phase, the fusion network \mathcal{F} undergoes a single update guided by the fusion loss, which

depends on the current state of network \mathcal{G} . This update primarily aims to compute the intermediate parameters $\theta_{\mathcal{F}'}$ and $\theta_{\mathcal{T}'}$, which are crucial for updating θ_G in the subsequent phase. The upper section of the purple region in Fig. 1 illustrates this process. During this phase, the images from the meta-training set $\{I_a^{mtr}, I_b^{mtr}\}$ are fed into the model:

$$\theta_{\mathcal{F}'} = \theta_{\mathcal{F}} - \eta_{\mathcal{F}'} \frac{\partial \mathcal{L}_f(I_a^{mtr}, I_b^{mtr}, I_f^{mtr}; \theta_G)}{\partial \theta_{\mathcal{F}}}, \quad (4)$$

where \mathcal{F} undergoes a single gradient descent update. θ_G represent the parameters of the loss generation module \mathcal{G} , which determine the parameters for the learnable fusion loss. And the notation $\eta_{\mathcal{F}'}$ refers to the step size. The module \mathcal{F}' temporarily substitutes \mathcal{F} , adjusting its parameters in one update step. Meanwhile, the parameters of \mathcal{F} , denoted as $\theta_{\mathcal{F}}$, remain unchanged. Similarly, \mathcal{T}' is updated in a single step using the parameters $\theta_{\mathcal{T}}$ from the current task network \mathcal{T} :

$$\theta_{\mathcal{T}'} = \theta_{\mathcal{T}} - \eta_{\mathcal{T}'} \frac{\partial \mathcal{L}_t(I_f^{mtr})}{\partial \theta_{\mathcal{T}}}. \quad (5)$$

The parameters $\theta_{\mathcal{F}'}$ and $\theta_{\mathcal{T}'}$ are both updated during the inner update, which also ensures that the computation graph of $\theta_{\mathcal{F}'}$ with respect to θ_G is preserved. This preserved graph is essential for optimizing θ_G during the outer update.

3.4.2. Outer Update

The primary objective of the outer update is to evaluate and refine the fusion guidance capability of \mathcal{G} , specifically by strengthening the influence of the loss function \mathcal{L}_f in steering the fusion module \mathcal{F} . In the framework diagram, this stage is represented in the lower part of the purple region. The modules \mathcal{F}' and \mathcal{T}' , derived from the inner update, represent the current guidance capacity of \mathcal{G} . In an ideal scenario, the optimal fusion loss should enhance the performance of the downstream task on the fused image. In this stage, the meta-test set $\{I_a^{mts}, I_b^{mts}\}$ is employed. The parameters θ_G are subsequently updated using the task loss \mathcal{L}_t , which is computed by \mathcal{F}' and \mathcal{T}' :

$$\theta_G = \theta_G - \eta_G \frac{\partial \mathcal{L}_t(I_f^{mts})}{\partial \theta_G}, \quad (6)$$

where $I_f^{mts} = \mathcal{F}'(I_a^{mts}, I_b^{mts})$, and the gradient $\partial \mathcal{L}_t / \partial \theta_G$ can be calculated as:

$$\frac{\partial \mathcal{L}_t}{\partial \theta_G} = \frac{\partial \mathcal{L}_t}{\partial \theta_{\mathcal{F}'}} * \left(-\eta_{\mathcal{F}'} \frac{\partial^2 \mathcal{L}_f(I_a^{mtr}, I_b^{mtr}, I_f^{mtr}; \theta_G)}{\partial \theta_{\mathcal{F}} \partial \theta_G} \right). \quad (7)$$

Eq. (6) holds because the task loss \mathcal{L}_t is determined by I_f^{mts} , which in turn relies on $\theta_{\mathcal{F}'}$. The optimization of θ_G via \mathcal{L}_t is realized by preserving the computational relationship between $\theta_{\mathcal{F}'}$ and θ_G throughout the inner update. The updated \mathcal{G} module gains the ability to generate enhanced fusion loss functions, enabling the fusion module to integrate relevant information from the source images more efficiently into the fused output for downstream tasks.

Algorithm 1 TDFusion Training Algorithm

Require: Training set $\{I_a^{ftr}, I_b^{ftr}\}$ with size N .
Output: Thoroughly trained $\theta_{\mathcal{F}}, \theta_{\mathcal{T}}, \theta_{\mathcal{G}}$.

- 1: Initialize $\theta_{\mathcal{F}}, \theta_{\mathcal{T}}, \theta_{\mathcal{G}}$.
- 2: **for** $epoch = 1$ **to** L **do**
- 3: Sample $\{I_a^{mtr}, I_b^{mtr}\}$ and $\{I_a^{mts}, I_b^{mts}\}$.
- 4: **for** $step = 1$ **to** M **do**
- 5: % Inner update: apply \mathcal{G} .
- 6: Sample (I_a^{mtr}, I_b^{mtr}) and get (I_f^{mtr}, y^{mtr}) .
- 7: Compute $\theta_{\mathcal{F}'}$ and $\theta_{\mathcal{T}'}$ by Eq. (4) and Eq. (5).
- 8: % Outer update: optimize \mathcal{G} .
- 9: Sample (I_a^{mts}, I_b^{mts}) and get (I_f^{mts}, y^{mts}) .
- 10: Update $\theta_{\mathcal{G}}$ by Eq. (6).
- 11: **end for**
- 12: **for** $step = 1$ **to** N **do**
- 13: % Fusion update: optimize \mathcal{F} and \mathcal{T} .
- 14: Sample (I_a^{ftr}, I_b^{ftr}) and get (I_f^{ftr}, y^{ftr}) .
- 15: Update $\theta_{\mathcal{F}}$ and $\theta_{\mathcal{T}}$ by Eq. (8) and Eq. (9).
- 16: **end for**
- 17: **end for**

3.5. Learning of Fusion Network

The alternating inner and outer update iterations form a flexible and effective mechanism to refine \mathcal{G} in response to the evolving state of \mathcal{F} . After refining \mathcal{G} , it is utilized to further improve the training of \mathcal{F} . This stage, denoted in blue in the diagrams, involves processing the fusion training set images $\{I_a^{ftr}, I_b^{ftr}\}$. Both \mathcal{F} and \mathcal{T} are updated through the application of the fusion loss \mathcal{L}_f and task loss \mathcal{L}_t :

$$\theta_{\mathcal{F}} = \theta_{\mathcal{F}} - \eta_{\mathcal{F}} \frac{\partial \mathcal{L}_f(I_a^{ftr}, I_b^{ftr}, I_f^{ftr}; \theta_{\mathcal{G}})}{\partial \theta_{\mathcal{F}}}, \quad (8)$$

$$\theta_{\mathcal{T}} = \theta_{\mathcal{T}} - \eta_{\mathcal{T}} \frac{\partial \mathcal{L}_t(I_f^{ftr})}{\partial \theta_{\mathcal{T}}}. \quad (9)$$

After completing several training sessions on the fusion network, the focus then shifts back to the learning phase of the fusion generation module. The fusion framework evolves through a series of alternating phases, with each phase fine-tuning the fusion loss based on the fusion network's current state. Such alternating phases ensure that the fusion network consistently applies the most suitable fusion loss during its progression. Ultimately, this results in the development of a highly efficient fusion network, optimized for peak performance. The complete training procedure is detailed in Algorithm 1.

3.6. Network Architecture

TDFusion is composed of three modules: the fusion network, the downstream task network, and the loss generation network. The fusion network shares the same architecture as [5], a lightweight model built upon the Restormer

Block (RTB) [74]. An adaptive fusion module is incorporated into this network to facilitate feature integration. Fig. 1 illustrates the structure of the loss generation module. It also employs the Restormer Block (RTB) [74] as its primary component, receiving inputs $\{I_a, I_b\}$. After applying $\text{Softmax}(\cdot)$, the final output guarantees $w_a^{ij} + w_b^{ij} = 1$ for each pixel. This design ensures that the fused image meets the similarity constraints and removes the reliance on initialization within the loss generation module. The architecture of the downstream task network $\mathcal{T}(\cdot)$ depends on the specific task. For learning of loss generation module, we chose the most lightweight models of SegFormer [9] and YOLOv8 [19] for semantic segmentation and object detection, respectively.

3.7. Theoretical Analysis

To better understand the weighting mechanism of the loss generation module \mathcal{G} , we investigate the optimization procedure of \mathcal{G} , which generates the weights $\{w_a, w_b\}$, denoted as $\theta_{\mathcal{G}}$. For clarity, we rewrite Eq. (2) as follows:

$$\begin{aligned} \mathcal{L}_f^{int} &= [w_a \odot (I_a - I_f) \odot (I_a - I_f) \\ &\quad + w_b \odot (I_b - I_f) \odot (I_b - I_f)] \times \frac{1}{HW} \\ &= [\mathcal{G}(I_a, I_b; \theta_{\mathcal{G}}) \odot (I_a - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b)) \\ &\quad \odot (I_a - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b)) \\ &\quad + (1 - \mathcal{G}(I_a, I_b; \theta_{\mathcal{G}})) \odot (I_b - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b)) \\ &\quad \odot (I_b - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b))] \times \frac{1}{HW}. \end{aligned} \quad (10)$$

Here, $w_a, w_b \in \mathbb{R}^{H \times W}$, $I_a, I_b \in \mathbb{R}^{H \times W}$, and \odot denotes the element-wise multiplication operation. Let Ω' be the set $\{\theta_{\mathcal{F}'}, \theta_{\mathcal{T}'}\}$, this leads to the following expression:

$$\begin{aligned} \theta_{\mathcal{G}} &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \frac{\partial \mathcal{L}_t^{mts}(\Omega'(\theta_{\mathcal{G}}))}{\partial \theta_{\mathcal{G}}} \\ &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \frac{\partial \mathcal{L}_t^{mts}(\Omega'(\theta_{\mathcal{G}}))}{\partial \Omega'} \frac{\partial \Omega'(\theta_{\mathcal{G}})}{\partial \theta_{\mathcal{G}}} \\ &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \underbrace{\eta_{\mathcal{F}'} \frac{\partial \mathcal{L}_t^{mts}(\Omega'(\theta_{\mathcal{G}}))}{\partial \Omega'}}_{(a)} \times \frac{\partial \mathcal{G}(I_a, I_b; \theta_{\mathcal{G}})}{\partial \theta_{\mathcal{G}}} \\ &\quad \times \underbrace{\left[\left(I_a - \frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}} \right) \odot \left(I_a - \frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}} \right) \right.}_{(b)} \\ &\quad \left. - \left(I_b - \frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}} \right) \odot \left(I_b - \frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}} \right) \right]}_{(b)} \\ &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \eta_{\mathcal{F}'} \mathbf{G} \times \frac{\partial \mathcal{G}(I_a, I_b; \theta_{\mathcal{G}})}{\partial \theta_{\mathcal{G}}}. \end{aligned} \quad (11)$$

Here, \mathbf{G} denotes the inner product between two gradients: (a) the first one is derived from the task loss using a **meta-testing set**, and (b) the second is calculated from the fusion loss based on a **meta-training set**. Consequently, the



Figure 2. Visual comparison of fusion results. The cases are “01258N” in MSRS dataset, “00122” in FMB dataset, “00449” in M3FD dataset and “200304” in LLVIP dataset.

optimization of the module θ_G is driven by the task loss, with the objective of preserving task-specific information throughout the fusion process.

4. Experiment

4.1. Setup

Experimental Setup. In our experiments, the epoch number L and the training iterations of the loss generation module M are set to 50 and 200. The learning iterations for the fusion network N depend on the size of the dataset. We use the Adam optimizer with a learning rate of $1e-4$, a batch size of 2, and a hyperparameter α set to 1. All experiments are conducted on a PC with a single NVIDIA RTX 3090 GPU.

Evaluation Metrics and Comparison Methods. The advanced fusion methods compared in our study include TarDAL [32], SegMIF [34], MURF [69], EMMA [86], DCINN [61], MRFS [76], and TIMFusion [40]. The fusion performance is evaluated using metrics including entropy (EN), spatial frequency (SF), sum of correlation differences (SCD), visual information fidelity (VIF), $Q^{AB/F}$, and structural similarity index metric (SSIM).

Dataset Split. We use four datasets annotated for downstream tasks, including MSRS [58], FMB [34], M3FD [32], and LLVIP [17]. MSRS contains 1083/361 image pairs for training/test, and FMB contains 1220/280 pairs for training/test. We follow the splits of the original papers for both datasets. M3FD dataset consists of 4200 pairs of images for detection, with 300 pairs designated for fusion evaluation.

The 4200 detection images are split into 3150 for training and 1050 for testing, ensuring that the 300 pairs for fusion evaluation are included within the detection test set. These 300 fusion images are then employed to assess the fusion performance. The original LLVIP dataset contains 12025/3463 image pairs as training/test set. Due to its large size, we select every 10th image to form our training and test sets, resulting in 1203/347 image pairs for training and testing. Our splits for M3FD and LLVIP will be available.

4.2. Fusion Experiments

Fig. 2 presents a visual comparison of different methods. The fused images generated by TDFusion excel in preserving discriminative details, achieving balanced brightness, and maintaining clear object contours. It effectively preserves the target features from the infrared images and the background details from the visible images, resulting in fused images that are more natural and exhibit greater clarity across different environments. These results highlight the advantages of TDFusion in detail preservation and visual performance. More results are available in the supplementary material. Tab. 1 presents the quantitative comparison of fusion over four datasets. TDFusion outperforms other methods across most metrics. This suggests that TDFusion not only enhances image details but also provides consistent fusion results across various scenarios. Compared to other methods, TDFusion shows superior adaptability and robustness, particularly in handling diverse image characteristics and challenging fusion tasks.

Table 1. Quantitative comparison of Infrared-visible image fusion. The **red** and **blue** markers represent the best and second-best values.

Infrared-visible Image Fusion on MSRS [58] Dataset							Infrared-visible Image Fusion on FMB [34] Dataset						
	EN \uparrow	SF \uparrow	SCD \uparrow	VIF \uparrow	$Q^{AB/F} \uparrow$	SSIM \uparrow		EN \uparrow	SF \uparrow	SCD \uparrow	VIF \uparrow	$Q^{AB/F} \uparrow$	SSIM \uparrow
TarDAL [32]	5.28	5.98	0.71	0.21	0.18	0.47	TarDAL [32]	6.63	6.94	1.03	0.28	0.29	0.74
SegMIF [34]	5.95	11.10	1.57	0.44	0.63	0.55	SegMIF [34]	6.83	13.69	1.72	0.39	0.65	0.60
MURF [69]	5.04	10.49	1.02	0.22	0.37	0.60	MURF [69]	6.37	13.88	1.34	0.22	0.37	0.68
EMMA [86]	6.73	11.56	1.62	0.49	0.64	0.70	EMMA [86]	6.77	15.00	1.50	0.42	0.65	0.72
DCINN [61]	6.00	10.51	1.49	0.41	0.57	0.52	DCINN [61]	6.47	11.47	1.39	0.38	0.59	0.74
MRFS [76]	7.00	8.86	1.42	0.37	0.49	0.55	MRFS [76]	6.78	12.42	1.24	0.38	0.62	0.73
TIMFusion [40]	6.27	9.67	1.34	0.32	0.48	0.68	TIMFusion [40]	6.51	12.23	1.24	0.35	0.59	0.73
TDFusion (Ours)	6.74	11.30	1.86	0.50	0.67	0.70	TDFusion (Ours)	6.86	14.16	1.76	0.43	0.68	0.75

Infrared-visible Image Fusion on M3FD [32] Dataset							Infrared-visible Image Fusion on LLVIP [17] Dataset						
	EN \uparrow	SF \uparrow	SCD \uparrow	VIF \uparrow	$Q^{AB/F} \uparrow$	SSIM \uparrow		EN \uparrow	SF \uparrow	SCD \uparrow	VIF \uparrow	$Q^{AB/F} \uparrow$	SSIM \uparrow
TarDAL [32]	6.87	7.63	1.29	0.27	0.30	0.71	TarDAL [32]	6.32	7.42	1.04	0.27	0.22	0.58
SegMIF [34]	6.85	14.14	1.72	0.37	0.60	0.59	SegMIF [34]	6.68	15.46	1.38	0.40	0.66	0.57
MURF [69]	6.50	12.55	1.46	0.21	0.32	0.64	MURF [69]	6.13	15.08	0.96	0.21	0.31	0.57
EMMA [86]	6.92	15.23	1.49	0.38	0.59	0.69	EMMA [86]	7.35	15.37	1.57	0.41	0.64	0.66
DCINN [61]	6.59	11.21	1.46	0.34	0.51	0.72	DCINN [61]	6.98	13.34	1.43	0.38	0.52	0.64
MRFS [76]	6.94	12.07	1.26	0.34	0.55	0.70	MRFS [76]	6.83	11.04	1.23	0.31	0.42	0.64
TIMFusion [40]	6.75	12.31	1.37	0.35	0.53	0.70	TIMFusion [40]	6.58	13.52	1.14	0.33	0.46	0.64
TDFusion (Ours)	6.99	14.49	1.83	0.41	0.65	0.72	TDFusion (Ours)	7.36	16.38	1.75	0.46	0.70	0.67

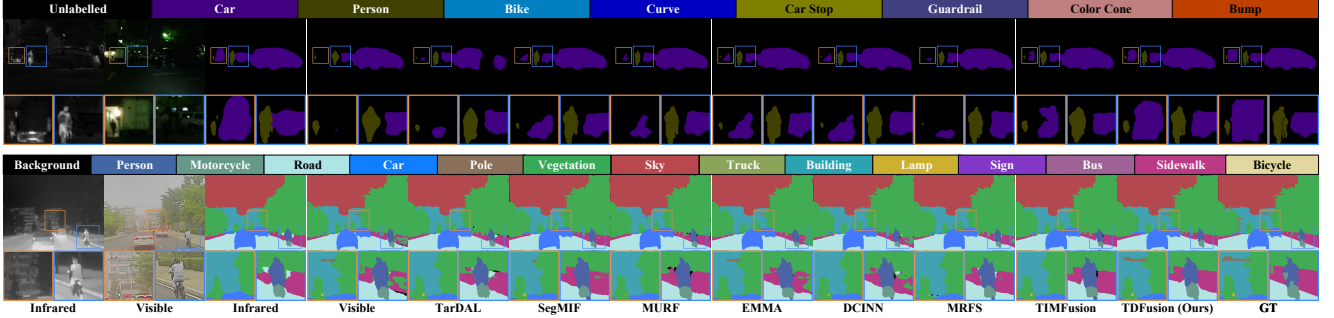


Figure 3. Visual comparison for Semantic Segmentation. The cases are “00726N” in MSRS dataset and “01438” in FMB dataset.

4.3. Downstream Applications

This section validates the adaptability of fusion methods to downstream tasks. For a fair comparison, we adopt SegFormer [9] and YOLOv8 [19] as backbones and retrain the task networks for each fusion method over 300 epochs to evaluate their adaptability to semantic segmentation and object detection. Fig. 3 and Fig. 4 present the visual comparisons of semantic segmentation and object detection, respectively. TDFusion outperforms in image detail retention, edge clarity, and object recognition, effectively identifying and segmenting objects. For semantic segmentation, the generated maps clearly distinguish different class regions, closely matching the ground truth. In object detection, the fused images exhibit more precise boundary localization for salient objects. This indicates that TDFusion maintains a better balance between fine details and overall context. More results can be found in the supplementary material.

Tab. 2 shows the performance comparison across different methods in semantic segmentation and object detection. TDFusion outperforms other methods on most metrics, particularly in mIoU and mAP. This demonstrates that TDFu-

Table 2. Performance comparison of downstream applications. The **red** and **blue** markers represent the best and second-best.

Methods	Semantic Segmentation				Object Detection			
	MSRS		FMB		M3FD		LLVIP	
	mAcc	mIoU	mAcc	mIoU	mAP50	mAP75	AP50	AP75
Infrared	83.23	69.49	58.85	51.98	79.12	53.05	96.03	72.07
Visible	83.44	73.76	65.12	57.96	82.21	54.82	91.78	48.66
TarDAL	81.93	71.35	62.86	55.33	83.16	56.39	93.79	62.71
SegMIF	85.73	74.25	65.97	58.41	83.61	58.23	93.95	66.45
MURF	85.03	74.08	64.10	56.96	80.58	54.22	94.24	68.04
EMMA	85.99	74.48	62.45	56.28	83.71	56.91	94.00	66.21
DCINN	84.11	74.35	61.09	54.81	82.69	57.37	94.92	68.34
MRFS	84.76	74.50	61.93	55.71	83.28	57.74	93.03	67.21
TIMFusion	83.67	73.58	63.70	57.24	83.22	56.08	93.76	61.33
TDFusion	86.04	75.09	67.17	60.50	86.27	59.71	95.00	69.18

sion effectively enhances the quality of fused images. It also improves the downstream task performance, especially in terms of accuracy and robustness in complex scenarios. Class-wise results are provided in the supplementary material.

4.4. Task-driven Learnable Loss

Our framework incorporates a learnable fusion loss that models the preferences of downstream tasks for information from source images. The models trained on FMB dataset

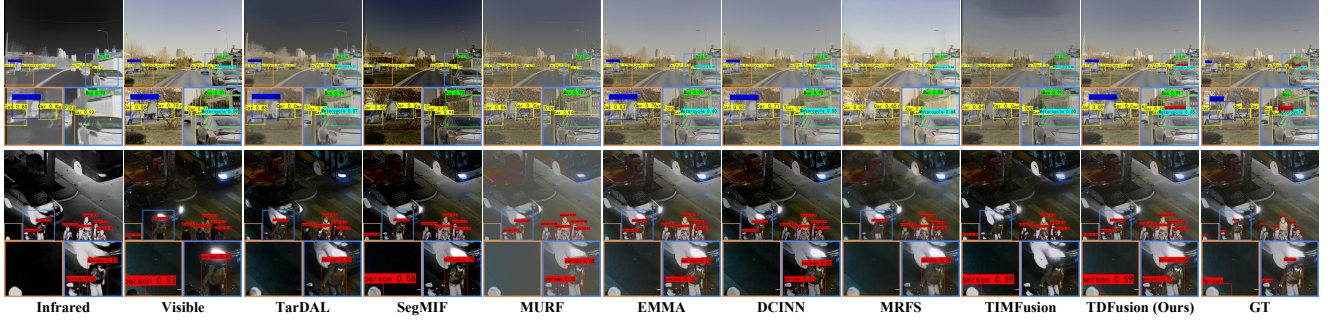


Figure 4. Visual comparison for Object Detection. The cases are “02236” in M3FD dataset and “210145” in LLVIP dataset.

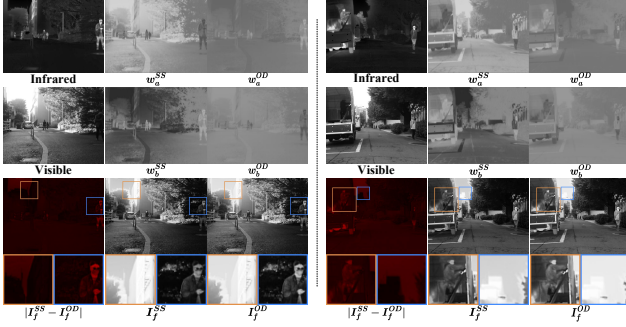


Figure 5. Visualisation of learnable loss for downstream tasks.

and LLVIP dataset labeled as *SS* and *OD*, are evaluated on MSRS dataset to simulate performance in unknown scenes, as shown in Fig. 5. The results demonstrate that the fusion model adaptively selects information from infrared and visible images to satisfy task requirements. In semantic segmentation, the model combines scene structure and texture, prioritizing boundaries. This improves segmentation under varying lighting conditions. Fusion weights $\{w_a^{SS}, w_b^{SS}\}$ indicate a preference for visible details and infrared advantages in low-light conditions. In object detection, the model focuses on edge and contrast information, especially for instances like pedestrians and vehicles. Higher fusion weights are assigned to bright regions in infrared images, enhancing target detection in low-light conditions. Fusion weights $\{w_a^{OD}, w_b^{OD}\}$ reflect this preference. Comparison of fusion losses across tasks reveals distinct differences, especially in highlighted regions, confirming that the model adapts to task-specific requirements by selecting the most relevant information from multimodal images. More results are provided in the supplementary material.

4.5. Ablation Studies

To thoroughly evaluate the performance of our proposed algorithm, we conduct a series of ablation experiments on FMB dataset, and the detailed results are shown in Tab. 3. In Exp. I, we exclude the learnable fusion loss by fixing w_a and w_b to $1/2$. In Exp. II, we omit the gradient loss from the loss function. In Exp. III, we also allow the fusion module parameters to be jointly optimized by both the task loss \mathcal{L}_t and fusion loss \mathcal{L}_f . In Exp. IV, we exclude the fusion

Table 3. Ablation experiment of fusion. The **red** denotes the best.

Ablation Studies of fusion on FMB Dataset							
Configurations	EN	SF	SCD	VIF	$Q^{AB/F}$	SSIM	
I fix w_a and w_b as $1/2$	6.60	13.73	1.58	0.39	0.60	0.72	
II w/o \mathcal{L}_f^{grad}	6.77	11.65	1.63	0.37	0.64	0.73	
III $\theta_{\mathcal{F}}$ influenced by \mathcal{L}_t	6.80	13.85	1.70	0.41	0.66	0.73	
IV w/o Fusion learning	6.82	14.07	1.72	0.41	0.67	0.72	
V $I_f = w_a * I_a + w_b * I_b$	6.75	11.49	1.65	0.38	0.62	0.73	
Ours	6.86	14.16	1.76	0.43	0.68	0.75	

module’s dedicated learning phase and update it during the outer update of the loss module, which tests the impact of the fusion module’s training schedule. In Exp. V, we replace our fusion method with a discriminative approach. The performance decline observed across different configurations confirms the rationality and effectiveness of our proposed method. Visualization and more analysis are provided in the supplementary material.

5. Conclusion

To overcome the limitations of predefined fusion losses, which often fail to effectively guide the fusion process for downstream tasks, we propose a meta-learning-based framework for task-guided fusion. This framework includes a loss generation module that outputs the parameters of the learnable fusion loss. The module is updated using a meta-learning approach, which alternates between inner and outer loop steps to enhance its ability to guide the fusion network. Under varying fusion conditions, this module generates the optimal fusion loss for the downstream task. This enables the fusion network to produce fused images that minimize the task-specific loss. The theoretical analysis explains how the downstream task loss guides the fusion loss in our framework. Experiments using four publicly available fusion datasets and downstream tasks including semantic segmentation and object detection, demonstrate the effectiveness of our approach.

Acknowledgement

This work has been supported by the National Natural Science Foundation of China under Grant 12201497 and 12371512.

References

- [1] Hessah Albanwan, Rongjun Qin, and Yang Tang. Image fusion in remote sensing: An overview and meta-analysis. *Photogrammetric Engineering & Remote Sensing*, 90(12): 755–775, 2024. **1**
- [2] Antreas Antoniou and Amos J. Storkey. Learning to learn by self-critique. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 9936–9946, 2019. **2**
- [3] Haowen Bai, Zixiang Zhao, Jianshe Zhang, Yukun Cui, Chunxia Zhang, Zhenbo Guo, and Yongjun Wang. Simultaneous automatic picking and manual picking refinement for first-break. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. **1**
- [4] Haowen Bai, Zixiang Zhao, Jianshe Zhang, Baisong Jiang, Lilun Deng, Yukun Cui, Shuang Xu, and Chunxia Zhang. Deep unfolding multi-modal image fusion network via attribution analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. **2**
- [5] Haowen Bai, Zixiang Zhao, Jianshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Baisong Jiang, and Shuang Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *International Journal of Computer Vision*, pages 1–21, 2024. **3, 5**
- [6] Haowen Bai, Jianshe Zhang, Zixiang Zhao, Lilun Deng, Yukun Cui, and Shuang Xu. Retinex-mef: Retinex-based glare effects aware unsupervised multi-exposure image fusion. *arXiv preprint arXiv:2503.07235*, 2025. **2**
- [7] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9445–9454. IEEE, 2021. **2**
- [8] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, and Yu Qiao. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *Information Fusion*, 46:206–217, 2019. **1**
- [9] Bo Cheng, Xiang Li, Yujie Wei, Cheng Huang, Xiaoyong Zhang, Yandong Jiang, Tianyu Zhang, Na Xu, Shuai Yu, Xinxin Zhan, et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12022–12032, 2021. **4, 5, 7**
- [10] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2021. **2**
- [11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International conference on machine learning (ICML)*, pages 1126–1135, 2017. **2**
- [12] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *Proceedings of the International conference on machine learning (ICML)*, pages 1920–1930, 2019. **2**
- [13] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115. IEEE, 2017. **1**
- [14] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *Neural Information Processing: 28th International Conference (ICONIP)*, pages 387–395. Springer, 2021. **1**
- [15] Rein Houthoofd, Yuhua Chen, Phillip Isola, Bradley C. Stadie, Filip Wolski, Jonathan Ho, and Pieter Abbeel. Evolved policy gradients. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5405–5414, 2018. **2**
- [16] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 539–555. Springer, 2022. **2**
- [17] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. **6, 7**
- [18] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 4369–4376, 2020. **1**
- [19] Glenn Jocher. Ultralytics YOLOv8. <https://github.com/ultralytics>, 2023. **4, 5, 7**
- [20] Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 831–847. Springer, 2018. **1**
- [21] 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, 2**
- [22] 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. **3**
- [23] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021. **1, 2**
- [24] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K. Tokuda, Roberto Hirata Junior, Roberto Marccondes Cesar Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3838–3847. Computer Vision Foundation / IEEE, 2019. **1**
- [25] Xiaoling Li, Yanfeng Li, Houjin Chen, Yahui Peng, and Pan Pan. Ccafusion: cross-modal coordinate attention network for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. **1**

- [26] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. [2](#)
- [27] Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li, Changsheng Gao, et al. Ustc-td: A test dataset and benchmark for image and video coding in 2020s. *arXiv preprint arXiv:2409.08481*, 2024. [1](#)
- [28] Zhuoyuan Li, Zikun Yuan, Li Li, Dong Liu, Xiaohu Tang, and Feng Wu. Object segmentation-assisted inter prediction for versatile video coding. *IEEE Transactions on Broadcasting*, 2024. [1](#)
- [29] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 719–735. Springer, 2022. [2](#)
- [30] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. [2](#)
- [31] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119, 2021. [1](#)
- [32] 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 (CVPR)*, pages 5792–5801, 2022. [1](#), [2](#), [4](#), [6](#), [7](#)
- [33] Jinyuan Liu, Jingjie Shang, Risheng Liu, and Xin Fan. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5026–5040, 2022. [2](#)
- [34] 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 (ICCV)*, pages 8115–8124, 2023. [1](#), [2](#), [6](#), [7](#)
- [35] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95:237–249, 2023. [2](#)
- [36] 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*, 132(5):1748–1775, 2024. [2](#)
- [37] 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](#)
- [38] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1600–1608. ACM, 2021. [3](#)
- [39] Risheng Liu, Long Ma, Tengyu Ma, Xin Fan, and Zhongxuan Luo. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5953–5969, 2022. [1](#)
- [40] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [1](#), [2](#), [3](#), [6](#), [7](#)
- [41] ShuMin Liu, Jiajia Chen, and Susanto Rahardja. A new multi-focus image fusion algorithm and its efficient implementation. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(5):1374–1384, 2020. [1](#)
- [42] Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1240–1248, 2023. [2](#)
- [43] Zhu Liu, Jinyuan Liu, Benzhuang Zhang, Long Ma, Xin Fan, and Risheng Liu. Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3706–3714, 2023.
- [44] Zhu Liu, Jinyuan Liu, Guanyao Wu, Zihang Chen, Xin Fan, and Risheng Liu. Searching a compact architecture for robust multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6224–6237, 2024. [2](#)
- [45] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. [1](#)
- [46] 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. [2](#)
- [47] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.
- [48] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiaoping (Steven) Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29: 4980–4995, 2020. [2](#)
- [49] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5637–5646, 2022. [1](#)
- [50] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. [2](#)

- [51] Seonghyun Park, An Gia Vien, and Chul Lee. Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):770–785, 2023. 1
- [52] Yanting Pei, Yaping Huang, Qi Zou, Yuhang Lu, and Song Wang. Does haze removal help cnn-based image classification? In *Proceedings of the European conference on computer vision (ECCV)*, pages 682–697, 2018. 1
- [53] Xinran Qin, Yuhui Quan, Tongyao Pang, and Hui Ji. Ground-truth free meta-learning for deep compressive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9947–9956, 2023. 2
- [54] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the International conference on machine learning (ICML)*, pages 4334–4343, 2018. 2
- [55] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [56] 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. 1, 2
- [57] 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. 4
- [58] 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. 6, 7
- [59] 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. 1
- [60] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3508–3515. ijcai.org, 2022. 2
- [61] Wu Wang, Liang-Jian Deng, Ran Ran, and Gemine Vivone. A general paradigm with detail-preserving conditional invertible network for image fusion. *International Journal of Computer Vision*, 132(4):1029–1054, 2024. 6, 7
- [62] Zeyu Xiao and Xinchao Wang. Event-based video super-resolution via state space models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1
- [63] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Xiaoyan Sun, and Zhiwei Xiong. Asymmetric event-guided video super-resolution. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 2409–2418, 2024. 1
- [64] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Zheng-Jun Zha, Xiaoyan Sun, and Zhiwei Xiong. Event-adapted video super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–235. Springer, 2024. 1
- [65] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. FusionDn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 12484–12491, 2020. 2
- [66] 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, 2022. 2
- [67] 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, 2022. 2
- [68] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19647–19656, 2022. 2
- [69] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6, 7
- [70] Shuang Xu, Ouafa Amira, Junmin Liu, Chun-Xia Zhang, Jianshe Zhang, and Guanghai Li. HAM-MFN: hyperspectral and multispectral image multiscale fusion network with RAP loss. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4618–4628, 2020. 1
- [71] Shuang Xu, Lizhen Ji, Zhe Wang, Pengfei Li, Kai Sun, Chunxia Zhang, and Jianshe Zhang. Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy. *IEEE Transactions Computational Imaging*, 6:1561–1570, 2020. 1
- [72] Yong Yang, Jiaxiang Liu, Shuying Huang, Weiguo Wan, Wenying Wen, and Juwei Guan. Infrared and visible image fusion via texture conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4771–4783, 2021. 1
- [73] 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
- [74] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. 5
- [75] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021. 2
- [76] 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 Com-*

- puter Vision and Pattern Recognition (CVPR), pages 26974–26983, 2024. 1, 2, 3, 6, 7
- [77] 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, 2
- [78] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13955–13965, 2023. 1, 2, 3
- [79] Yangyang Zhao, Qingchun Zheng, Peihao Zhu, Xu Zhang, and Wenpeng Ma. Tufusion: A transformer-based universal fusion algorithm for multimodal images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1
- [80] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jianshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 970–976, 2020. 2
- [81] Zixiang Zhao, Shuang Xu, Jianshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2022. 2
- [82] 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 (CVPR)*, pages 5906–5916, 2023. 1, 2
- [83] 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 (CVPR)*, pages 5906–5916, 2023. 1, 2, 3, 4
- [84] 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. *CoRR*, abs/2303.06840, 2023. 2
- [85] Zixiang Zhao, Jianshe Zhang, Haowen Bai, Yicheng Wang, Yukun Cui, Lilun Deng, Kai Sun, Chunxia Zhang, Junmin Liu, and Shuang Xu. Deep convolutional sparse coding networks for interpretable image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2369–2377, 2023. 2
- [86] 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 (CVPR)*, pages 25912–25921, 2024. 6, 7
- [87] Huabing Zhou, Wei Wu, Yanduo Zhang, Jiayi Ma, and Haibin Ling. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. *IEEE Transactions on Multimedia*, 25:635–648, 2023. 1, 2
- [88] Man Zhou, Naishan Zheng, Xuanhua He, Danfeng Hong, and Jocelyn Chanussot. Probing synergistic high-order interaction for multi-modal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [89] 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*, pages 7099–7108, 2024. 1, 2, 3