A General Image Fusion Approach Exploiting Gradient Transfer Learning and Fusion Rule Unfolding

Wu Wang, Liang-Jian Deng, *Senior Member, IEEE*, Qi Cao, Gemine Vivone, *Senior Member, IEEE*

Abstract—The goal of a deep learning-based general image fusion method is to solve multiple image fusion tasks with a single model, thereby facilitating the deployment of models in practical applications. However, existing methods fail to provide an efficient and comprehensive solution from both model training and network design perspectives. Regarding model training, current approaches cannot effectively leverage complementary information across different tasks. In terms of network design, they rely on experience-based network designs. To address these issues, we propose a comprehensive framework for general image fusion using the newly proposed gradient transfer learning and fusion rule unfolding. To leverage complementary information across different tasks during training, we propose a sequential gradient-transfer framework based on the idea that different image fusion tasks often exhibit complementary structural details and that image gradients effectively capture these details. To move beyond heuristic-based network design, we evolved a fundamental image fusion rule and integrated it into a deep equilibrium model, resulting in a more efficient and versatile image fusion network capable of uniformly handling various fusion tasks. Considering three different image fusion tasks, i.e., multi-focus image fusion, multi-exposure image fusion, and infrared and visible image fusion, our method not only produces images with richer structural information but also achieves highly competitive objective metrics. Furthermore, the results of generalization experiments on previously unseen image fusion tasks, i.e., medical image fusion, demonstrate that our method significantly outperforms competing approaches. The code will be available upon possible acceptance.

Index Terms—Image fusion, general learning framework, gradient transfer learning, fusion rule unfolding, multi-focus image fusion, multi-exposure image fusion, infrared and visible image fusion.

1 Introduction

Individual imaging sensors have physical limitations that restrict them to capturing only specific parts of a scene. Image fusion algorithms aim to overcome these limitations by combining information from different source images to create a single image taking the best from the input data. This paper focuses on image fusion algorithms specifically designed for typical image fusion tasks, including multi-focus image fusion (MFIF), multi-exposure image fusion (MEIF), infrared and visible image fusion (IVF), and medical image fusion (MIF). A schematic illustration is shown in Fig. 1.

Traditional image fusion methods [5], [6], [7], [8] use handcrafted features and fusion rules to generate fused outcomes. Among them, the transformation-based image fusion method has been widely studied because it can model multiple image fusion methods. Typical examples of transformations are the wavelet transformation [8] and the nonsubsampled contourlet transformation [5]. These methods, while offering interpretability and leveraging domain knowledge, struggle with modeling the non-linear relationships necessary for effective image fusion. As a result, their performance can be limited.

Deep learning (DL) methods have become prominent in image fusion research thanks to their strong representational capabilities. DL-based image fusion methods [9], [10], [11], [12], [13], [14], [15] can be categorized into non-general and general approaches. Since non-general image fusion methods can only solve single image fusion tasks, recent research has focused on developing general image fusion methods that can solve multiple image fusion tasks with one model. Despite notable progress, existing general image fusion methods fail to provide an efficient and comprehensive solution from both the perspectives of model training and network design.

In terms of model training, most methods adopt either task-specific or multi-task learning approaches. Task-specific learning involves the training of separate models with the same neural network for each task, but this approach is limited by its inability to leverage information across tasks. To solve this problem, U2Fusion [9] and TC-MOE [16] adopted multi-task learning to learn from multiple image fusion data. However, multi-task learning methods face the task conflict problem [17], causing the model's performance to degrade for some tasks. For example, in the MEIF task, the source image may have underexposed or overexposed regions, and the fused image is required to have normal exposure. Besides, the MFIF task demands that

W. Wang is with the School of Computing and Artificial Intelligence, and with the Engineering Research Center of Intelligent Finance, Ministry of Education, Southwestern University of Finance and Economics, Chengdu, Sichuan, 611130, China.
L.-J. Deng and Q. Cao are with the Yingcai Honors College and School of Mathematical Sciences, respectively, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China.
G. Vivone is with the Institute of Methodologies for Environmental Analysis, CNR-IMAA, Tito Scalo, 85050, Italy and with the National Biodiversity Future Center, NBFC, Palermo, 90133, Italy. (e-mail: gemine.vivone@imaa.cnr.it).
Corresponding author: L.-J. Deng (e-mail: liangjian.deng@uestc.edu.cn).

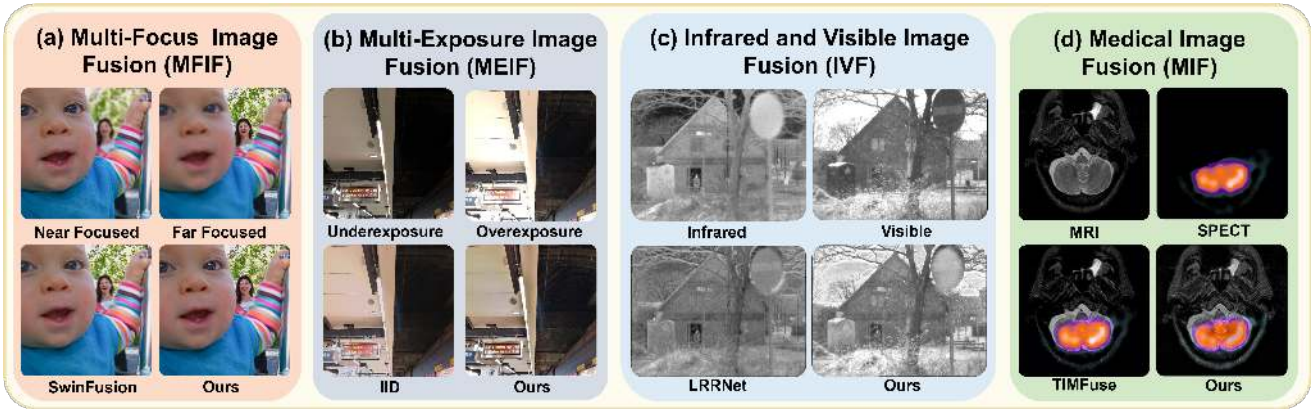

Fig. 1: A schematic illustration of various image fusion tasks using real examples. The first row depicts flowcharts for four typical image fusion tasks: MFIF, MEIF, IVF, and MIF. The second row presents the results of these tasks using state-of-the-art image fusion methods, including SwinFusion [1], IID [2], LRRNet [3], TIMFuse [4], and the proposed method. Compared to the state-of-the-art methods, our approach produces fusion outcomes with better structural information.

the fused image preserves the exposure of the source image. As a result, the objectives of the MEIF task and the MFIF task are conflicting, making it challenging for multi-task learning-based methods to achieve promising performance on multiple image fusion tasks simultaneously. In summary, while task-specific approaches cannot exploit cross-task information, multi-task learning methods are hindered by task conflicts that limit their overall effectiveness.

In terms of network design, existing DL-based general image fusion methods often rely on empirical experience for network design, which can be time-consuming and difficult to explain [3]. For example, CDDFuse [4] integrates convolutional neural networks (CNNs), invertible neural networks (INNs), and Transformers into a hybrid structure, requiring significant effort to effectively coordinate these architectures. To address these issues, we propose a comprehensive framework from both perspectives of model training and network design.

Specifically, to effectively leverage cross-task complementary information without suffering from task conflicts during model training, we conduct a series of experiments to analyze the visual characteristics of fusion images generated by task-specific learning methods during cross-task inferencing. The experimental results showed that the models trained on single tasks exhibit significant distortions in aspects such as global brightness and exposure, but there is complementary information in the local structures. From the experimental results, we drew two conclusions: i) there are conflicts among different image fusion tasks, making it difficult for a single model to achieve good performance across multiple image fusion tasks; ii) there is information complementarity among different image fusion tasks. We further validated this observation by analyzing the gradient maps of fusion images. Based on these observations, we propose a training framework that incorporates sequential gradient transfer. In this framework, we sequentially train the models using data from MFIF, MEIF, and IVF tasks, where the model trained in the previous stage serves as the teacher model in the next stage, and its generated image gradients are utilized in the subsequent stage. By training a separate model for each type of task, we effectively avoid the issue of conflicting task objectives. Meanwhile, the cross-task gradient transfer enables the model to extract complementary structure information across tasks.

Rather than relying on experience-based network design, we derived our model from a fundamental image fusion rule applicable across a wide range of image fusion tasks, unfolding it into a learnable iterative algorithm. By incorporating a deep equilibrium (DEQ) model, we further refined the algorithm to avoid explicit unfolding, resulting in an efficient and general image fusion network. This approach significantly reduces the model's parameters and training memory while maintaining comparable performance.

We incorporate the sequential gradient transfer training algorithm and an implicitly unfolded network based on fundamental fusion rules into a unified framework. This integrated approach not only improves fusion performance across various tasks but also increases computational efficiency, thereby offering a more balanced and practical solution.

Our main contributions are as follows:

- Motivated by the widely observed complementary structural information across different tasks and models, we propose a novel and first-of-its-kind training paradigm for general image fusion. Unlike task-specific models, our framework explicitly exploits inter-task correlations via a sequential gradient transfer mechanism.
- Building on the insight that multiple image fusion tasks can be uniformly modeled using fundamental fusion rules, we introduce a unified deep unfolding network derived from a classical fusion formulation.
- To enhance efficiency, we integrate the DEQ model into the network, constructing an implicit structure that eliminates the need for iterative unfolding. This advancement significantly improves memory and parameter efficiency while preserving competitive performance.
- Extensive experiments conducted on four image fusion tasks demonstrate that our method achieves highly competitive performance in terms of both visual quality and objective metrics, while also exhibit-

ing notable efficiency. Furthermore, the approach shows strong generalization capability to unseen data and tasks. Ablation studies offer additional validation, confirming the effectiveness of each core component of the proposed framework.

2 Related Works

In this section, we review prior research mostly related to our study, thus highlighting the novelty of our work.

2.1 Image Fusion Methods

2.1.1 Traditional Image Fusion Methods

Traditional image fusion methods typically perform image fusion in a certain transform domain. They first project the source images into the transformed domain using a predefined transformation. Fusion is then carried out using simple handcrafted rules, such as chosen-max and weighted-average. The final fused image is obtained by performing an inverse transformation. Typical applied transformations include the wavelet transformation [18], shift feature transformation [19], nonsubsampled contourlet transformation [5], and target-enhanced multi-scale transformation [20]. These methods rely on handcrafted transformations for feature extraction and simple fusion rules, which often lack the capability to effectively represent complex features.

2.1.2 DL-based Non-general Image Fusion Methods

DL-based non-general image fusion methods are designed to handle specific tasks. These methods can be broadly classified into two categories: CNN-based and Transformer-based. CNN-based methods are better at extracting local features. Some examples include [3], [10], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. For instance, LRRNet [3] proposed a convolutional sparse dictionary coding network to address the IVF problem. Transformer-based methods, like [31], [32], and [33], are more adept at capturing global features. For example, YDTR [32] employs a Y-shaped Transformer architecture to tackle the IVF task. The recently proposed EAT model [34] introduces a focal Transformer network designed to focus on salient regions while establishing long-range multi-exposure relationships thereby addressing MEIF through adversarial learning. Moreover, ITFuse [35], an interactive transformer framework for IVF, employs feature interaction modules and reconstruction mechanisms to mitigate information loss. Besides, GAN-based image fusion methods such as [36], [37], [38], [39], [40], [41], [42], [43], [44] learn the distribution mapping with a generator and a discriminator. FusionGAN [42] is the first GAN-based method designed to address the IVF task.

2.1.3 DL-based General Image Fusion Methods

IFCNN [45] is the first general image fusion model based on deep learning. It trains a robust model through extensive data augmentation on a simulated MFIF dataset, enabling the model to generalize to other image fusion tasks. Subsequently, U2Fusion [9] employed continual learning to perform unsupervised learning for multiple image fusion tasks simultaneously. In contrast, models like SwinFusion [1], CddFuse [4], CUNet [46], PMGI [47], SDNet [48], SFINet [49], and DCINN [50] use task-specific training, independently training models for each task. Recently, TCMOA [16] and PSLPT [51] have used multi-task learning to develop general image fusion models. TIMFuse [52] initially uses multi-task learning to obtain a pre-trained model, which is then fine-tuned for each specific task to enhance performance. Among these methods, U2Fusion [9] is the most similar to the proposed one. However, there are two key differences between our approach and U2Fusion [9]. First, we train a separate model for each image fusion task, while U2Fusion [9] uses a single model for all tasks. This single-model approach in U2Fusion [9] can lead to task conflict issues (causing performance degradation) avoided by our method. Second, we incorporate transfer learning during sequential training to extract complementary structural information, thereby better preserving structural details.

2.2 Transfer Learning

Transfer learning aims to utilize the knowledge learned from previous tasks or domains to improve learning performance in new tasks or domains. While widely applied to computer vision tasks such as image classification [53], object detection [54], and image segmentation [55], it has been less explored in the context of image fusion. In this work, we employ transfer learning techniques to leverage the complementary structure information across different image fusion tasks.

2.3 Deep Equilibrium Model

Most foundation models, e.g., ResNet [56], DenseNet [57], and SwinTransformer [58], consist of a series of basic modules with different parameters that perform forward computation sequentially to learn a mapping. Therefore, the feature maps of each layer must be stored to enable feed-forward computation and gradient calculation. As a result, when the model becomes deeper or wider, the number of parameters and the computational cost increase accordingly. Unlike these neural networks, the DEQ model [59] iteratively uses the same basic module to learn nonlinear mapping. Since this iterative process can be regarded as the fixed point iteration, when the number of iterations is infinite, the iterative process will converge to the equilibrium point. The DEQ model [59] directly solves the equilibrium point with numerical analysis methods, which avoids explicit feed-forward computation. Therefore, the DEQ model [59] does not need to store intermediate results, thus reducing the memory demand during training. The DEQ model [59] has demonstrated promising performance on various computer vision tasks such as semantic segmentation [60] and object detection [61]. In this work, we leverage the insight that deep unfolding models can be viewed as a fixed-point iteration process, which can be efficiently solved using a DEQ model. Based on this idea, we introduce a novel DEQ solver tailored to our fusion-rule-driven unfolded model to improve efficiency.

2.4 Deep Unfolding for Image Fusion

Most DL-based image fusion approaches often rely on experience to design network structures, which can be

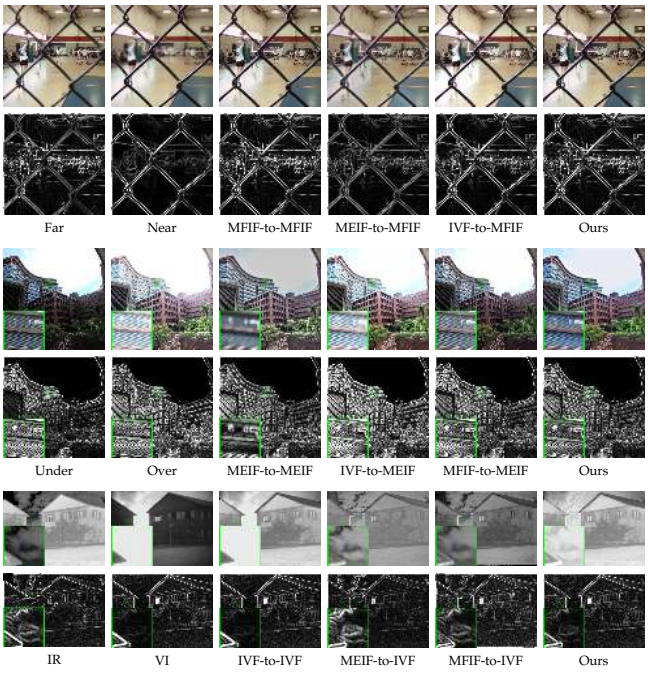

Fig. 2: Results of some task-specific trained models on three typical image fusion tasks. Rows 1, 3, and 5 display the fused images produced by the models for the MFIF, MEIF, and IVF tasks, respectively. Rows 2, 4, and 6 show the corresponding gradient maps. "MEIF-to-MFIF" denotes the result produced by the model trained on the MEIF task and evaluated on the MFIF task and so forth.

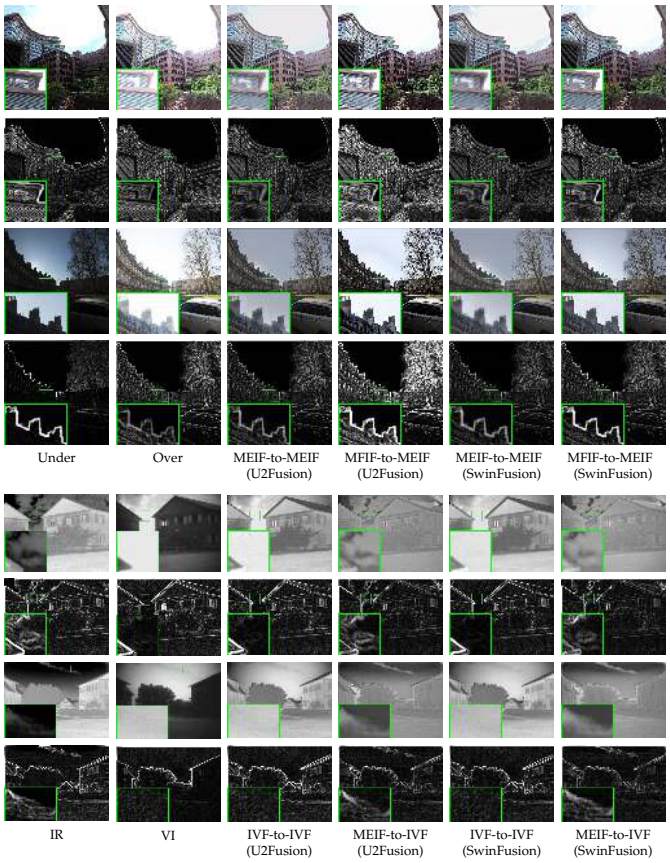

Fig. 3: Experimental results of U2Fusion [9] and SwinFusion [1] trained for specific tasks under both intra-task and inter-task testing scenarios. The first four rows display fused images and their corresponding gradient maps for the MEIF task, while the remaining four rows present results for the IVF task, including fused images and associated gradient maps.

time-consuming and can lead to a lack of interpretability. In contrast, the deep unfolding method addresses these challenges by building image fusion networks through the unfolding of optimization algorithms. In MCSC [46], the authors first proposed a multimodal convolutional sparse dictionary learning model. LRRNet [3] introduced a low-rank representation learning model specifically for multi-modal image fusion problems. The work in [62] addressed the multimodal image fusion problem by unfolding a bilevel optimization algorithm, while AUIF [63] developed an image fusion network by unfolding an image decomposition algorithm. Additionally, M2CDL [64] presented a multi-scale, multimodal convolutional dictionary learning model. While existing deep unfolding methods mainly focus on building networks by unfolding specific optimization algorithms, their application has been limited to multimodal image fusion problems. However, a notable gap remains in the development of deep unfolding methods for general image fusion tasks. Unlike the aforementioned approaches, we built a general image fusion network by unfolding the fundamental fusion rule. This approach not only retains the advantages of existing deep unfolding methods but also enables unified modeling across various image fusion tasks. As a result, our network can adapt to a wide range of image fusion scenarios beyond multimodal image fusion.

3 Proposed Method

In this work, we propose a comprehensive image fusion framework that encompasses both network design and model training. First, we introduce the sequential gradient transfer framework for model training. Then, concerning the network design, we explain the deep-unfolded network based on the fundamental fusion rule.

3.1 Sequential Gradient Transfer Framework

Our approach is motivated by an observation, i.e., models trained for specific tasks often struggle to preserve structural information when tested within the same task. However, these models tend to show complementary structural information when tested across different tasks. Motivated by this observation, we propose a sequential gradient transfer framework to leverage cross-task structural information, thereby enhancing the preservation of structural details from source images.

3.1.1 Key Observation

To illustrate our key observation, we independently trained three models using the same network architecture for the MFIF, MEIF, and IVF tasks (details of the network architecture are discussed in Sect. 3.2). For the MFIF task, we utilized the RealMFF [65] dataset. For the MEIF task, we employed the widely used SICE [66] dataset. For the IVF task,

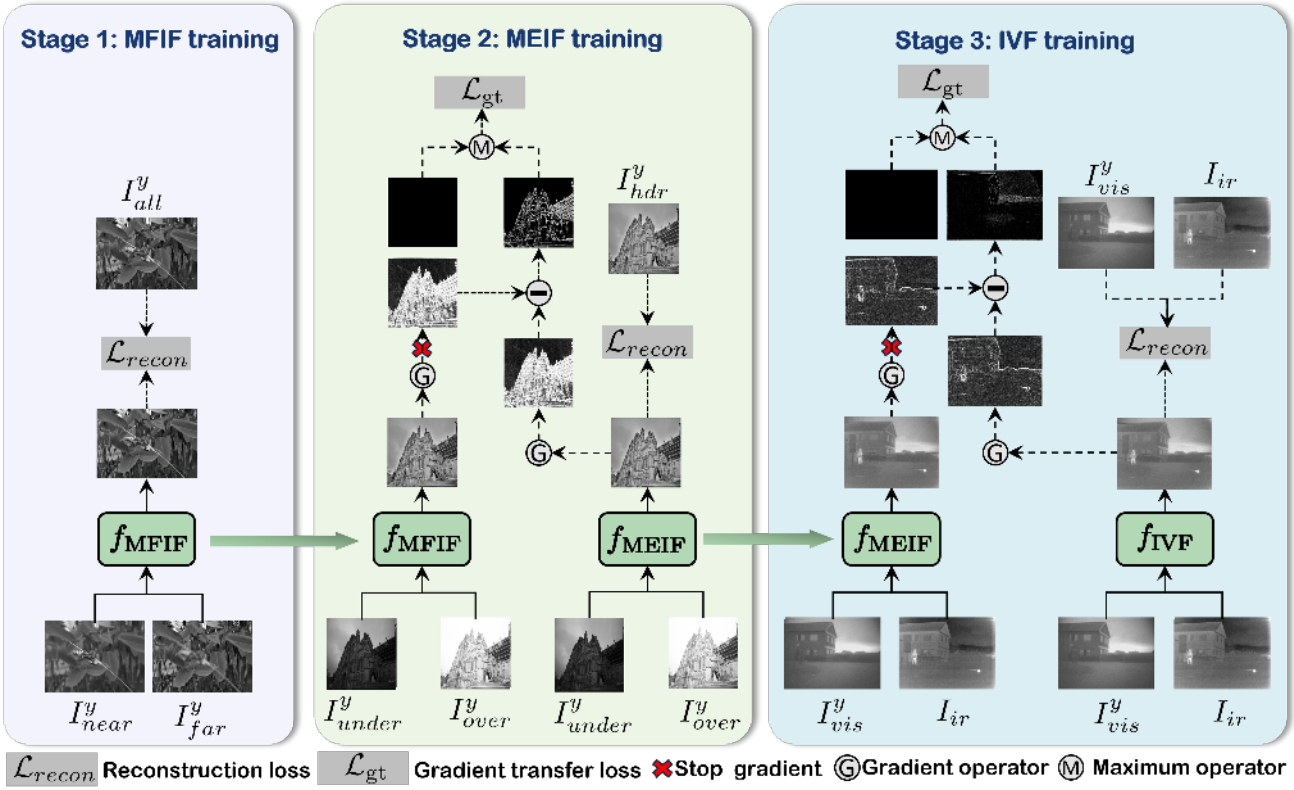

Fig. 4: A schematic diagram of our sequential gradient transfer framework. The framework comprises three training stages, each specifically designed for the MFIF, MEIF, and IVF tasks. We transfer the complementary structural information from the previous stage to the next stage via gradient maps by minimizing the gradient transfer loss according to (3). During this process, only the parameters of the model for the current stage are updated, while the parameters from the previous stage remain unchanged.

we used the TNO [67] dataset. Since the RealMFF and SICE datasets are labeled, we employed a loss function based on the SSIM [68] metric for supervised training. For the unlabeled TNO [67] dataset, we adopted the unsupervised loss function used by SwinFusion [1], which demonstrated high performance.

The test results within the same task are shown in the third column of Fig. 2. For the MFIF task, the performance is satisfactory. However, in the MEIF task, the fusion images lack local details, likely due to significant exposure differences between source images, introducing noise from underexposure and overexposure. For the IVF task, the generated images do not retain the structure from the infrared images, possibly due to significant modality differences and the low visual quality of the source images. Overall, task-specific training methods struggle to retain the structural information of the source images.

The cross-task test results are presented in the fourth and fifth columns of Fig. 2. For the MFIF task, the intra-task performance is strong, and cross-task results do not show complementary information. In the MEIF task, while the model trained on MFIF data introduces noticeable global brightness distortions compared to intra-task results, it demonstrates significant structural information complementarity. Similarly, the IVF task results also reveal such complementarity when cross-tasking is tested. We attribute these observations to the higher data quality in the MFIF task compared to the IVF and MEIF tasks, and models trained on high-quality data tend to better preserve structural information. Evidence supporting this cue includes pre-training models on high-quality simulated data before fine-tuning IVF data as in [4], [22], [23]. By analyzing the gradient maps of the corresponding images, we further verify the existence of structural information complementarity. To verify that this observation reflects a general phenomenon rather than an isolated occurrence and is independent of network architecture, we further trained two representative general image fusion methods, U2Fusion [9] and SwinFusion [1], using the same dataset and training procedure. The trained models were evaluated under both intra-task and cross-task settings. Experimental results are presented in Fig. 3, where two samples from each task are shown for illustration. It can be observed that both U2Fusion [9] and SwinFusion [1] exhibit cross-task transfer effects, a phenomenon consistently evident across multiple samples. This observation led us to develop a learning framework specifically designed to extract and transfer this complementary structural information, aiming to improve the preservation of structural details.

3.1.2 *The Training Framework*

Inspired by the aforementioned observation and considering that image gradients can effectively represent the structural information of images, we propose a three-stage

sequential gradient transfer learning framework (as shown in Fig. 4) to get complementary structure information across tasks. In the three stages, we sequentially trained three models for the MFIF, MEIF, and IVF tasks[1]. We define I_{near}, I_{far}, and I_{all} as the near-focus image, the far-focus image, and the all-focus image (i.e., the reference image), respectively, for the MFIF tasks. Let I_{under}, I_{over}, and I_{hdr} stand for the under-exposure image, the over-exposure image, and the high-dynamic range image (i.e., the reference image), respectively, for the MEIF task. Let I_{vis} and I_{ir} represent the visible image and infrared image, respectively, for the IVF task. We aim to train three models, i.e., f_{MFIF}, f_{MEIF}, and f_{IVF}, using the same network architecture for each task. Given that the datasets include both red-green-blue (RGB) and grayscale images, we convert all RGB images to the YCrCb color space and use only the Y component for training. After obtaining the Y component of the target image, we apply a simple weighted-average combination on the Cr and Cb components of the source images to derive the Cr and Cb components of the target image. The loss functions used at each stage include an image reconstruction loss term and a gradient transfer loss term:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{gt}}, \tag{1}$$

where $\mathcal{L}_{\text{recon}}$ denotes the reconstruction loss, $\mathcal{L}_{\text{gt}}$ indicates the gradient transfer loss, and λ is a weighting coefficient.

To effectively train the network, we divide the whole procedure into three training stages (more details in Fig. 4):

- In the first training stage, considering that the MFIF task can provide complementary structural information for the MEIF and IVF tasks, as well as MFIF can be effectively addressed through task-specific training (as shown in Fig. 2), we initially focus on the training of the MFIF task. The training process optimizes the following reconstruction loss function:

$$\mathcal{L}_{\text{recon}}^{\text{s1}} = 1 - \mathsf{SSIM}(f_{\text{MFIF}}(I_{\text{near}}^{y}, I_{\text{far}}^{y}),\ I_{\text{all}}^{y}), \tag{2}$$

  where $s1$ refers to stage 1, SSIM represents the structural similarity index metric (SSIM) [68], and the variables I_{near}^{y}, I_{far}^{y}, and I_{all}^{y} denote the Y component of the near-focus, the far-focus, and the all-focus images, respectively. At this stage, since the gradient transfer strategy is not applied, the weight of the gradient transfer loss function is set to 0.
- In the second training stage, we proceed to train the model for the MEIF task, leveraging the complementary structural information contained in f_{MFIF}. Given that image gradients effectively capture structural information, we use these image gradients to transfer complementary structural information while avoiding the transfer of potentially negative information. This is achieved by optimizing the following gradient transfer ℓ_1 loss function:

$$\begin{aligned}\mathcal{L}_{\text{gt}} =& \|\mathsf{Max}(\nabla f_{\text{MFIF}}(I_{\text{under}}^{y}, I_{\text{over}}^{y}) \\ & - \nabla f_{\text{MEIF}}(I_{\text{under}}^{y}, I_{\text{over}}^{y}), 0)\|_1,\end{aligned} \tag{3}$$

  where $\|\cdot\|_1$ is the ℓ_1 norm, Max is the maximum operator, ∇ is the gradient operator[2], and I_{under}^{y} and I_{over}^{y} correspond to the Y components of the under-exposed and over-exposed images, respectively. This formulation ensures that only regions where the gradient of $\nabla f_{\text{MFIF}}(I_{\text{under}}^{y}, I_{\text{over}}^{y})$ is larger than that of $\nabla f_{\text{MEIF}}(I_{\text{under}}^{y}, I_{\text{over}}^{y})$ are transferred, as larger gradients indicate more significant structural information. Additionally, we continue to optimize the reconstruction loss as outlined in (2).
- In the third stage, we train the model for the IVF task. The complementary information from the MEIF task is also utilized through gradient transfer to assist in the learning process. The gradient transfer loss is calculated similarly to (3). Due to the unavailability of labeled data, we follow the approach used in SwinFusion [1] and adopt the following reconstruction loss function:

$$\mathcal{L}_{\text{recon}}^{s3} = \mathcal{L}_{\text{int}} + \beta_1 \mathcal{L}_{\text{text}} + \beta_2 \mathcal{L}_{\text{stru}}, \tag{4}$$

  where $\mathcal{L}_{\text{int}}$ denotes the intensity loss, $\mathcal{L}_{text}$ represents the texture loss in the gradient domain, $\mathcal{L}_{\text{stru}}$ is the structural loss, and β_1 and β_2 are weighting coefficients. $\mathcal{L}_{\text{int}}$ can be calculated as:

$$\mathcal{L}_{\text{int}} = \|f_{\text{IVF}}(I_{\text{vis}}^{y}, I_{\text{ir}}) - \mathsf{Max}(I_{\text{vis}}^{y}, I_{\text{ir}})\|_1, \tag{5}$$

  where I_{vis}^{y} denotes the Y component of the visible image[3]. $\mathcal{L}_{\text{text}}$ can be calculated as:

$$\mathcal{L}_{\text{text}} = \|\nabla f_{\text{IVF}}(I_{\text{vis}}^{y}, I_{\text{ir}}) - \mathsf{Max}(\nabla I_{\text{vis}}^{y}, \nabla I_{\text{ir}})\|_1. \tag{6}$$

  Moreover, $\mathcal{L}_{\text{stru}}$ can be calculated as:

$$\begin{aligned}\mathcal{L}_{\text{stru}} = 1 - \mathsf{SSIM}(f_{\text{IVF}}(I_{\text{vis}}^{y}, I_{\text{ir}}), I_{\text{vis}}^{y}) + \\ 1 - \mathsf{SSIM}(f_{\text{IVF}}(I_{\text{vis}}^{y}, I_{\text{ir}}), I_{\text{ir}}).\end{aligned} \tag{7}$$

Remark 1 *(Color Issues for the MEIF task)* As previously discussed, during the training for the MEIF task, the source images are converted to the YCrCb color space, with only the Y components used for training. A simple weighted-average method is then employed to calculate the target Cr and Cb components. However, in the MEIF task, due to significant exposure differences between the source images, this approach can lead to severe color distortion [69], [70]. To address this issue, we designed a ColorNet (details in Sect. 1 of the Supplementary Materials) to perform color correction on the fused images generated during the second stage of training. ColorNet utilizes the Y component of the fused image along with the Cr and Cb components of the source images to predict accurate Cr and Cb components. The ColorNet is trained using the following ℓ_1 loss function:

$$\mathcal{L}_{\text{color}} = \|f_{\text{color}}(I_{\text{under}}^{\text{crcb}}, I_{\text{over}}^{\text{crcb}}, \widehat{I}_{\text{hdr}}^{y}) - I_{\text{hdr}}^{\text{crcb}}\|_1, \tag{8}$$

where f_{color} represents the ColorNet, $I_{\text{under}}^{\text{crcb}}$ and $I_{\text{over}}^{\text{crcb}}$ denote the corresponding CrCb component of the under-exposed and over-exposed images, respectively, $\widehat{I}_{\text{hdr}}^{y} = f_{\text{MEIF}}(I_{\text{under}}^{y}, I_{\text{over}}^{y})$ indicates the fused Y component, and $I_{\text{hdr}}^{\text{crcb}}$ is the reference CrCb component.

1. Since we found that models trained for the MEIF and IVF tasks generalized well to the MIF task, we did not train a model specifically for the MIF task.

2. We use the Sobel operator to obtain the image gradient.

3. Note that infrared images are gray-scale images. Therefore, there is no need to project them into the YCrCb space.

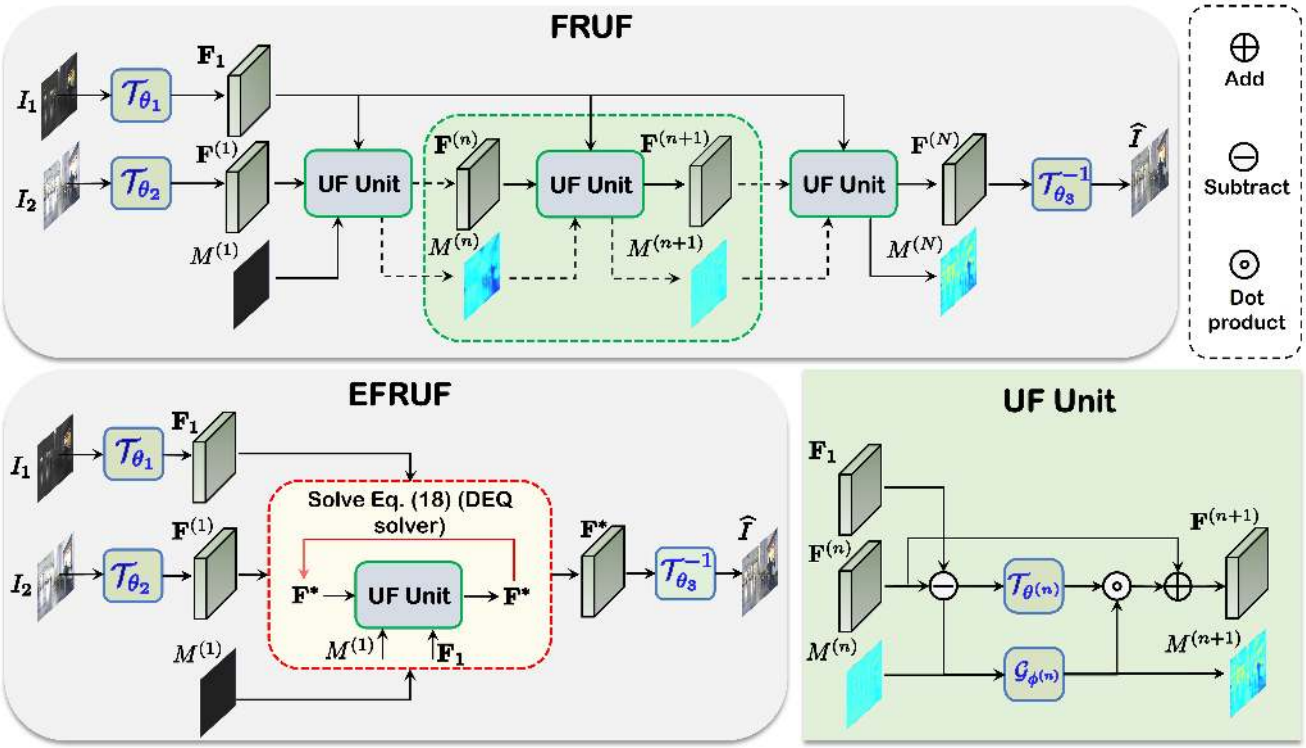

Fig. 5: The overall structure of the proposed FRUFN and EFRUFN. The FRUFN uses a sequential UF Unit to obtain the target image, while the EFRUFN directly solves $\mathbf{F}^*$ with a DEQ solver.

3.2 Network Design

After detailing the sequential gradient transfer training algorithm, we will describe in detail the general image fusion network that we trained using this training algorithm. Our objective is to develop a general image fusion network based on a fundamental fusion principle. We begin by presenting the general formulation of this rule. Next, we expand this formulation to create a learnable image fusion algorithm. We then integrate this algorithm with the DEQ model to propose a more efficient solution. Finally, we parametrize these algorithms to derive two general image fusion networks named FRUFN and EFRUFN.

3.2.1 Fusion Rule Unfolding Network (FRUFN)

(I) Mathematical Derivation

For a wide range of image fusion tasks, including MFIF, MEIF, IVF, and MMF, there exists a fundamental fusion principle that can facilitate a unified modeling approach to address these problems. Let matrix I_1 and I_2 denote the single-band source images, and $\widehat{I}$ indicate the fusion result. The fundamental image fusion rule, see e.g., [5], [45], [71], can be generally formulated as:

$$\widehat{I} = \mathcal{T}^{-1}(M \odot \mathcal{T}_1(I_1) + (1 - M) \odot \mathcal{T}_2(I_2)), \tag{9}$$

where $\mathcal{T}_1(\cdot)$ and $\mathcal{T}_2(\cdot)$ are transformations[4] that project the single-band source images into a high-dimensional space, $\mathcal{T}^{-1}$ is the inverse transformation, $\odot$ denotes the element-wise multiplication, and M is a mask corresponding to the fusion rule[5]. In the transformed high-dimensional space, we have:

$$\mathbf{F} = M \odot \mathbf{F}_1 + (1 - M) \odot \mathbf{F}_2, \tag{10}$$

where $\mathbf{F}_1 = \mathcal{T}_1(I_1)$ and $\mathbf{F}_2 = \mathcal{T}_2(I_2)$ are the transformed source images, and $\widehat{I} = \mathcal{T}^{-1}(\mathbf{F})$.

We then unfold (10) to an iterative process:

$$\mathbf{F}^{(n+1)} = M^{(n)} \odot \mathbf{F}_1 + (1 - M) \odot \mathbf{F}^{(n)}, \tag{11}$$

where $n \in \mathbb{N}^+$ is the n-th iteration and the initialization is $\mathbf{F}^{(1)} = \mathbf{F}_2$. Compared to (10), this new formulation allows for the use of different fusion rules in different iterations, making it more flexible. It is worth noting that many mainstream neural networks are now composed of a series of identical basic units with residual connections. Our goal is to design a general image fusion network based on (11). Inspired by these popular network architectures, we rewrite (11) in a residual form:

$$\mathbf{F}^{(n+1)} = \mathbf{F}^{(n)} + M^{(n+1)} \odot (\mathbf{F}_1 - \mathbf{F}^{(n)}). \tag{12}$$

However, in (12), $\mathbf{F}^{(n)}$ and $\mathbf{F}_1$ only interact indirectly through $M^{(n+1)}$, which is not efficient enough. To address this issue, we propose to introduce an additional transformation, i.e., $\mathcal{T}^{(n)}$, to enhance the interaction capability:

$$\mathbf{F}^{(n+1)} = \mathbf{F}^{(n)} + M^{(n+1)} \odot \mathcal{T}^{(n)}(\mathbf{F}_1 - \mathbf{F}^{(n)}). \tag{13}$$

4. $\mathcal{T}_1$ and $\mathcal{T}_2$ can be the same or different.

5. This formulation encompasses both traditional transformation-based algorithms and modern deep learning methods, such as U2Fusion [9], YDTR [32], and SwinFusion [1], when transformations are learnable.

Afterward, we choose to parametrize the fixed mask, $M^{(n+1)}$, and the transformation, $\mathcal{T}^{(n)}$, to derive a learnable algorithm:

$$\mathbf{F}^{(n+1)} = \mathbf{F}^{(n)} + \mathcal{G}_{\phi^{(n)}}(\mathbf{F}_1 - \mathbf{F}^{(n)}, M^{(n)}) \odot \mathcal{T}_{\theta^{(n)}}(\mathbf{F}_1 - \mathbf{F}^{(n)}), \tag{14}$$

where $\mathcal{T}_{\theta^{(n)}}$ is a learnable transformation with $\theta^{(n)}$ being the parameters, $\mathcal{G}_{\phi^{(n)}}$ is the mask generator with $\phi^{(n)}$ being the parameters, and $M^{(n+1)} = \mathcal{G}_{\phi^{(n)}}(\mathbf{F}_1 - \mathbf{F}^{(n)}, M^{(n)})$. $\mathcal{G}_{\phi^{(n)}}$ not only uses the residual, i.e., $\mathbf{F}_1 - \mathbf{F}^{(n)}$, but also uses the learned mask of the last iteration, i.e., $M^{(n)}$, to predict the mask for the current iteration. In addition, we also parametrize the initial transformations, i.e., $\mathcal{T}_1$ and $\mathcal{T}_2$, to be learnable transformations, i.e., $\mathcal{T}_{\theta_1}$ and $\mathcal{T}_{\theta_2}$. After reaching the predetermined maximum number of iterations, i.e., N, the fusion result can be obtained by applying a learnable inverse transformation, $\mathcal{T}_{\theta_3}^{-1}$:

$$\widehat{I} = \mathcal{T}_{\theta_3}^{-1}(\mathbf{F}^{(N)}). \tag{15}$$

The proposed algorithm leverages the strengths of both deep learning models and traditional transform-based approaches, as summarized in Algorithm 1 (see supplementary materials). It allows the use of state-of-the-art networks such as ResNet [56] and Vision Transformers [72] to parametrize the transformations and mask generators. At the same time, it adapts to different fusion rules through learnable masks, enhancing the flexibility of the fusion process.

(II) The Derivation-based Network Architecture Design

In this section, we propose a corresponding deep-learning network architecture, namely the fusion rule unfolding network (FRUFN), based on the derivation analysis in Sect. 3.2.1(I).

As illustrated in Fig. 5, the FRUFN accepts the source image I_1 and the initial mask, $M^{(1)}$, to iteratively refine the transformed source image I_2 and predict a new mask. We initialize $M^{(1)}$ as a zero matrix. The FRUFN comprises three main components: i) initial transformations, i.e., $\mathcal{T}_{\theta_1}$ and $\mathcal{T}_{\theta_2}$, which project the source images into the feature space; ii) the unfolding unit (UF Unit), which iteratively refines the fused outcome by extracting complementary information from I_1; and iii) the inverse transformation $\mathcal{T}_{\theta_3}^{-1}$, which reconstructs the final fused image from the features.

We empirically design $\mathcal{T}_{\theta_1}$ and $\mathcal{T}_{\theta_2}$ to have a convolutional layer and two residual blocks as shown in Fig. 6. The UF Unit is designed by parametrizing $\mathcal{T}_{\theta^{(n)}}$ and $\mathcal{G}_{\phi^{(n)}}$ is based on (14). Regarding $\mathcal{T}_{\theta^{(n)}}$ whose goal is to extract abundant complementary information from two source images, we model it with the global-local transformation (GLT) based on a CNN-Transformer hybrid structure. The dynamic learnable mask, $\mathcal{G}_{\phi^{(n)}}$, is responsible for learning semantic attention weights to facilitate the feature interaction from the two source images. We model it with the multi-scale mask generators (MSMG).

Global-Local Transformation (GLT). Since $\mathcal{T}_{\theta^{(n)}}$ is used to learn complementary features from source images, which include both local features (e.g., high-frequency details in visible images) and global features (e.g., saliency targets in infrared images), we propose the GLT (see the lower part of Fig. 6) to model $\mathcal{T}_{\theta^{(n)}}$. GLT has a CNN-Transformer hybrid structure in which the CNN is responsible for learning local features and the transformer is responsible for learning global features. Specifically, we use two convolutional layers to learn local features. We use the Swin-Transformer block (SwinBlock) [58] to learn long-range features as it can deal with images with various resolutions and it is memory efficient due to the shift-window mechanism. Owing to the unstable convergence of the SwinBlock during training, two-layer normalization [73] layers are inserted in-between the convolutional layers and the SwinBlock to stabilize the training.

Multi-Scale Mask Generator (MSMG). To model the adaptively learnable mask, the neural network should be able to explore the semantic information. Therefore, the neural network must have a large receptive field. Inspired by the architectures addressing the semantic segmentation task [74], we design the mask generators to be a multi-scale structure (see Fig. 6). Specifically, the features are passed to a convolution layer and then downsampled with max pooling by a factor of 2, 4, and 8 to produce multi-scale features. In each scale, we use a ConvNext Block [75] to enrich multi-scale features. Afterward, multi-scale features are aggregated and further sent to a GLT module to extract more semantic information. Finally, the extracted semantic information is projected to a two-channel feature map and normalized by SoftMax along channel dimension to produce the semantic map. Thus, the semantic map can automatically distinguish information from the two source images and flexibly assign appropriate weights to them.

3.2.2 *Efficient Fusion Rule Unfolding Network (EFRUFN)*

(I) Mathematical Derivation

Since the proposed Algorithm 1 employs distinct parameters in each iteration, escalating the iteration count will not only augment the number of parameters but also increase memory usage and computational complexity. To tackle this challenge, we propose an efficient strategy based on (14) to facilitate parameter sharing across iterations, thereby substantially reducing the number of parameters required:

$$\mathbf{F}^{(n+1)} = \mathbf{F}^{(n)} + \mathcal{G}_{\phi}(\mathbf{F}_1 - \mathbf{F}^{(n)}, M^{(n)}) \odot \mathcal{T}_{\theta}(\mathbf{F}_1 - \mathbf{F}^{(n)}), \tag{16}$$

where the learnable parameters ϕ and θ are shared for all iterations.

While this weight-sharing approach is more parameter-efficient, it retains the original algorithm's memory and computational demands. By considering the weight-share algorithm as a fixed-point process, we can apply the DEQ [59] model to solve this problem. Specifically, taking the right side of (16) as a uniformed variable, $\mathbf{F}^{(n)}$, it can be reformulated as follows:

$$\mathbf{F}^{(n+1)} = \mathcal{H}_{\theta,\phi}(\mathbf{F}^{(n)}), \tag{17}$$

where $\mathcal{H}_{\theta,\phi}(\mathbf{F}^{(n)}) = \mathbf{F}^{(n)} + M^{(n+1)} \odot \mathcal{T}_{\theta}(\mathbf{F}_1 - \mathbf{F}^{(n)})$ with $M^{(n+1)} = \mathcal{G}_{\phi}(\mathbf{F}_1 - \mathbf{F}^{(n)}, M^{(n)})$. If the iterative process converges, updating the iterative process will not alter $\mathbf{F}^{(n+1)}$. Let $\mathbf{F}^*$ represent the equilibrium point (or convergence point), the fixed-point iteration process can be expressed as:

$$\mathbf{F}^* = \mathcal{H}_{\theta,\phi}(\mathbf{F}^*), \tag{18}$$

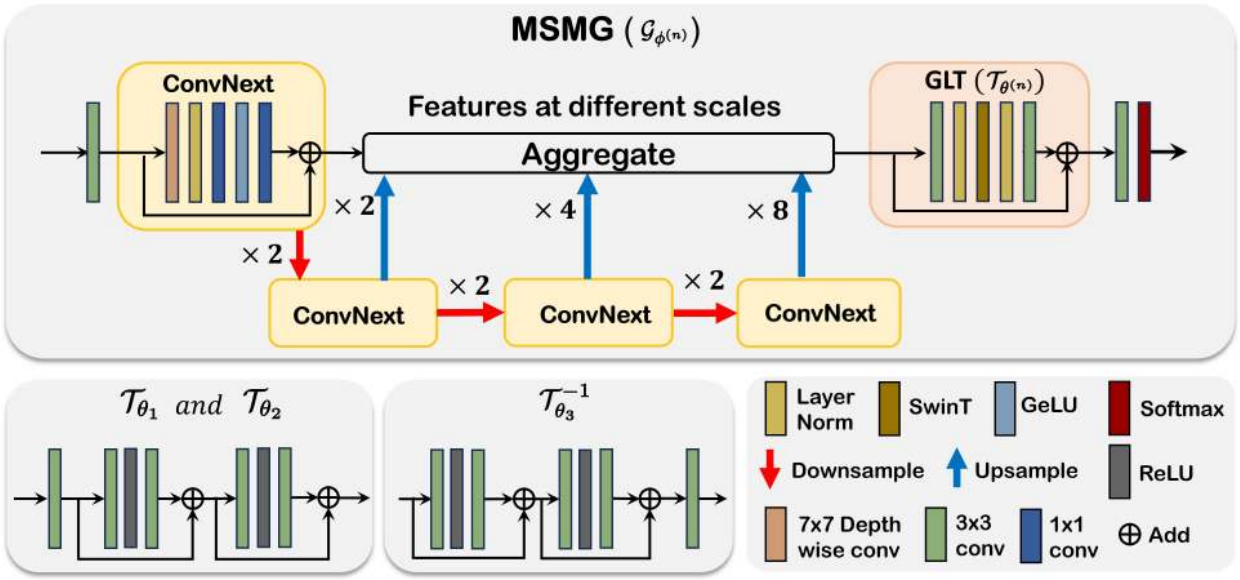

Fig. 6: The detailed structure of the MSMG ($\mathcal{G}_{\phi^{(n)}}$), the GLT ($\mathcal{T}_{\theta^{(n)}}$), $\mathcal{T}_{\theta_1}$, $\mathcal{T}_{\theta_2}$, and $\mathcal{T}_{\theta_3}^{-1}$. The GLT is employed to extract complementary long-range and short-range features from the source images, while the MSMG is used to explore semantic information to learn dynamic fusion rules.

which can be efficiently solved by DEQ [59]. We summarize the final efficient algorithm in Algorithm 2 of the Supplementary Materials. Specifically, the DEQ algorithm mainly consists of the forward and backward processes.

In the forward process, the DEQ [59] model can utilize accelerated algorithms[6], e.g., Broyden's method [76] or Anderson mixing [77], to solve the fixed point iteration problem. In the meantime, the DEQ [59] model does not store all intermediate results. Thus, the computational complexity and memory consumption can be largely reduced. Theoretically, the training memory of the efficient algorithm remains constant while the training memory of the original algorithm grows linearly with the number of iterations.

In the backward process, since the DEQ [59] model does not store all feature maps in the forward propagation, the gradient of the model parameters cannot be directly calculated with the chain rule. Instead, the gradient of the model parameters is calculated indirectly through implicit differentiation. For example, given the fixed-point representation as shown in (18), the gradient with respect to θ[7] of the corresponding loss $\mathcal{L}$ is given by:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \mathbf{F}^*}\left(E - \frac{\partial \mathcal{L}}{\partial \mathbf{F}^*}\right)^{-1}\frac{\partial \mathcal{H}_{\theta,\phi}(\mathbf{F}^*)}{\partial \theta}, \tag{19}$$

where E is the identity matrix. According to this equation, we only need the final output $\mathbf{F}^*$ and do not need to acquire all intermediate results to calculate the gradients. However, computing $(E - \frac{\partial \mathcal{L}}{\partial \mathbf{F}^*})^{-1}$ in (19) is still difficult as calculating the inverse of a very high-dimensional matrix is not easy. In the implementation, this term can only be approximately calculated. A recent work in [78] proposed to approximate $(E - \frac{\partial \mathcal{L}}{\partial \mathbf{F}^*})^{-1}$ by $E \approx (E - \frac{\partial \mathcal{L}}{\partial \mathbf{F}^*})^{-1}$ and experimentally verified that this strategy does not harm the performance. Specifically, the gradient can be approximately calculated by:

$$\frac{\partial \mathcal{L}}{\partial \theta} \approx \frac{\partial \mathcal{L}}{\partial \mathbf{F}^*}\frac{\partial \mathcal{H}_{\theta,\phi}(\mathbf{F}^*)}{\partial \theta}. \tag{20}$$

(II) The Derivation-based Network Architecture Design Based on (18), we propose an efficient version of the FRUFN, named the EFRUFN. Fig. 5 (b) demonstrates the overall flow of the proposed EFRUFN. The EFRUFN shares similar factors with respect to the FRUFN, i.e., T_{θ_1}, $\mathcal{T}_{\theta_2}$, and $\mathcal{T}_{\theta_3}^{-1}$. We also model $\mathcal{H}_{\theta,\phi}(\mathbf{F}^*)$ with a UF Unit. Instead of obtaining the transformed fused outcome by sequentially applying the UF Unit like the FRUFN, the EFRUFN directly calculates $\mathbf{H}^*$ with a root-finding method. Thus, the EFRUFN is more parameter-efficient and memory-efficient than the FRUFN.

4 Experiments

This section is devoted to the description of the experimental setup and the outcomes of our experiments. We will first focus on the MFIF, MEIF, and IVF tasks. Afterward, we will highlight the generalization capabilities of our model testing it on the MIF task. Moreover, we will verify the efficiency of the proposed designs via some ablation studies. Finally, we will provide some discussions, including critical aspects as number of parameters and training memory requirements.

4.1 Experimental Setup

4.1.1 Network Configuration

Regarding the network configurations, the number of feature maps of the SwinBlock [58], the ConvNext block, and the ResBlock [56] in the FRUFN and the EFRUFN is set to 32. The depths and heads of the SwinBlock [58] for the GLT and the MSMG are set to 2 and 6, respectively. The window sizes of the SwinBlock [58] for the GLT and the MSMG are set to 2 and 4, respectively. The larger window

6. Note that the fixed-point process can be solved by updating (17) until it converges. However, accelerated algorithms can take much fewer steps to reach the fixed point.

7. Calculating the gradient of θ_2 or ϕ is the same as θ.

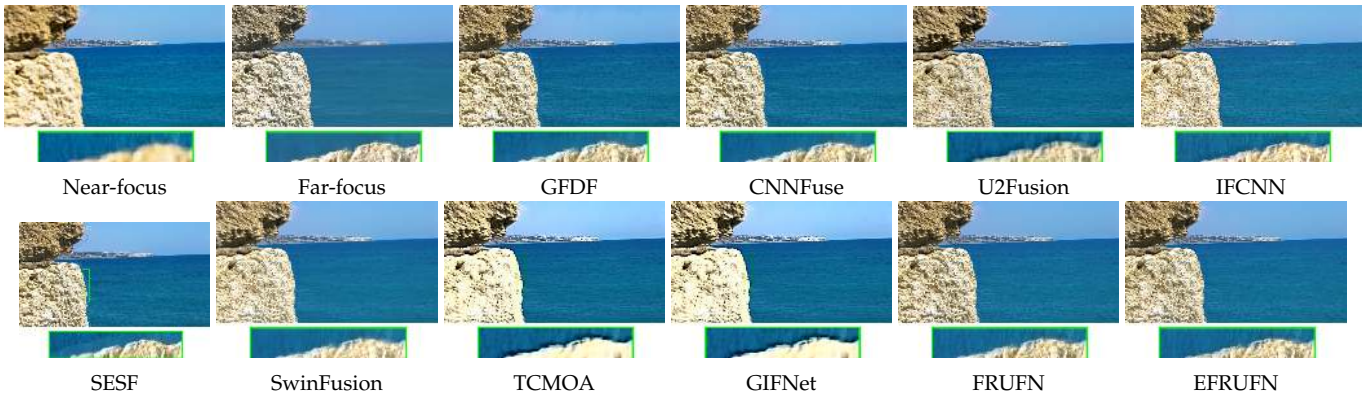

Fig. 7: Visual comparisons for all compared approaches on the Lytro dataset.

size of the MSMG is designed to increase the receptive field. The number of iterations for the FRUFN is set to 5. For the EFRUFN, the maximum number of iterations is also set to 5 (For more implementation details, please refer to Section 2 of the Supplementary Materials.).

4.1.2 Dataset Information

For the MFIF task, we employ the RealMFF [65] and MFI-WHU [79] datasets. The RealMFF [65] dataset consists of real-world samples and we selected 700 samples for training (The dataset configuration is provided in Table 1 of the Supplementary Materials.). The WHU-MFI [79] dataset is a simulated dataset and we selected 96 samples for training. We used the SICE [66] dataset for the training of the MEIF task. SICE [66] is a simulated dataset and we selected 201 pairs of samples for training, while the remaining sample pairs are used for testing. Since each sample pair contains multiple sets of underexposed and overexposed images, we randomly selected one pair of underexposed and overexposed images from each sample pair for training. We selected 190 samples from the RoadScene (RS) [80] dataset and 30 samples from the TNO [67] dataset for the training of the IVF task.

4.1.3 Metrics

For the MFIF, IVF, and MIF tasks, where reference (ground-truth) data on the test set are unavailable, we use EN [83], MI [84], SCD [85], and MS-SSIM[8] [86] for quantitative evaluation. In addition to these metrics calculated using shallow features or information theory, we have also incorporated UNIQUE [87], a deep learning-based metric. Unlike the former indicators, this one places a greater emphasis on perceptual quality. For the MEIF task with reference images, we used PSNR, MEF-SSIM [88], and MS-SSIM [86] to evaluate the fidelity between the fused image and the reference image. The code for the metrics is provided by [89][9].

4.2 MFIF Experiments

For the MFIF experiments, the Lytro [94] dataset is used. The original Lytro [94] dataset has 38 image pairs, we remove the low-visual quality and unregistered image pairs and use the remaining 22 image pairs for testing. We compare our methods with non-general image fusion methods, which include GFDF [7], CNNFuse [91], SESF[10] [90], ZMFF[11] [92], and general image fusion methods, which include SwinFusion[12] [1], U2Fusion[13] [9], IFCNN[14] [45], TCMOA[15] [16], and GIFNet[16] [93].

TABLE 1: Average metrics of all compared DL-based approaches on 22 samples from the Lytro dataset. The results for the general methods are shown in light yellow. The best, second-best, and third-best results are highlighted in red, blue, and bold, respectively.

Method	MS-SSIM↑	EN↑	MI↑	SCD↑	UNIQUE↑	#Param. (M)
GFDF [7]	0.9917	7.4883	14.9767	0.5622	1.16	\
SESF [90]	0.9910	**7.4890**	**14.9779**	0.5638	**1.175**	\
CNNFuse [91]	0.9919	7.4865	14.9730	0.5603	1.146	0.2M
ZMFF [92]	0.9888	7.4835	14.9670	0.4519	1.026	4.67M
SwinFusion [1]	0.9892	7.4674	14.9348	0.6023	0.867	0.973M
U2Fusion [9]	0.9777	7.2757	14.5513	1.0203	0.803	2.636M
IFCNN [45]	**0.9922**	7.4693	14.9387	0.6340	1.17	0.083M
TCMOA [16]	0.9908	7.4833	14.9667	0.5768	1.11	9.58M
GIFNet [93]	0.8833	7.4141	14.8282	0.5404	0.896	3.291M
FRUFN	0.9938	7.4943	14.9887	0.6858	1.24	0.864M
EFRUFN	0.9937	7.4943	14.9887	**0.6768**	1.24	**0.236M**

4.2.1 Visual Results

Fig. 7 shows the fused images on the Lytro [94] dataset. An important criterion for evaluating the performance of an MFIF algorithm is whether it can generate small distortions near the decision boundary (the boundary between the far-focus area and the near-focus area). For ease of analysis, we also present enlarged image patches. From the figure, it can be seen that GFDF [7], CNNFuse [91], and U2Fusion [9] lost far-focus details near the decision boundary. SESF [90] and IFCNN [45] generate inaccurate decision boundaries, while our methods can generate accurate decision boundaries and preserve both far-focus and near-focus details. Images fused by TCMOA [16] and GIFNet [93] exhibit not only an ill-

8. It is worth noting that the MS-SSIM metric is generally considered to better reflect the structural information fidelity of the fused results.
9. https://github.com/xingchenzhang/MFIF
10. https://github.com/Keep-Passion/SESF-Fuse
11. https://github.com/junjun-jiang/ZMFF
12. https://github.com/Linfeng-Tang/SwinFusion
13. https://github.com/hanna-xu/U2Fusion
14. https://github.com/uzeful/IFCNN
15. https://github.com/YangSun22/TC-MoA
16. https://github.com/AWCXV/GIFNet

defined decision boundary but also a significant blurring in near-focus regions.

4.2.2 Quantitative Results

Tab. 1 reports the quantitative assessment. Our method outperforms the other approaches considering MS-SSIM [86], EN [83], MI [84], and UNIQUE [87] as quality metrics. Although U2Fusion [9] achieves the best results in the SCD [85] metric, it performs poorly in the remaining three metrics. Based on the visual comparison experiments, the fusion results generated by U2Fusion [9] are excessively blurry. Overall, without significantly increasing the parameter count, our method clearly showed advantages in the used quality metrics. Furthermore, the EFRUFN achieves nearly identical quantitative results to the FRUFN with only approximately a quarter of the parameters.

4.3 MEIF Experiments

For the MEIF experiments, we exploited the SICE dataset [66]. To manage computational resources and ensure consistency, we resized the high-resolution original images to 512×512 before conducting the testing. We compare our methods with several general methods, including IFCNN [45], U2Fusion [9], and SwinFusion [1], as well as non-general methods, including BHFMEF[17] [70], DPEMEF [18] [69], IID[19] [2], TransMEF[20] [95], TCMOA [16], and GIFNet [93].

4.3.1 Visual Results

In Fig. 8, we show the fused images generated by the compared approaches. From the region indicated by the red arrow, it can be observed that the compared methods fail to effectively preserve structural details in the background sky area. Moreover, the images generated by IID [2], TransMEF [95], SwinFusion [1], and U2Fusion [9] exhibit underexposure, while the fusion image generated by IFCNN [45] appears overexposed. Images fused by TCMOA [16] exhibit structural distortions, while those fused by GIFNet [93] appear noticeably under-exposed. In comparison to these competitors, our method can better preserve the structural information of the source images and generate images with the proper exposure.

4.3.2 Quantitative Results

From Tab. 2, it can be observed that both the FRUFN and EFRUFN significantly outperform the other approaches for all the metrics. The FRUFN and the EFRUFN obtain similar performance. In terms of model parameter count, most methods have parameter counts within the same order of magnitude[21].

TABLE 2: Average metrics of all compared DL-based approaches on 27 samples from the SICE [66] dataset. The results for the general image fusion methods are shown in light yellow. The best, second-best, and third-best results are highlighted in red, blue, and bold, respectively.

Method	MS-SSIM↑	MEF-SSIM↑	PSNR↑	#Param. (M)
DPEMEF [69]	0.8178	0.7008	19.1137	13.603M
BHFMEF [70]	0.8154	0.7022	**19.9319**	0.03M
IID [2]	0.8149	0.6966	19.6203	**0.36M**
TransMEF [95]	0.8127	0.6584	18.9058	19.052M
SwinFusion [1]	**0.8393**	**0.7218**	19.5695	0.973M
IFCNN [45]	0.8310	0.7112	16.5957	0.083M
U2Fusion [9]	0.8267	0.6996	17.2494	2.636M
TCMOA [16]	0.8085	0.6729	19.5999	9.58M
GIFNet [93]	0.7829	0.6239	14.5419	3.291M
FRUFN	0.8507	0.7280	21.6278	0.994M
EFRUFN	0.8500	0.7305	21.7333	0.368M

4.4 IVF Experiments

We used the RS [80] and TNO [67] datasets in this case. We compare our method with both general image fusion methods and non-general image fusion methods. The non-general image fusion methods are RecoNet[22] [26], LRRNet[23] [3], EMMA[24] [96], and MMIF-INet[25] [97]. The general image fusion methods include SwinFusion [1], CDDFuse[26] [4], U2Fusion [9], IFCNN [45], TIMFuse[27] [52], TCMOA [16], LFDT[28] [98], and GIFNet [93].

4.4.1 Visual Results

The visual results on the TNO dataset are presented in Fig. 9, while the corresponding results for the RS dataset are provided in Fig. 2 of the Supplementary Materials. As shown in Fig. 9, from the close-ups in the green rectangular boxes, it can be seen that ReCoNet [26], LRRNet [3], and TIMFuse [52] lost the structural information of the IR image. From the close-ups in the red rectangular boxes, it is evident that ReCoNet [26], SwinFusion [1], CDDFuse [4], EMMA [96], and TIMFuse [52] lost the structural information of the IR image. The close-ups in the blue rectangular boxes show that the fused results produced by IFCNN [45], SwinFusion [1], and CDDFuse [4] are too blurry. Images fused by TCMOA [16] exhibit insufficient detail rendition in background regions, while those fused by GIFNet [93] suffer from excessively low brightness in salient foreground structures. Additionally, the outcomes generated by U2Fusion [9] have too low brightness. Overall, the results generated by our method can better preserve the structural information of the source images.

4.4.2 Quantitative Results

Tab. 3 reports the quantitative assessment. Our method demonstrates competitive performance for most metrics with respect to the other general image fusion methods, though it slightly performs worse than the non-general

17. https://github.com/ZhiyingDu/BHFMEF
18. https://github.com/dongdong4fei/DPE-MEF
19. https://github.com/HaoZhang1018/IID-MEF
20. https://github.com/miccaiif/TransMEF
21. For the MEIF task, we additionally designed a ColorNet to correct the color of the fused images. As a result, the parameter count of the FRUFN and the EFRUFN is slightly higher.
22. https://github.com/dlut-dimt/ReCoNet
23. https://github.com/hli1221/imagefusion-LRRNet
24. https://github.com/Zhaozixiang1228/MMIF-EMMA
25. https://github.com/HeDan-11/MMIF-INet
26. https://github.com/Zhaozixiang1228/MMIF-CDDFuse
27. https://github.com/LiuZhu-CV/TIMFusion
28. https://github.com/BOYang-pro/LFDT-Fusion

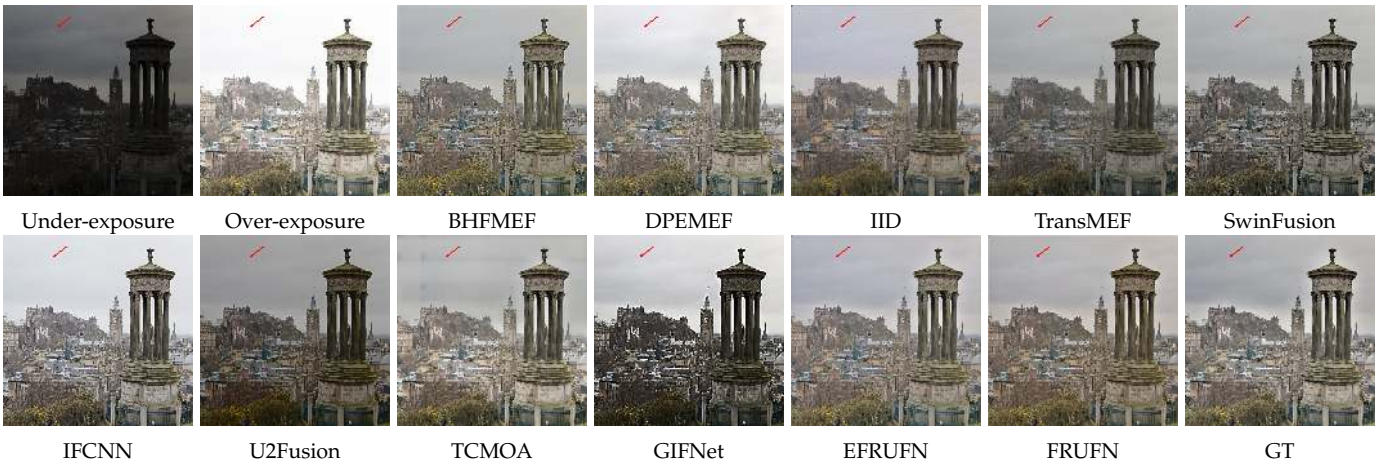

Fig. 8: Visual comparisons of all compared approaches on the SICE dataset.

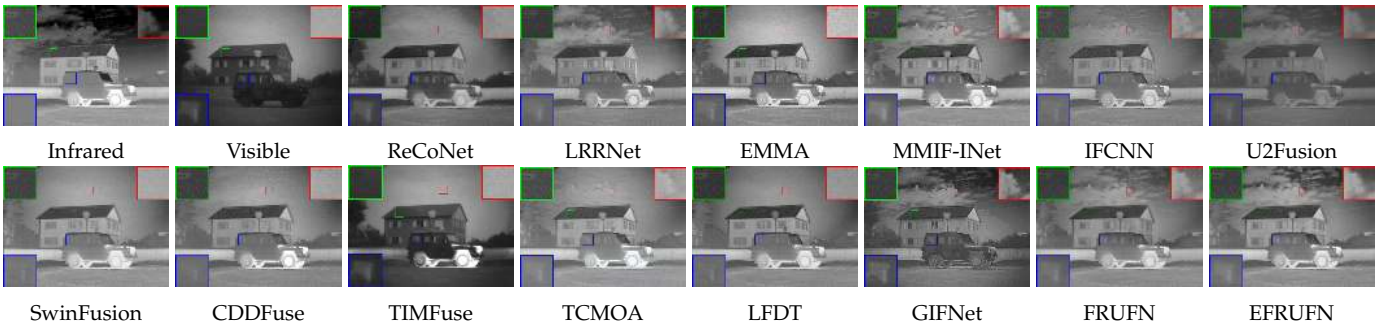

Fig. 9: Visual comparisons of all compared approaches on the TNO dataset.

image fusion method EMMA [96][29]. However, our method offers two notable advantages compared to EMMA. Firstly, the EFRUFN has a significantly smaller parameter count. Secondly, both the FRUFN and EFRUFN outperform EMMA in the MS-SSIM [86] metric, indicating a superior ability to preserve structural information.

4.5 Generalization Experiments

In this section, we directly apply the model trained in the second stage to the MIF task to test the model's generalization ability. We selected two widely studied MIF tasks, i.e., MRI-CT image fusion and MRI-SPECT image fusion. The images used for testing are from the Harvard Medical Image dataset[30]. Specifically, we randomly chose 20 pairs of images for the testing of the MRI-CT image fusion task and 20 for the testing of the MRI-SPECT image fusion task. We compare our methods with non-general methods, which include GeSeNet[31] [99], FusionMamba[32] [100], MMIF-INet [97], and general methods, which include SwinFusion [1], CDDFuse [4], U2Fusion [9], IFCNN [45], TIMFuse [52], LFDT [98], and GIFNet [93].

29. It is noteworthy that the UNIQUE metric is predicted by a pre-trained neural network; therefore, negative values are normal.
30. http://www.med.harvard.edu/AANLIB/home.html
31. https://github.com/lok-18/GeSeNet
32. https://github.com/millieXie/FusionMamba?tab=readme-ov-file

4.5.1 *Visual Results*

The visual results for the MRI-SPECT image fusion task are shown in Fig. 10. From the close-ups in the green rectangular boxes, it can be observed that SwinFusion [1] loses structural information in the MRI image, and the results generated by U2Fusion [9] and TIMFuse [52] appear excessively blurry in terms of structural details. From the close-ups in the red rectangular boxes, it can be seen that MATR [33] and CDDFuse [4] lose structural information in the SPECT image. Images produced by GIFNet [93] and FusionMamba [100] exhibit loss of fine-grained detail information from the original MRI image. Overall, our method generates fusion images that simultaneously preserve the structural information of both the MRI and SPECT images.

Fig. 11 depicts the fused products for the MRI-CT image fusion task. From the close-ups in the rectangular boxes, the outcomes of the FRUFN and EFRUFN show sharper edges and higher contrast than the other approaches. It is worth noting that our model has not been trained on MIF data, yet it still generates high-quality fusion images. This indicates that our method exhibits very good generalization capabilities, allowing it to perform well even on unseen MIF data.

4.5.2 *Quantitative Results*

Tab. 4 reports the quantitative assessment showing that the FRUFN clearly overcomes the other compared approaches in almost all cases. The EFRUFN also shows strong performance for many quality metrics. It is noteworthy that

TABLE 3: Average metrics of all compared DL-based approaches on 26 samples from the TNO dataset and 20 samples from the RS dataset. The results for the general methods are presented in light yellow. The best, second-best, and third-best results are highlighted in red, blue, and bold, respectively.

Method	Venue	TNO					RS					#Param. (M)
		MS-SSIM↑	EN↑	MI↑	SCD↑	UNIQUE↑	MS-SSIM↑	EN↑	MI↑	SCD↑	UNIQUE↑	
ReCoNet [26]	ECCV'22	0.8953	6.8035	13.6070	1.7323	-0.886	0.8642	7.3858	14.7716	1.2747	-1.15	0.007M
LRRNet [3]	TPAMI'23	0.7716	6.7207	13.4414	1.6810	-0.98	0.7791	7.2669	14.5338	0.9236	-1.177	0.049M
EMMA [96]	CVPR'24	0.8939	7.1802	14.3604	1.7050	-0.905	0.8665	7.6774	15.3549	1.6081	-0.926	1.52M
MMIF-INet [97]	IF'25	0.9530	6.9260	13.8520	1.6992	-0.943	**0.9294**	7.4843	14.9687	1.1078	-0.793	0.749M
IFCNN [45]	IF'20	0.9132	6.6554	13.3107	1.6927	-0.806	0.9154	7.3933	14.7867	1.1155	-0.76	**0.083M**
U2Fusion [9]	TPAMI'20	0.8405	6.2690	12.5380	1.6423	-0.91	0.8768	7.2037	14.4074	1.0146	-1.04	2.636M
SwinFusion [1]	JAS'22	0.8999	6.7819	13.5639	1.7290	-0.775	0.8469	7.2072	14.4144	1.3199	-1.12	0.973M
CDDFuse [4]	CVPR'23	0.9069	6.7819	**14.0701**	**1.7918**	-0.82	0.8977	**7.6017**	**15.2033**	1.6102	-0.803	1.188M
TIMFuse [52]	TPAMI'24	0.7745	7.1492	14.2985	1.2740	-0.88	0.6245	7.0683	14.1367	0.8264	-1.27	1.232M
TCMOA [16]	CVPR'24	0.9224	6.7713	13.5427	1.6286	**-0.76**	0.9345	7.4447	14.8895	1.0427	-1.126	9.58M
LFDT [98]	IF'25	0.8670	6.8402	13.6804	1.5817	-0.74	0.8585	7.3815	14.7631	1.0819	**-0.735**	20.28M
GIFNet [93]	CVPR'25	0.9023	6.8772	13.7544	1.8515	-0.38	0.9055	7.6521	15.3042	1.5391	-0.848	3.291M
FRUFN	-	**0.9374**	6.8625	13.7251	1.7573	-0.837	0.9537	7.5641	15.1283	1.4607	-0.66	0.864M
EFRUFN	-	0.9489	**6.8991**	13.7983	1.8410	-0.867	0.9498	7.5762	15.1524	**1.5555**	-0.70	0.236M

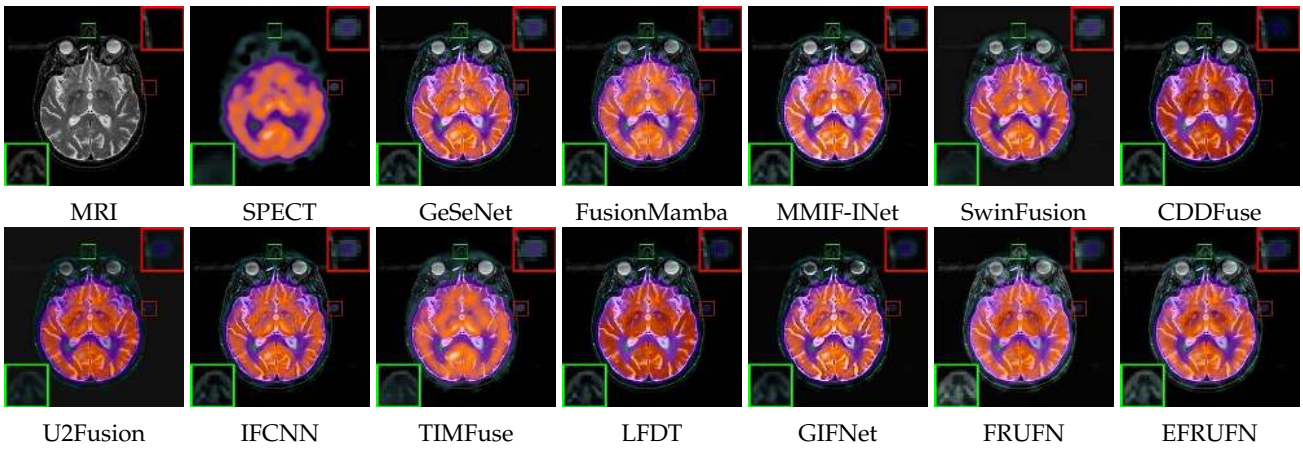

Fig. 10: Visual comparisons on the Harvard medical image dataset for the MRI-SPECT medical image fusion task.

although FusionMamba [100] achieves a better MS-SSIM score than our method, its parameter count is significantly larger. The outstanding quantitative results clearly highlight the excellent generalization capabilities of our method.

4.6 Comparison of Computational Complexity and Runtime

In this section, we compare the computational complexity of general image fusion methods. Table 5 summarizes the FLOPs, MACs, and runtime of each method. Note that U2Fusion [9] is implemented using an older version of TensorFlow, while all other methods are developed in PyTorch. Therefore, directly comparing U2Fusion [9] with other methods in terms of computational complexity and runtime would be methodologically inconsistent. For this reason, we exclude U2Fusion's complexity and runtime measurements from our comparative analysis. As shown in the table, our method exhibits higher computational complexity and longer runtime than IFCNN [45], but outperforms TCMOA [16] and remains competitive with other approaches. Additionally, EFRUFN shows slightly increased FLOPs, MACs, and runtime compared to FRUFN. This is attributable to its numerical algorithm, which requires the storage of intermediate results during inference, introducing a modest computational overhead.

4.7 Ablation Experiments: Sequential Gradient Transfer Framework

To validate the effectiveness of our proposed sequential transfer framework, we conducted a series of ablation experiments on two tasks, i.e., MEIF and IVF, with the proposed FRUFN. On the MEIF and IVF tasks, we compare the results of directly testing the models independently trained for the MFIF, MEIF, and IVF tasks with the results of the models trained using sequential gradient transfer. We present the quantitative results for the MEIF and IVF tasks in Tabs. 6 and 7, respectively. It can be observed from the tables that the sequential gradient transfer learning improves the performance for both the MEIF and IVF tasks.

4.8 Ablation Experiments: Network Structure

In this work, we propose the unfolding paradigm and dynamic learnable mask to increase the generalization ability of our models. To study the effects of these novel designs, we perform both numerical and visual ablation experiments. More specifically, we additionally train three variant

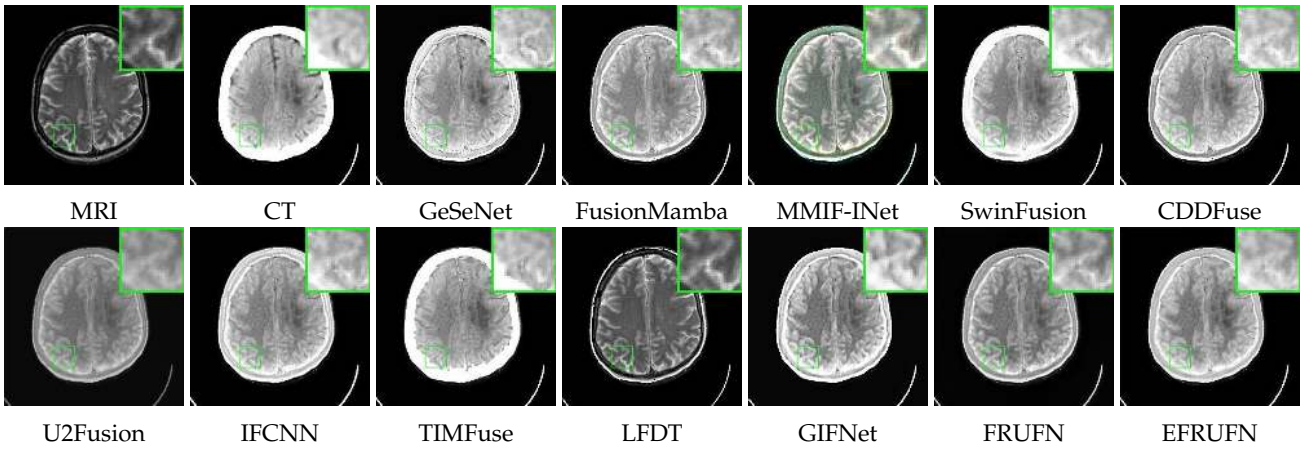

Fig. 11: Visual comparisons on the Harvard medical image dataset for the MRI-CT medical image fusion task.

TABLE 4: Average metrics of all compared approaches on 20 samples from the Harvard medical image dataset for the tasks of MRI-SPECT image fusion and MRI-CT image fusion. The results for the general methods are shown in light yellow. The best, second-best, and third-best results are highlighted in red, blue, and bold, respectively.

Method	Venue	MRI-SPECT					MRI-CT					#Param. (M)
		MS-SSIM↑	EN↑	MI↑	SCD↑	UNIQUE↑	MS-SSIM↑	EN↑	MI↑	SCD↑	UNIQUE↑	
GeSeNet [99]	TNNLS'23	0.9534	4.6066	9.2132	1.1758	**0.487**	0.9207	**4.7744**	**9.5488**	1.2974	-0.044	**0.241M**
FusionMamba [100]	VI'24	0.9704	4.2786	8.5572	1.4780	-0.11	0.9442	4.1012	8.2024	1.4686	0.069	51.647M
MMIF-INet [97]	IF'25	0.9426	4.3263	8.6525	1.8776	0.373	0.9135	4.3232	8.6464	1.3750	**0.152**	0.749M
U2Fusion [9]	TPAMI'20	0.8448	4.0814	8.1628	0.1108	-0.175	0.8625	4.4808	8.9616	0.7168	0.14	2.636M
IFCNN [45]	IF'20	0.9654	4.4966	8.9932	1.0646	0.47	**0.9410**	4.4958	8.9916	1.1755	0.007	0.083M
SwinFusion [1]	JAS'22	0.9385	**4.7227**	**9.4454**	0.8908	-0.161	0.9352	4.0171	8.0342	1.4001	0.032	0.973M
CDDFuse [4]	CVPR'23	0.9316	4.3568	8.7136	1.2204	0.470	0.9275	4.2861	8.5722	1.3207	0.018	1.188M
TIMFuse [52]	TPAMI'24	0.9105	4.5642	9.1285	**1.4802**	0.247	0.8818	4.2810	8.5620	1.2532	0.083	1.232M
LFDT [98]	IF'25	0.9316	4.5014	9.0028	1.0419	-0.067	0.7395	4.2624	8.5248	0.6816	-0.004	20.28M
GIFNet [93]	CVPR'25	0.8530	4.3458	8.6915	1.5984	0.514	0.9380	4.6068	9.2136	1.1512	-0.234	3.291M
FRUFN	-	0.9328	5.0450	10.0901	1.2293	0.521	0.9112	5.2384	10.4768	1.0985	0.24	0.864M
EFRUFN	-	**0.9583**	5.2384	10.4769	1.4471	0.402	0.9415	4.8855	9.7711	**1.3487**	0.303	0.236M

TABLE 5: Computational complexity and runtime comparison of general image fusion methods. FLOPs and MACs were computed using the `calflops` package on source images with a resolution of 224×448. Runtime is reported as the average time per test sample over the entire TNO dataset.

Method	SwinFusion	U2Fusion	IFCNN	TCMOA	CDDFuse	TIMFuse	LFDT	GIFNet	FRUFN	EFRUFN
#Flops. (G)	189.02	\	26.07	2120	358.07	102.73	27.5	122.1	130.22	148.71
#MACs. (G)	92.78	\	13.01	1050	178.92	51.15	13.49	60.08	64.36	73.61
#Times. (S)	1.369	\	0.019	0.556	0.302	0.073	0.278	0.192	0.333	0.356

models of the FRUFN and EFRUFN by either removing the unfolding paradigm or removing the dynamic learnable mask.

TABLE 6: Average metrics of all compared DL-based approaches on 27 samples from the SICE [66] dataset. The best results are highlighted in red.

Method	PSNR↑	MS-SSIM↑	MEF-SSIM↑
MFIF-to-MEIF	13.9429	0.7814	0.6664
IVF-to-MEIF	13.9429	0.7814	0.6664
MEIF-to-MEIF	19.1939	0.8418	0.7357
Ours	21.6278	0.8507	0.7280

Tab. 8 illustrates the numerical ablation experimental results. From the table, removing the proposed mechanisms leads to a performance drop on both the IVF and MIF tasks. However, the effects on the MFIF task are minimal. Additionally, we depict the visual ablation experimental results on the IVF task in Fig. 12. The figure shows that omitting the unfolding paradigm or resorting the simple parameter sharing reduces the model's representational capacity, leading to excessively blurred structures in the fused images. Furthermore, without a dynamically learnable mask, the

TABLE 7: Average metrics of all compared DL-based approaches on 26 samples from the TNO [67] dataset. The best results are highlighted in red.

Method	EN↑	MI↑	SCD↑	MS-SSIM↑
MFIF-to-IVF	6.4519	12.9038	1.6024	0.8972
MEIF-to-IVF	6.1661	12.3322	1.4028	0.9028
IVF-to-IVF	6.8161	13.6321	1.6072	0.8925
Ours	6.8625	13.7251	1.7573	0.9374

TABLE 8: Ablation study results for network structures on the MFIF, the MEIF, and the IVF tasks. The best results are highlighted in red.

Method	MFIF			MEIF			IVF (TNO)		
	MS-SSIM↑	EN↑	MI↑	MS-SSIM↑	MEF-SSIM↑	PSNR↑	MS-SSIM↑	EN↑	MI↑
FRUFN shared params	0.9940	7.4920	14.9839	0.8496	0.7275	21.3389	0.9112	6.8044	13.6089
FRUFN w/o unfolding	0.9938	7.4875	14.9749	0.8480	0.7258	22.1921	0.9127	6.7653	13.5307
EFRUFN w/o mask	0.9934	7.4834	14.9668	0.8395	0.7223	21.5043	0.9300	6.5709	13.1419
FRUFN	0.9938	7.4943	14.9887	0.8507	0.7280	21.6278	0.9374	6.8625	13.7251
EFRUFN	0.9937	7.4943	14.9887	0.8500	0.7305	21.7333	0.9489	6.8991	13.7983

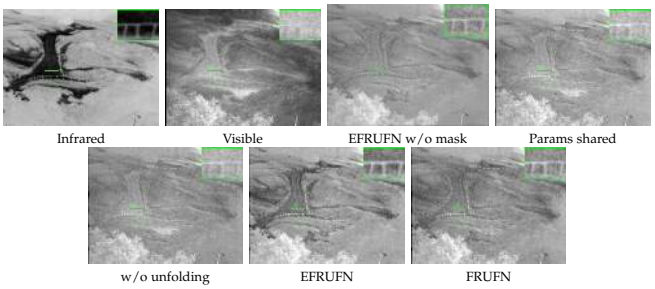

Fig. 12: Visual comparisons of different variants of the EFRUFN and FRUFN for the IVF task.

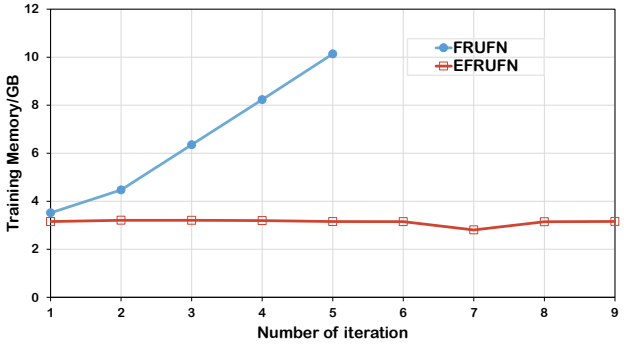

Fig. 13: A comparison of the training memory usage between the FRUFN and EFRUFN, benchmarked on a single RTX 4070Ti GPU with 12GB of memory. As the number of iterations increases, the GPU memory consumption for training the FRUFN rises linearly, whereas the GPU memory usage for training the EFRUFN remains constant.

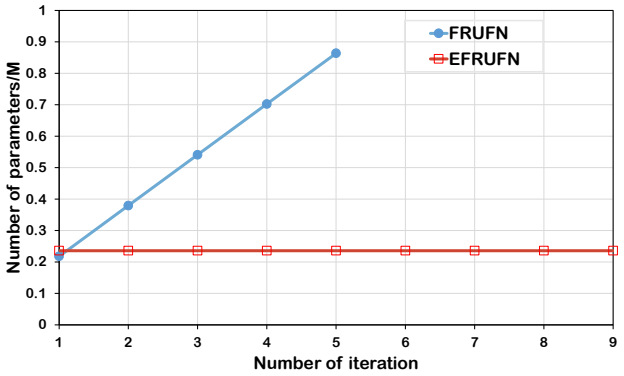

Fig. 14: A comparison of the number of parameters between the FRUFN and the EFRUFN. As the number of iterations increases, the parameter count for training the FRUFN grows linearly, whereas the parameter count for training the EFRUFN remains constant.

model fails to distinguish between the foreground and background, resulting in a noticeable decrease of image contrast.

4.9 Efficiency comparison between EFRUFN and FRUFN

In this section, we compare the efficiency between the EFRUFN and the FRUFN. We report the total training memory with the batch size being set to 4 in Fig. 13. It is evident that the FRUFN's memory consumption increases linearly with the number of iterations, whereas the EFRUFN's memory usage remains constant regardless of the iteration count. Due to GPU memory limitations, the FRUFN can only handle a maximum of 5 iterations, while the EFRUFN can support an unlimited number of iterations. Consequently, the EFRUFN is more memory-efficient than the FRUFN. Fig. 14 shows that the parameter count of the EFRUFN remains stable as the number of iterations increases, in contrast to the FRUFN, where the parameter count grows linearly. This indicates that the EFRUFN is more parameter-efficient than the FRUFN. Moreover, Fig. 15 shows that the performance of the EFRUFN and FRUFN is similar across different iteration counts. However, the FRUFN is constrained to 5 iterations due to GPU memory limits, while the EFRUFN can be iterated more times without such constraints, potentially leading to better performance. In summary, the EFRUFN requires fewer model parameters and less GPU memory during training while achieving performance comparable to the FRUFN.

5 Conclusion

In this work, we presented a comprehensive approach to general image fusion, incorporating innovations in both model training and network architecture design. For model training, we introduced a sequential gradient transfer framework that leverages cross-task structural complementary information, substantially improving the model's ability to preserve the structural integrity of source images. In terms of network design, we proposed a versatile image fusion network based on the deep unfolding of core fusion principles, streamlining the traditionally time-consuming and labor-intensive network design process. Extensive experiments demonstrated that our approach produces outcomes with superior structural detail preservation compared to existing methods. Furthermore, the proposed

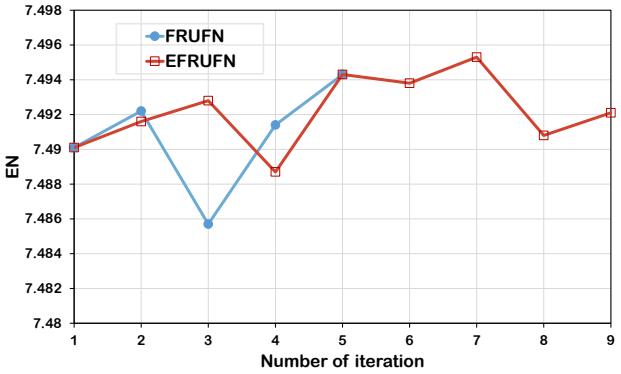

Fig. 15: A comparison of performance between the FRUFN and EFRUFN. We report the EN [83] quality metric for the fused images from both the EFRUFN and FRUFN at various iteration counts.

method exhibits strong generalization capabilities, performing well even when dealing with previously unseen image fusion tasks.

REFERENCES

[1] J.-Y. Ma, L. Tang, F. Fan, J. Huang, X.-G. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[2] H. Zhang and J. Ma, "IID-MEF: A multi-exposure fusion network based on intrinsic image decomposition," *Information Fusion*, vol. 95, pp. 326–340, 2023.

[3] H. Li, T.-Y. Xu, X.-J. Wu, J. Lu, and J. Kittler, "LRRNet: A novel representation learning guided fusion network for infrared and visible images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11 040–11 052, 2023.

[4] Z.-X. Zhao, H.-W. Bai, J.-S. Zhang, Y.-L. Zhang, S. Xu, Z. Lin, R. Timofte, and L. V. 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)*, 2023, pp. 5906–5916.

[5] J.-J. Adu, J.-H. Gan, Y. Wang, and J. Huang, "Image fusion based on non-subsampled contourlet transform for infrared and visible light image," *Infrared Physics and Technology*, vol. 61, pp. 94–100, 2013.

[6] J.-Y. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, pp. 100–109, 2016.

[7] X. Qiu, M. Li, L. Zhang, and X. Yuan, "Guided filter-based multi-focus image fusion through focus region detection," *Signal Processing: Image Communication*, vol. 72, pp. 35–46, 2019.

[8] S. S, P. Latha, and S. A. P, "Image fusion technique using dt-cwt," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 2013, pp. 160–164.

[9] H. Xu, J.-Y. Ma, J.-J. Jiang, X.-J. Guo, and H.-B. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[10] H. Xu, J. Yuan, and J. Ma, "Murf: mutually reinforcing multi-modal image registration and fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, 2023.

[11] J. Liu, S. Li, L. Tan, and R. Dian, "Denoiser learning for infrared and visible image fusion," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.

[12] J. Liu, S. Li, H. Liu, R. Dian, and X. Wei, "A lightweight pixel-level unified image fusion network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2023.

[13] L. Pang, X. Rui, L. Cui, H. Wang, D. Meng, and X. Cao, "Hir-diff: unsupervised hyperspectral image restoration via improved diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3005–3014.

[14] G. Yang, X. Cao, W. Xiao, M. Zhou, A. Liu, X. Chen, and D. Meng, "Panflownet: a flow-based deep network for pan-sharpening," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 857–16 867.

[15] J. Liu, S. Li, R. Dian, and Z. Song, "Focus relationship perception for unsupervised multi-focus image fusion," *IEEE Transactions on Multimedia*, vol. 26, pp. 6155–6165, 2024.

[16] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7099–7108.

[17] B. Liu, Y. Feng, P. Stone, and Q. Liu, "Famo: Fast adaptive multitask optimization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 57 226–57 243, 2024.

[18] A. Ben Hamza, Y. He, H. Krim, and A. Willsky, "A multiscale approach to pixel-level image fusion," *Integrated Computer-Aided Engineering*, vol. 12, no. 2, pp. 135–146, 2005.

[19] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense sift," *Information Fusion*, vol. 23, pp. 139–155, 2015.

[20] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Information Sciences*, vol. 508, pp. 64–78, 2020.

[21] R.-C. Hou, D.-M. Zhou, R.-C. Nie, D. Liu, L. Xiong, Y.-B. Guo, and C.-B. Yu, "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[22] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

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

[24] Z. Wang, X. Li, H. Duan, and X. Zhang, "A self-supervised residual feature learning model for multifocus image fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 4527–4542, 2022.

[25] J.-Y. Liu, R.-W. Dian, S.-T. Li, and H.-B. Liu, "SGFusion: A saliency guided deep-learning framework for pixel-level image fusion," *Information Fusion*, vol. 91, pp. 205–214, 2023.

[26] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, "Reconet: Recurrent correction network for fast and efficient multi-modality image fusion," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*. Springer, 2022, pp. 539–555.

[27] Z.-X. Zhao, S. Xu, C.-X. Zhang, J.-M. Liu, P.-F. Li, and J.-S. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 7 2020, pp. 970–976.

[28] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, and X. Fan, "Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," *International Journal of Computer Vision*, vol. 132, no. 5, pp. 1748–1775, 2024.

[29] M. Zhou, N. Zheng, X. He, D. Hong, and J. Chanussot, "Probing synergistic high-order interaction for multi-modal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024.

[30] A. Guo, R. Dian, and S. Li, "A deep framework for hyperspectral image fusion between different satellites," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 7939–7954, 2023.

[31] Y. Fu, T. Xu, X. Wu, and J. Kittler, "Ppt fusion: Pyramid patch transformerfor a case study in image fusion," *arXiv preprint arXiv:2107.13967*, 2021.

[32] W. Tang, F.-Z. He, and Y. Liu, "YDTR: Infrared and visible image fusion via y-shape dynamic transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 5413–5428, 2023.

[33] W. Tang, F. He, Y. Liu, and Y. Duan, "Matr: Multimodal medical image fusion via multiscale adaptive transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5134–5149, 2022.

[34] W. Tang and F. He, "Eat: Multi-exposure image fusion with adversarial learning and focal transformer," *IEEE Transactions on Multimedia*, 2025.

[35] W. Tang, F. He, and Y. Liu, "Itfuse: An interactive transformer for infrared and visible image fusion," *Pattern Recognition*, vol. 156, p. 110822, 2024.

[36] H. Zhang, J. Yuan, X. Tian, and J. Ma, "Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection

and dual markovian discriminators," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1134–1147, 2021.

[37] Y. Rao, D. Wu, M. Han, T. Wang, Y. Yang, T. Lei, C. Zhou, H. Bai, and L. Xing, "At-gan: A generative adversarial network with attention and transition for infrared and visible image fusion," *Information Fusion*, vol. 92, pp. 336–349, 2023.

[38] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, vol. 66, pp. 40–53, 2021.

[39] C. Zhao, P. Yang, F. Zhou, G. Yue, S. Wang, H. Wu, G. Chen, T. Wang, and B. Lei, "Mhw-gan: Multidiscriminator hierarchical wavelet generative adversarial network for multimodal image fusion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13 713–13 727, 2024.

[40] J.-Y. Ma, P.-W. Liang, W. Yu, C. Chen, X.-J. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85–98, 2020.

[41] Y. Gao, S. Ma, and J. Liu, "Dcdr-gan: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 549–561, 2022.

[42] J.-Y. Ma, W. Yu, P.-W. Liang, C. Li, and J.-J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[43] D. Rao, T. Xu, and X.-J. Wu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Transactions on Image Processing*, 2023.

[44] J. Zhang, L. Jiao, W. Ma, F. Liu, X. Liu, L. Li, P. Chen, and S. Yang, "Transformer based conditional gan for multimodal image fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 8988–9001, 2023.

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

[46] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3333–3348, 2020.

[47] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 797–12 804.

[48] H. Zhang and J. Ma, "Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion," *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2761–2785, 2021.

[49] M. Zhou, J. Huang, K. Yan, D. Hong, X. Jia, J. Chanussot, and C. Li, "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024.

[50] W. Wang, L.-J. Deng, R. Ran, and G. Vivone, "A general paradigm with detail-preserving conditional invertible network for image fusion," *International Journal of Computer Vision*, pp. 1–26, 2023.

[51] W. Wang, L.-J. Deng, and G. Vivone, "A general image fusion framework using multi-task semi-supervised learning," *Information Fusion*, vol. 108, p. 102414, 2024.

[52] R. Liu, Z. Liu, J. Liu, X. Fan, and Z. Luo, "A task-guided, implicitly-searched and metainitialized deep model for image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6594–6609, 2024.

[53] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 2020, pp. 491–507.

[54] Y. Li, S. Xie, X. Chen, P. Dollar, K. He, and R. Girshick, "Benchmarking detection transfer learning with vision transformers," *arXiv preprint arXiv:2111.11429*, 2021.

[55] J. Dong, Y. Cong, G. Sun, B. Zhong, and X. Xu, "What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020, pp. 4023–4032.

[56] K.-M. He, X.-Y. Zhang, S.-Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[57] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[58] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 12 009–12 019.

[59] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[60] S. Bai, V. Koltun, and J. Z. Kolter, "Multiscale deep equilibrium models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5238–5250, 2020.

[61] T. Wang, X. Zhang, and J. Sun, "Implicit feature pyramid network for object detection," *arXiv preprint arXiv:2012.13563*, 2020.

[62] R.-S. Liu, J.-Y. Liu, Z.-Y. Jiang, X. Fan, and Z.-X. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 1261–1274, 2020.

[63] Z. Zixiang, X. Shuang, Z. Jiangshe, L. Chengyang, Z. Chunxia, and L. Junmin, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1186–1196, 2022.

[64] X. Deng, J. Xu, F. Gao, X. Sun, and M. Xu, "DeepM^2m2cdl: Deep multi-scale multi-modal convolutional dictionary learning network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2770–2787, 2024.

[65] J. Zhang, Q. Liao, S. Liu, H. Ma, W. Yang, and J.-H. Xue, "Realmff: A large realistic multi-focus image dataset with ground truth," *Pattern Recognition Letters*, vol. 138, pp. 370–377, 2020.

[66] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[67] T. Alexander, "The TNO multiband image data collection," *Data In Brief*, vol. 15, pp. 249–251, 2017.

[68] Z. 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*, vol. 13, no. 4, pp. 600–612, 2004.

[69] D. Han, L. Li, X. Guo, and J. Ma, "Multi-exposure image fusion via deep perceptual enhancement," *Information Fusion*, vol. 79, pp. 248–262, 2022.

[70] P. Mu, Z. Du, J. Liu, and C. Bai, "Little strokes fell great oaks: Boosting the hierarchical features for multi-exposure image fusion," in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 2985–2993.

[71] A. Ben Hamza, Y. He, H. Krim, and A. Willsky, "A multiscale approach to pixel-level image fusion," *Integrated Computer-Aided Engineering*, vol. 12, no. 2, pp. 135–146, 2005.

[72] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representation (ICLR)*, 2021.

[73] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[74] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7242–7252.

[75] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[76] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Mathematics of Computation*, vol. 19, no. 92, pp. 577–593, 1965.

[77] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *Journal of the ACM (JACM)*, vol. 12, no. 4, pp. 547–560, 1965.

[78] S. W. Fung, H. Heaton, Q. Li, D. McKenzie, S. Osher, and W. Yin, "Jfb: Jacobian-free backpropagation for implicit networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, no. 6, 2022, pp. 6648–6656.

[79] Z. Hao, L. Zhuliang, S. Zhenfeng, X. Han, and M. Jiayi, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, vol. 66, pp. 40–53, 2021.

[80] H. Xu, J.-Y. Ma, Z.-L. Le, J.-J. Jiang, and X.-J. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 12 484–12 491.

[81] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference On Learning Representations (ICLR)*, vol. 1412, 2014, p. 80.

[82] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *International Conference on Learning Representation (ICLR)*, 2020.

[83] R. Wesley, van Aardt Jan, and A. Fethi, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, pp. 1–28, 2008.

[84] G.-H. Qu, D.-L. Zhang, and P.-F. Yan, "Information measure for performance of image fusion," *Electronics Letters*, vol. 38, pp. 1–7, 2002.

[85] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *Aeu-international Journal of Electronics and Communications*, vol. 69, no. 12, pp. 1890–1896, 2015.

[86] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers (ACSSC)*, vol. 2, 2003, pp. 1398–1402.

[87] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[88] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.

[89] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4819–4838, 2022.

[90] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, and M. Mukeshimana, "Sesf-fuse: An unsupervised deep model for multi-focus image fusion," *Neural Computing and Applications*, vol. 33, pp. 5793–5804, 2021.

[91] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[92] X. Hu, J. Jiang, X. Liu, and J. Ma, "ZMFF: Zero-shot multi-focus image fusion," *Information Fusion*, vol. 92, pp. 127–138, 2023.

[93] C. Cheng, T. Xu, Z. Feng, X. Wu, Z. Tang, H. Li, Z. Zhang, S. Atito, M. Awais, and J. 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*, 2025, pp. 28 102–28 112.

[94] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, pp. 72–84, 2015.

[95] L. Qu, S. Liu, M. Wang, and Z. Song, "Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, no. 2, 2022, pp. 2126–2134.

[96] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, and L. Van Gool, "Equivariant multi-modality image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 912–25 921.

[97] D. He, W. Li, G. Wang, Y. Huang, and S. Liu, "Mmif-inet: Multimodal medical image fusion by invertible network," *Information Fusion*, vol. 114, p. 102666, 2025.

[98] B. Yang, Z. Jiang, D. Pan, H. Yu, G. Gui, and W. Gui, "Lfdt-fusion: A latent feature-guided diffusion transformer model for general image fusion," *Information Fusion*, vol. 113, p. 102639, 2025.

[99] J. Li, J. Liu, S. Zhou, Q. Zhang, and N. K. Kasabov, "Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 16 248–16 261, 2024.

[100] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu, "Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba," *Visual Intelligence*, vol. 2, no. 37, 2024.

Wu Wang received the B.S. degree in electronic science and technology from the School of Electronic Science and Applied Physics, Hefei University of Technology, Anhui, China in 2015, and the Ph.D. degree in communication engineering from the School of Informatics, Xiamen University, Fujian, China in 2021. He is currently an Assistant Professor with the School of Computing and Artificial Intelligence, Southwest University of Finance and Economics. His research interests mainly focus on image fusion, including hyperspectral image fusion, pansharpening, infrared and visible image fusion, multi-focus image fusion, multi-exposure image fusion, and medical image fusion.

Liang-Jian Deng (Senior Member, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2016, respectively. He is currently a Research Fellow with the School of Mathematical Sciences, UESTC. From 2013 to 2014, he was a Joint-Training Ph.D. student with the Case Western Reserve University, Cleveland, OH, USA. In 2017, he was a Postdoc with Hong Kong Baptist University (HKBU). In addition, he also stayed at Isaac Newton Institute for Mathematical Sciences, Cambridge University and HKBU for short visits. His research interests include the use of partial differential equations (PDE), optimization modeling, and deep learning to address several tasks in image processing, and computer vision, e.g., resolution enhancement and restoration.

Qi Cao received the B.S. degrees in mathematics-physics fundamental science from the Yingcai Honors College, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2024. He is currently pursuing Ph.D. degree of machine learning and data science. His research interests include the use of machine learning and reinforcement learning techniques in cutting-edge topics, e.g., computer vision, large language models and image processing.

Gemine Vivone (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (summa cum laude) and the Ph.D. degree in information engineering from the University of Salerno, Fisciano, Italy, in 2008, 2011, and 2014, respectively. He is currently a Researcher with the Institute of Methodologies for Environmental Analysis (IMAA), National Research Council (CNR), Tito, Italy, and the National Biodiversity Future Center (NBFC), Palermo, Italy. His main research interests include statistical signal processing, detection of remotely sensed images, data fusion, and tracking algorithms.