Information Fusion 108 (2024) 102414

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

A general image fusion framework using multi-task semi-supervised learning

Wu Wang [a], Liang-Jian Deng [a,*], Gemine Vivone [b,c]

[a] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China
[b] Institute of Methodologies for Environmental Analysis, National Council of Research, CNR-IMAA, Tito Scalo, 85050, Italy
[c] NBFC, National Biodiversity Future Center, Palermo, 90133, Italy

ARTICLE INFO

Keywords:
Image fusion
Multi-task
Semi-supervised learning
Laplacian pyramid
Fusion rule
Remote sensing
Medical images

ABSTRACT

Existing image fusion methods primarily focus on solving single-task fusion problems, overlooking the potential information complementarity among multiple fusion tasks. Additionally, there has been no prior research in the field of image fusion that explores the mixed training of labeled and unlabeled data for different fusion tasks. To address these gaps, this paper introduces a novel multi-task semi-supervised learning approach to construct a general image fusion framework. This framework not only facilitates collaborative training for multiple fusion tasks, thereby achieving effective information complementarity among datasets from different fusion tasks, but also promotes the (unsupervised) learning of unlabeled data via the (supervised) learning of labeled data. Regarding the specific network module, we propose a so-called pseudo-siamese Laplacian pyramid transformer (PSLPT), which can effectively distinguish information at different frequencies in source images and discriminatively fuse features from distinct frequencies. More specifically, we take datasets of four typical image fusion tasks into the same PSLPT for weight updates, yielding the final general fusion model. Extensive experiments demonstrate that the obtained general fusion model exhibits promising outcomes for all four image fusion tasks, both visually and quantitatively. Moreover, comprehensive ablation and discussion experiments corroborate the effectiveness of the proposed method. The code is available at https://github.com/wwhappylife/A-general-image-fusion-framework-using-multi-task-semi-supervised-learning.

1. Introduction

Image fusion has extensive applications in various fields such as medicine, remote sensing [5–8], and industry [9]. Due to the limitations of optical imaging devices, a single source of sensor can only capture a portion of the scene information. For example, the images obtained from infrared imaging devices only contain saliency information in the infrared spectrum and lack the detailed information captured in the visible light spectrum. The objective of image fusion is to generate a synthesized image by integrating complementary information from multiple source images. Typical image fusion tasks include multi-focus image fusion (MFF), multi-exposure image fusion (MEF), infrared and visible light image fusion (IVF), multi-modal medical image fusion (MMF), remote sensing image pansharpening [10,11], and hyperspectral image fusion [12–14]. However, designing a general image fusion framework that incorporates both pansharpening and hyperspectral image fusion is challenging due to the large number of spectral bands in hyperspectral and multispectral images. In this work, we primarily focus on MFF, MEF, IVF, and MMF. Fig. 1 illustrates different image fusion tasks in a schematic diagram format.

Traditional image fusion methods rely on fixed image transformations, image decomposition techniques, and handcrafted fusion rules that lack representation ability. For example, the Laplacian pyramid approach decomposes the source image into components of different frequencies through multi-scale decomposition and performs fusion on these components individually. Due to its powerful representation capabilities, deep learning (DL) -based image fusion methods have become mainstream. Although deep learning-based image fusion methods have achieved good results, they still face two issues.

On one hand, many existing methods primarily focus on learning individual image tasks and overlook the potential complementary information among different image tasks. For example, an important goal of many image fusion tasks, such as IVF and MMF, is to preserve detailed information from the source images. However, data associated with the IVF and MMF tasks often lack sufficient detailed information. Consequently, models trained solely on IVF or MMF task-related data may lack the ability to preserve detailed information effectively. On the contrary, natural images in the MEF and MFF tasks contain rich detailed information. Leveraging such data, we can help the model to improve its ability to preserve detailed information. More specifically,

* Corresponding author.
E-mail addresses: wangwu@uestc.edu.cn (W. Wang), liangjian.deng@uestc.edu.cn (L.-J. Deng), gemine.vivone@imaa.cnr.it (G. Vivone).

https://doi.org/10.1016/j.inffus.2024.102414
Received 10 January 2024; Received in revised form 14 March 2024; Accepted 6 April 2024
Available online 8 April 2024

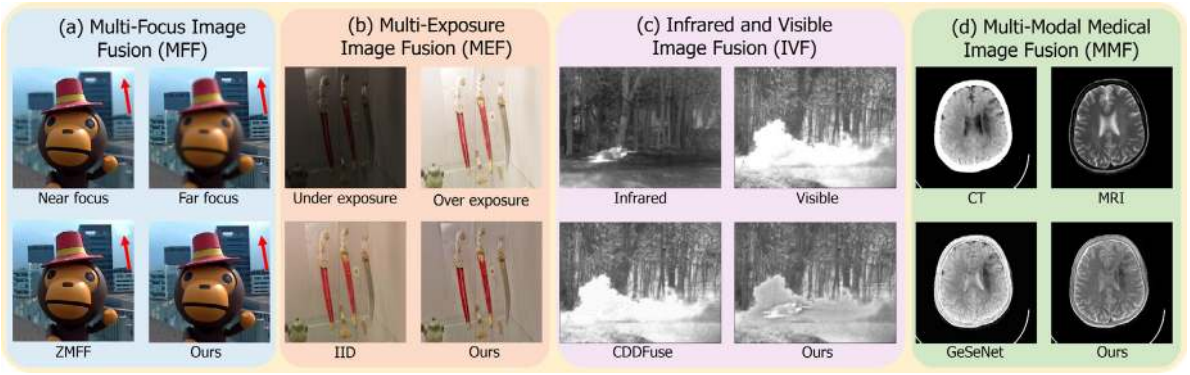

Fig. 1. Schematic illustration of various image fusion tasks, *i.e.,* MFF, MEF, IVF, and MMF. First row, source images. Second row, results of corresponding image fusion tasks obtained by state-of-the-art (SOTA) single-task training-based image fusion methods, *i.e.,* zero-shot multi-focus image fusion (ZMFF) [1], intrinsic image decomposition (IID) [2], correlation-driven dual-branch image fusion (CDDFuse) [3], general semantic-guided network (GeSeNet) [4], and the proposed multi-task semi-supervised learning method.

most image fusion methods rely on unsupervised learning, which poses challenges due to the lack of labeled data. These methods often rely on complex loss functions and model-tuning techniques to compensate for the absence of labels. We believe that supervised training can compensate for the limitations of unsupervised training. One piece of evidence supporting this is the fact that the image fusion framework based on convolutional neural network (IFCNN) [15], which is solely trained on data related to the MFF task, can generalize to multiple image fusion tasks.

Beyond the above-mentioned methods, many current approaches employ manually designed simple fusion rules, *e.g.,* the direct-average and choose-max rules, to fuse the features of source images, without distinguishing between the multi-frequency information present in the source images. However, in image fusion tasks, an important objective is to preserve the high-frequency details of the source images, *e.g.,* the visible details for the IVF tasks, and the foreground and background details for the MFF task. Therefore, some image fusion methods attempt to decompose the source image into high-frequency and low-frequency components to better preserve high-frequency information. For example, the work in [16] employs mean filtering to decompose the source image into high-frequency and low-frequency components, which are then fused separately. On the other hand, the work in [17] uses an autoencoder to learn a well-designed loss function for decomposing the source image into high-frequency and low-frequency features. The former relies on manually designed filters, which lack flexibility, while the latter is overly complex and unable to capture multiple frequency features. Additionally, many image fusion methods utilize simple fusion rules, and different methods employ different fusion rules. For example, IFCNN [15] uses the choose-max rule, the work in [16] employs the direct-average rule, and the work in [18] exploits the ℓ_1 norm as a fusion rule. Instead of relying on a single fusion rule, it may be beneficial to explore more advanced fusion strategies that can adapt to the specific characteristics of the images and the fusion task at hand. This flexibility can help improve the overall performance and quality of the image fusion results.

Inspired by the traditional Laplacian pyramid, we propose the PSLPT to incorporate different fusion rules for fusing the multi-frequency features of source images. Contrary to previous Laplacian pyramid networks that directly send all the source images into a single network to extract features with convolutional layers, PSLPT decomposes a pair of source images into features at multiple frequencies using Transformers. The benefit is that it allows for a more flexible fusion of local and global features at different frequencies from the source images. More in detail, PSLPT consists of two encoders, fusion modules, and a decoder. Each encoder and decoder form a Laplacian pyramid network, which automatically decomposes the source images into features at different frequencies by learning to reconstruct the source images. The fusion modules then adaptively fuse features at different frequencies from the source images. The fused features are subsequently passed through the decoder to generate the fusion result. This approach enables dynamic fusion of multi-frequency features with different learned fusion rules.

To extract complementary information from multiple image fusion tasks and leverage supervised learning to enhance unsupervised learning, we propose a multi-task semi-supervised training framework that includes two training stages. In the first stage, we pre-train the model using labeled data from the MFF and MEF tasks. Since these labeled data are natural images with rich details, the model can better preserve the details of the source images during this stage. Considering the significant exposure differences in the source images of the MEF task, it becomes challenging for the model to fit such data. On the other hand, the images related to the MFF task are all normally exposed. Therefore, multi-task learning can facilitate learning in the MEF task.

Although the model trained in the first stage performs well on the MFF and MEF tasks, it does not generalize well to the IVF task. This is because the IVF task requires the model to generate fused images with an intensity that closely matches the intensity of the source images. However, this goal conflicts with the objective of the MEF, which aims to fuse overexposed and underexposed images into normally exposed images, meaning that the MEF requires the fused image to have a different intensity compared to the source images. To address this issue, in the second training stage, we fix the parameters of the encoder and decoder and save the parameters of the fusion modules of the PSLPT trained in the first stage for handling the MEF and MFF tasks. We only train another set of fusion modules separately to handle the IVF task and the MMF task. Since we only train new fusion modules, a minimal number of iterations is sufficient for the model to fit the data in the IVF task. Therefore, we perform semi-supervised fine-tuning using unlabeled data from the IVF and MMF tasks, as well as the labeled data from the first stage. More specifically, we incorporate unlabeled data from both IVF and MMF tasks for unsupervised training, aiming to enhance the model's generalization ability for both the IVF and MMF tasks. At the same time, we still leverage the multi-task supervised training from the first stage to enhance unsupervised learning.

We train a single PSLPT model with two sets of fusion modules using the proposed multi-task semi-supervised training framework on nine datasets related to four image fusion tasks. Afterward, we directly test this single PSLPT model on these four image fusion tasks. Extensive experiments demonstrate that our PSLPT has achieved highly competitive performance. The results of the ablation experiments lead to the conclusions that (i) multi-task semi-supervised learning can improve the generalization ability across various image fusion tasks, (ii) the fused images generated using our proposed learned fusion rules exhibit more details, higher contrast, and lower noise in terms of visual quality compared to the images generated using simple fusion rules, validating the effectiveness of the learned fusion rule.

In summary, the contribution of our work is three-fold:

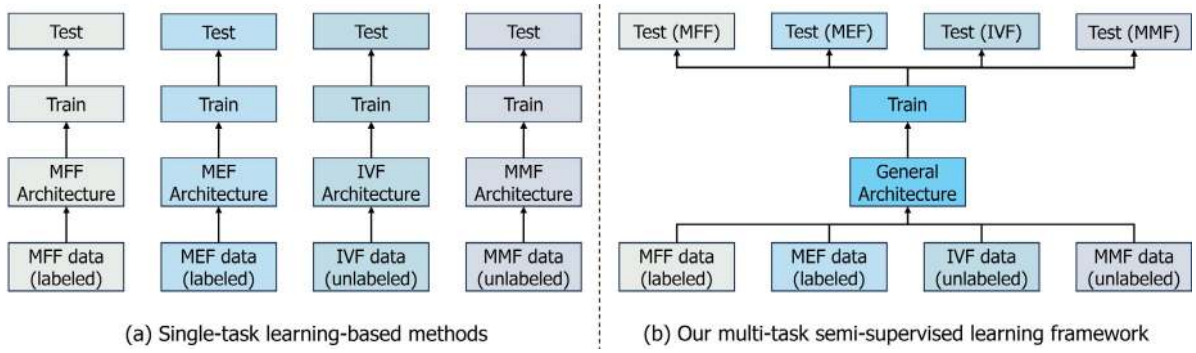

Fig. 2. A comparison of (a) the single-task learning-based methods and (b) our multi-task semi-supervised learning-based learning framework. Compared to methods based on single-task learning, our approach can extract complementary information from multi-task data and effectively utilize labeled data to guide the learning of unlabeled data.

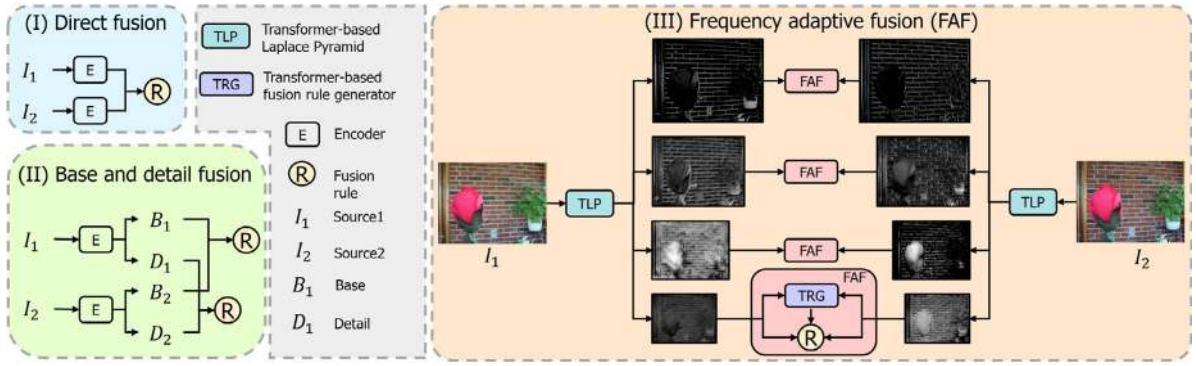

Fig. 3. Comparison of different feature fusion methods. Compared to direct fusion methods and methods that separately fuse base components and detail components, the proposed frequency-adaptive fusion method utilizes a transformer-based Laplacian pyramid to decompose the source image into different frequency bands. Subsequently, a transformer-based fusion rule generator is employed to generate fusion rules, which are then used to fuse the features of different frequency bands separately. This approach enables a more refined fusion.

- To address the general image fusion problem, we propose a new paradigm for general image fusion from both network architecture design and model training perspectives. Experimental results on four image fusion tasks show that the proposed method can achieve very competitive performance. The ablation experiments verify the novelty of the proposed method.
- In terms of model training, we propose a multi-task semi-supervised training framework to extract complementary information from multiple image fusion tasks and leverage supervised learning to enhance unsupervised learning, thereby improving the model's generalization ability.
- In terms of network architecture design, we propose the so-called PSLPT, which decomposes the source images into features at multiple frequencies and employs learned fusion rules to fuse them separately. This enables us to obtain more accurate fusion results.

The paper is organized as follows. Section 2 discusses related works to provide context for this study; Section 3 presents a detailed explanation of the proposed method; Section 4 showcases the effectiveness of the proposed method through a series of extensive experiments. Finally, Section 5 concludes the paper, summarizing the findings and contributions.

2. Related work

2.1. Laplacian pyramid for image fusion

The Laplacian pyramid is an image pyramid structure used to decompose an image into different frequency bands. It consists of a Gaussian pyramid and a difference pyramid, which are created through successive downsampling and upsampling operations. The traditional Laplacian pyramid uses hand-designed filters and downsampling to decompose images into components of different frequencies. It has been widely applied in many image fusion tasks, such as remote sensing image fusion [19–21], MFF [22,23], MMIF [24,25], IVF [26], and MEF [27,28]. Inspired by these traditional methods, some Laplacian pyramid networks [29–33] for image fusion have been proposed. For example, the work in [30] decomposes first the source image using the traditional Laplacian pyramid obtaining the multi-frequency components of the source image, and then uses a neural network to fuse them, finally getting the fused image through inverse transformation. Instead, the work in [31] exploits a learnable network to replace the hand-designed filters in the traditional Laplacian pyramid and builds an end-to-end Laplacian pyramid network.

Our method is also inspired by the traditional Laplacian pyramid, but it differs from the existing Laplacian pyramid networks for two points: (i) Unlike previous methods that solely rely on a single Laplacian network for image fusion, our PSLPT consists of two encoders and a shared decoder. This unique architecture enables PSLPT to not only learn image fusion but also learn to reconstruct the source images. This dual encoding scheme ensures that the network can capture and represent the diverse frequency characteristics present in the source images. (ii) We use Transformer to extract features, while these methods use convolutional networks to extract features. We believe that Transformer is better at extracting global features and is more suitable for image fusion tasks. Before our method, PPT [34] also employed a pyramid transformer structure. However, there are two main differences between our work and the one in [34]. First, our PSLPT is based on the Laplacian pyramid structure. It is used to decompose the source images into features of different frequencies. On the other hand, PPT [34] does not rely on the Laplacian pyramid structure. Instead, it

adopts a multi-scale structure to aggregate multi-scale features from the source images. Second, PPT [34] adopts a patch-based transformer approach. Specifically, the source image is first divided into different image patches, which are then processed by the transformer to extract features. Finally, these feature patches are used to generate the output image. In other words, this type of transformer can only handle image patches rather than the entire image. On the other hand, our PSLPT directly processes the source image as a whole without the need for partitioning it into image patches.

2.2. Learned fusion rules

Currently, most image fusion methods employ simple fusion rules, and only a very few works have explored the use of learned fusion rules. For example, the fusion rules in the work of [35] are divided into spatial attention-based fusion rules and channel attention-based fusion rules, with only the channel attention-based fusion rule being learnable. In contrast to [35], we learn the fusion rules based on spatial attention. Although the work in [36] learns the fusion rules based on spatial attention, their fusion rule is applied to the image domain, whereas our learned fusion rules are used for feature fusion. Additionally, we employ transformers to learn fusion rules (see Fig. 3), while the work in [35] and the work in [36] both use CNNs to learn fusion rules.

Compared to these two methods, we believe that our method has the following three advantages: First, employing the module for learning fusion rules based on spatial attention allows our method to effectively capture and combine relevant spatial information from different source images. Second, fusing in the feature space maximizes the utilization of deep learning's representation power, and this approach is also adopted by most deep learning-based image fusion methods. Moreover, by utilizing transformers to learn fusion rules, we leverage their strengths in capturing long-range dependencies and modeling complex relationships. This allows us to capture more comprehensive and context-aware fusion patterns, potentially leading to better fusion performance compared to methods that rely on CNNs.

2.3. Image fusion methods based on single-task training

Most image fusion methods are based on single-task training, which means that such methods can either only solve a single image fusion task (see Fig. 2(a)) or require the training of separate models with the same network architecture but different parameters for each image fusion task. These kinds of methods can generally be categorized into three types. The first type [37–40], directly learns the mapping for image fusion. These methods focus on constructing networks with strong representation capabilities and designing more effective loss functions. Due to the powerful modeling capabilities of transformers in capturing global features, there have been several image fusion studies based on them. The second type [3,41–44] first learns to reconstruct the source images and then learns the mapping for image fusion, enabling a better understanding of the information from different modalities in the source images. For example, the work in [41] utilizes a low-rank representation model to decompose a pair of source images into corresponding dictionaries and low-rank coefficients. The dictionaries and coefficients are fused separately to generate the fused image. The third type of method employs generative adversarial networks (GANs) to learn the mapping of the distribution for image fusion. These methods aim to improve the perceptual quality of the fused images. Representative methods in this category include [45–47], and [48].

In general, single-task training-based methods can only exploit data information that is relevant to the specific task at hand and cannot leverage cross-task data information.

2.4. Image fusion methods based on multi-task training

At present, there are indeed very few image fusion methods based on multi-task or semi-supervised learning. To our knowledge, the unified unsupervised image fusion network (U2Fusion) [49] is the only image fusion method based on multi-task learning. U2Fusion [49] employs a continual learning approach to utilize unlabeled data from multiple image fusion tasks to train a unified fusion model. However, U2Fusion [49] does not make use of valuable labeled data.

Compared to U2Fusion [49], our method has several advantages. First, we employ a multi-task semi-supervised training approach, which allows us to effectively utilize valuable labeled data to train the model. Second, unlike U2Fusion [49], which trains a single model for all tasks and may not generalize well to multiple image fusion tasks, we train universal encoders and decoders for all image fusion tasks. Additionally, we use two sets of fusion modules to handle MFF, MEF and IVF, MMF tasks separately. This flexible design balances both efficiency and enhances the model's generalization capability.

2.5. Motivation

In this work, our objective is to train a single model to handle multiple image fusion tasks. Existing image fusion methods primarily focus on single-task training using unsupervised learning, overlooking the potential complementary information among multiple image fusion tasks. Furthermore, they neglect the fact that the datasets related to image fusion tasks contain both labeled and unlabeled data. To address these issues, we propose a multi-task semi-supervised learning framework, see Fig. 2(b). In our framework, we extract complementary information from different datasets associated with image fusion tasks by simultaneously learning multiple image fusion tasks. We leverage semi-supervised learning to enhance the unsupervised learning of unlabeled data by utilizing the supervision provided by labeled data.

Additionally, most existing image fusion networks indiscriminately fuse different frequency features from source images using a simple fusion rule, see *e.g.,* Fig. 3. To overcome this limitation, we introduce PSLPT to decompose source images into features at different frequencies. We then exploit the learned fusion rules to achieve a more precise and flexible fusion of features at different frequencies.

3. Proposed method

This section introduces the pseudo-siamese Laplacian pyramid transformer (PSLPT), which decomposes the source image into multi-frequency features at multiple levels and employs learned fusion rules to fuse these features to get a fused outcome, see Section 3.1. To extract complementary information from different image fusion tasks and leverage the learning from labeled data to assist the learning from unlabeled data, we propose a multi-task semi-supervised training framework for better addressing the general image fusion problem, see Section 3.2.

3.1. PSLPT

To apply different fusion rules to the components of varying frequencies in the source images, inspired by the Laplacian pyramid, we propose the PSLPT as shown in Fig. 4.

Overall pipeline. Given a pair of source images, I_1 and I_2, PSLPT first extracts multi-frequency features from the source images using a pair of independent encoders. Then, the extracted features can be further fused or directly fed into the encoder to generate the fused outcome or reconstruct the source images. To obtain multi-scale features, inspired by the decomposition of the Laplacian pyramid, the encoder repeatedly extracts features from the source image and downsamples them, resulting in different levels or frequency bands of features. Conversely, the decoder extracts features from the encoder's features layer by layer

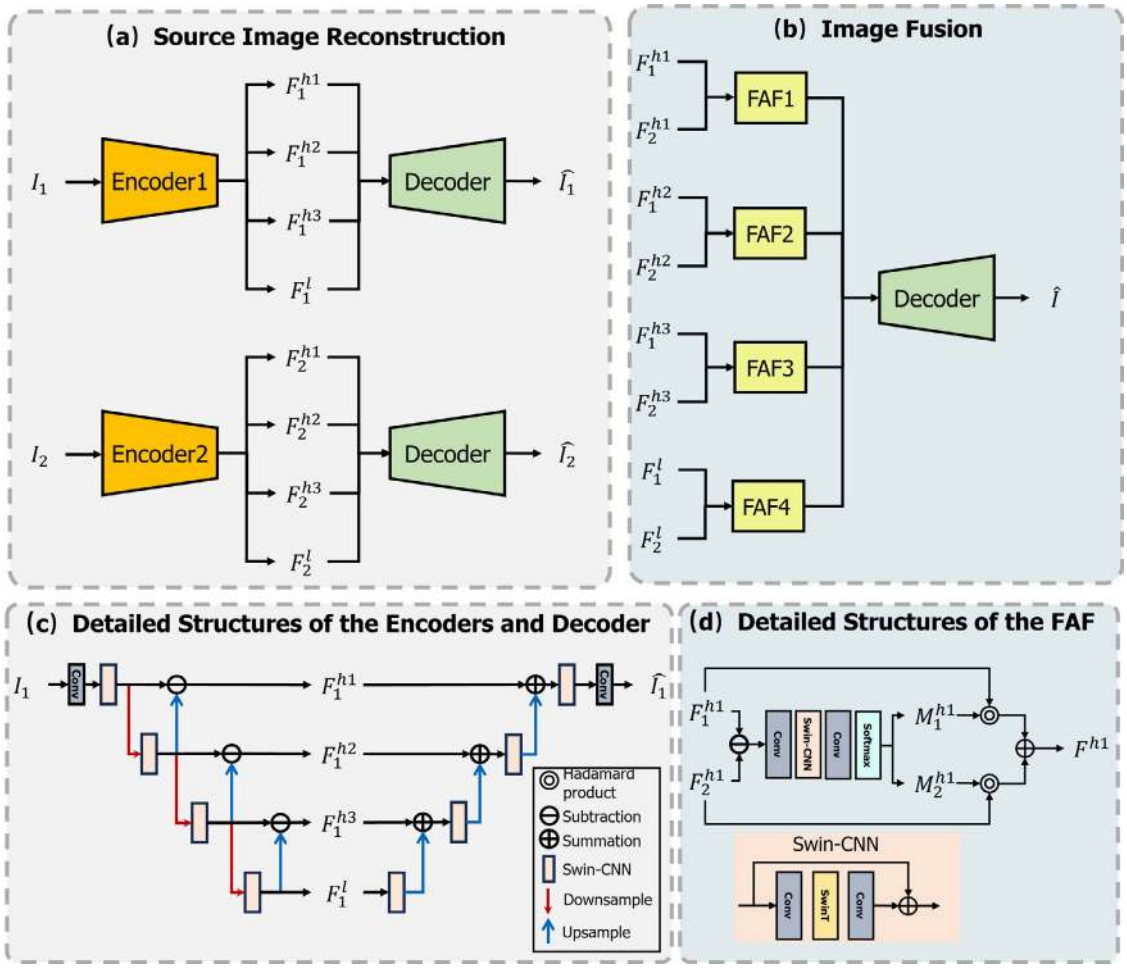

Fig. 4. The overall structure of the proposed pseudo-siamese Laplacian pyramid transformer (PSLPT). It is composed of two encoders that decompose the source images (*i.e.,* I_1 and I_2) into the low-frequency features (F_1^l and F_2^l) and the high-frequency components ($F_1^{h1}, F_1^{h2}, F_1^{h3}$ and $F_2^{h1}, F_2^{h2}, F_2^{h3}$). A shared decoder is responsible for image reconstruction and image fusion. (a) The flowchart of the reconstruction of the source images; (b) the flowchart of the generation of the fusion product; (c) the detailed structures of the encoders and the decoder; (d) the detailed structure of the frequency-adaptive fusion (FAF) modules.

and upsamples them to reconstruct the image. A pair of encoder and decoder forms the structure of the Laplacian pyramid, and a transformer is used to extract features at each level. Therefore, we refer to this structure as the **P**seudo **S**iamese **L**aplacian **P**yramid **T**ransformer (PSLPT). In the following section, we demonstrate the detailed skim of the reconstruction of source images and the generation of the fused outcome.

3.1.1. Multi-frequency decomposition by learning image reconstruction

The diagram for reconstructing the source image in PSLPT is shown in Fig. 4(c). Note that since the reconstruction process for both the source images is identical, we only show the reconstruction process for one of the source images. The encoders first extract shallow-level features from one source image, I_1, using a single-layer convolution. Then, the encoders gradually extract deep-level features with the hybrid module of Swin Transformer and CNN (Swin-CNN) from the shallow-level features and downsample the deep-level features, resulting in multi-scale features. We utilize max pooling to perform downsampling. The features at the lowest scale, *i.e.,* F_1^l, are considered the low-frequency features of the source image. Subsequently, the low-scale features are upsampled by a factor of 2, and their residuals are computed with the same-scale features to generate high-frequency features, *i.e.,* F_1^{h1}, F_1^{h2}, and F_1^{h3}. Besides, we exploit bilinear interpolation to perform upsampling. The decoder utilizes the multi-frequency features generated by the encoder to reconstruct the source images. The decoder and encoder exhibit completely symmetric structures. By optimizing a reconstruction loss function, the encoder is compelled to extract the desired multi-frequency features.

For the Swin-CNN modules, to simultaneously capture local and global features, we employ a hybrid structure combining CNN and transformers. The specific structure is illustrated in Fig. 4(d). It includes a Swin-Transformer (SwinT) [50] module and two convolutional layers to extract global and local features, respectively. Both the convolutional layers and the SwinT have 48 feature maps. The convolutional layers have a kernel size of 3×3. The depth of the SwinT module is 2, the window size referring to the window attention is 2, and the number of heads for the multi-head attention is 6. For more details about the SwinT module, please refer to [50]. Additionally, since transformers can be challenging to converge during training, we incorporate a shortcut to accelerate convergence.

3.1.2. Image fusion by frequency-adaptive fusion (FAF)

Previous image fusion methods usually use fixed or predefined simple fusion rules, *e.g.,* the direct-average and choose-max rules, which are not flexible enough. Instead, we propose the frequency adaptive fusion (FAF) modules to learn the fusion rules. Given the extracted multi-frequency features from two source images, *i.e.,* F_1^{h1}, F_1^{h2}, F_1^{h3} and F_2^{h1}, F_2^{h2}, F_2^{h3}, the FAF modules aim to generate the fused features that can be used by the decoder to obtain the target fused outcome. Taking the fusion of the highest frequency features as an example (see Fig. 4(d)), as the fusion pays more attention to the complementary information of the source images rather than the common features,

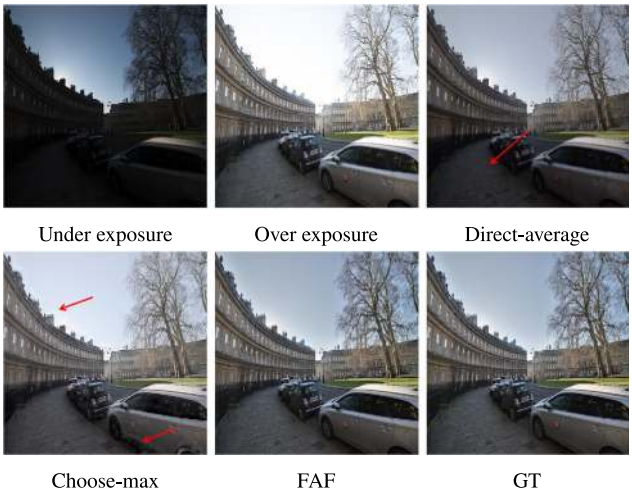

Fig. 5. Comparison of the visual quality of the fused images generated using the direct-average, the choose-max, and the proposed FAF rules.

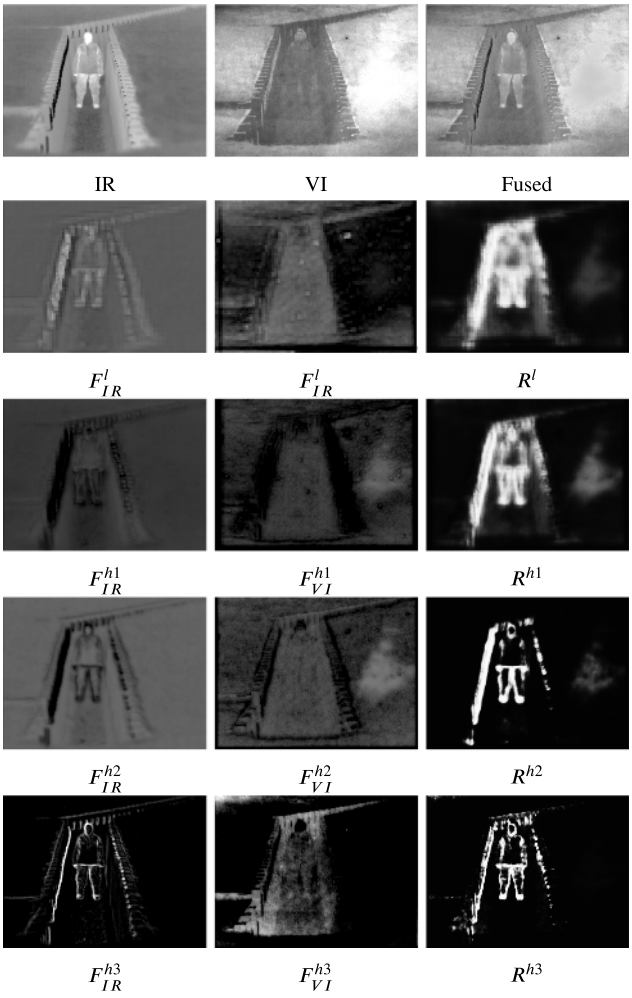

Fig. 6. Visualizing the multiple frequency features learned by the PSLPT and their corresponding fusion rules. The first row displays the source images and the fused result. From the second row to the fifth row, the first two columns show the features at different frequencies decomposed by the PSLPT, while the third column depicts the fusion rules.

the residuals from the two source image features are first calculated. Then the residual is sent to the Swin-CNN module for the extraction of semantic features. After that, a single layer of convolution is used to map the multi-channel semantic features into 2-channel features. Finally, the SoftMax operation is applied to normalize the features, resulting in two masks, *i.e.*, M_1^{h1} and M_1^{h2}. These masks are then element-wise multiplied by the features of the source images to obtain the fused features F^{h1}. Thus, we have:

$$F^{h1} = M_1^{h1} \odot F_1^{h1} + M_2^{h1} \odot F_2^{h1}, \tag{1}$$

where $\odot$ denotes the Hadamard product.

We present the fused images generated using the direct-average rule, the choose-max rule, and our FAF modules in Fig. 5. From the figure, it can be observed that the images generated using the direct-average rule fail to effectively recover the details in the dark areas, while the fused images generated using the choose-max rule exhibit noticeable distortions at the boundary between the sky and the buildings. In contrast, the images generated using the FAF modules demonstrate the highest quality. Furthermore, taking the IVF task as an example, we visualize the features at different frequencies learned by the PSLPT and their corresponding learned fusion rules in Fig. 6. From the figure, it can be seen that the low-frequency features decomposed by the PSLPT successfully extract salient objects from the infrared image (IR) and capture detailed structures from the visible image (VI). The high-frequency features decomposed by the PSLPT also effectively extract detailed information from the source images.

3.2. Multi-task semi-supervised learning framework

Our goal is to train a single PSLPT model using data from multiple image fusion tasks. By leveraging complementary information from these tasks, we aim to better address the image fusion problem. However, data from different image fusion tasks may consist of labeled or unlabeled data. More specifically, for multi-modal image fusion tasks, such as MMF and IVF, obtaining labeled data is challenging, while for tasks like MEF and MFF, labeled data can be obtained through reasonable simulation. To address these issues, we propose a multi-task semi-supervised learning framework.

Let I_1^{s1}, I_2^{s1}, and I^{s1} denote the source images and the corresponding label refers to the MFF task, I_1^{s2}, I_2^{s2}, and I^{s2} indicate the source images and the corresponding label refers to the MEF task, I_1^{u} and I_2^{u} denote the unlabeled source images referring to the IVF task and MMF task.[1] The flowchart of the framework is shown in Fig. 7. In the proposed framework, we extract complementary information from data related to different image fusion tasks through multi-task learning and we also employ semi-supervised learning to guide the learning of unlabeled data through labeled data. The proposed framework includes two training stages. In the first stage, we perform multi-task supervised pre-training using labeled data from MFF and MEF, which enhances the model's ability to extract complementary features from MFF and MEF tasks. Moreover, the availability of labeled data facilitates better convergence during this stage. In the second stage, we conduct semi-supervised fine-tuning using both labeled and unlabeled data. The unsupervised training improves the model's generalization ability in tasks like IVF, where labeled data may be scarce. Simultaneously, supervised training ensures that the model retains its capability to preserve detailed information, which helps the learning using unlabeled data.

[1] Since data related to the IVF and MMF tasks do not have labels, we merge the datasets from both the tasks into a single dataset and use a unified notation to represent these data.

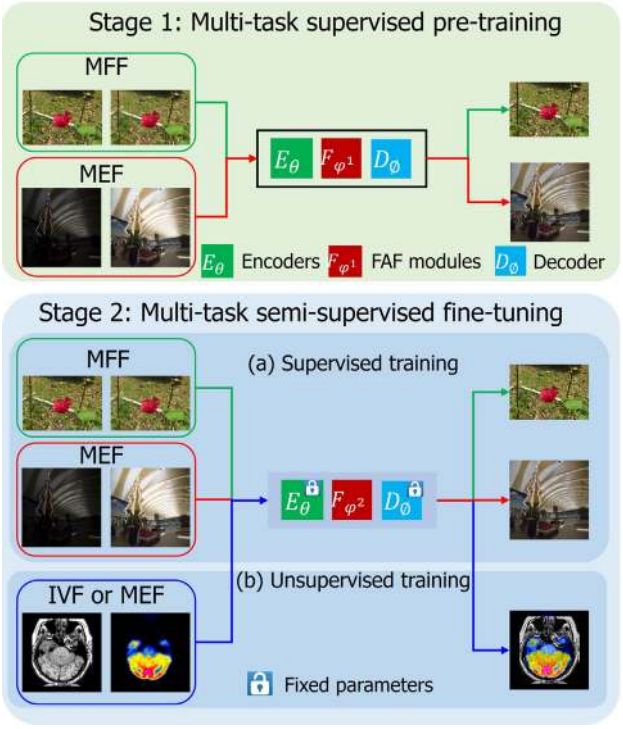

Fig. 7. The flowchart of the proposed two-stage multi-task semi-supervised learning framework. In the first stage of training, we employ multi-task supervised training to enable the model to learn to preserve rich detailed information from the source images. At this stage, we utilize labeled data from both the MFF and MEF tasks for training, and all parameters of PSLPT are updated. In the second stage of training, we utilize semi-supervised training to fine-tune the trained model in the first stage. At this stage, we not only utilize labeled data from MFF and MEF for supervised training but also leverage data from IVF and MMF for unsupervised training. More specifically, we fix the parameters of the encoders (E_θ) and decoder ($D_\emptyset$) of the PSLPT and save the parameters of the FAF modules (F_{φ^1}) of the proposed PSLPT to deal with the MFF and MEF tasks. Afterward, we update the parameters of the fusion modules to obtain another set of FAF modules (F_{φ^2}) to address the IVF and MMF tasks. This two-stage training strategy ensures that the model can generalize to the unlabeled data while still being able to preserve details.

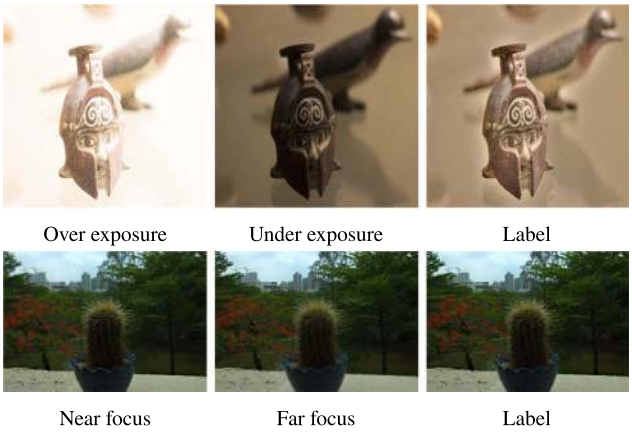

Fig. 8. Visual comparisons of images for the MEF (first row) and MFF (second row) tasks.

3.2.1. Stage 1: Multi-task supervised training

For supervised training, we found that the images referring to the MEF task (see Fig. 8) suffer from poor image quality due to overexposure and underexposure, making it difficult for the model to converge during training. On the other hand, the images in the MFF task (see Fig. 8) are normally exposed and have more details. Therefore, training simultaneously on both the MFF and MEF tasks can help the model to converge better for the MEF task.

Specifically, we sample data separately from the MFF and MEF datasets and perform forward propagation for each task independently. This allows us to leverage the higher-quality images from the MFF task to facilitate the model's convergence during training. Therefore, in the supervised training stage, the overall loss function is a weighted sum of the loss functions for the two tasks:

$$\mathcal{L}_{s1} = \mathcal{L}_{\text{MFF}} + \lambda \mathcal{L}_{\text{MEF}}, \tag{2}$$

where $\mathcal{L}_{\text{MFF}}$ and $\mathcal{L}_{\text{MEF}}$ are the loss functions for the MFF task and MEF tasks, respectively, and λ is a weight parameter that is employed to adjust the importance of the two tasks. Specifically, the loss function for the MFF and MEF tasks consists of an image fusion loss and an image reconstruction loss function. We adopt the structural similarity index measure (SSIM) [51] based loss function as the image fusion loss function and the ℓ_1 norm as the reconstruction loss function. Therefore, the loss function for the MFF task can be summarized as:

$$\mathcal{L}_{\text{MFF}} = 1 - \texttt{SSIM}(I^{s1}, \hat{I}^{s1}) + \lambda_1(\|\hat{I}_1^{s1} - I_1^{s1}\|_1 + \|\hat{I}_2^{s1} - I_2^{s1}\|_1), \tag{3}$$

where $\hat{I}^{s1}$ is the fused image, $\hat{I}_1^{s1}$ and $\hat{I}_2^{s1}$ are the reconstructed source images, λ_1 is a trade-off parameter, $\texttt{SSIM}(\cdot)$ is the SSIM function, and $\|\cdot\|_1$ is the ℓ_1 norm. The loss function for the MEF task is as follows:

$$\mathcal{L}_{\text{MEF}} = 1 - \texttt{SSIM}(I^{s2}, \hat{I}^{s2}) + \lambda_1(\|\hat{I}_1^{s2} - I_1^{s2}\|_1 + \|\hat{I}_2^{s2} - I_2^{s2}\|_1). \tag{4}$$

3.2.2. Stage 2: Semi-supervised fine-tuning

After the first stage of training, the model can effectively deal with the MFF and MEF tasks but may struggle to handle the IVF and MMF tasks. The reason is that the goal of the MEF task is to fuse over-exposed and under-exposed images into a properly exposed image, while the IVF and MMF tasks require to the fused result to retain the intensity of the source images. The objectives of these two types of tasks conflict with each other. Direct fine-tuning of the model using datasets referring to the IVF and MMF tasks can improve performance for the IVF and MMF tasks but may degrade the performance of the MEF task. Since the objectives of these two types of tasks conflict with each other, we propose using two sets of FAF modules to handle different types of image fusion tasks. Specifically, in the second stage, we fix the parameters of the encoders and decoder of the PSLPT, and we save the FAF modules trained in the first stage to handle the MEF and MFF tasks. Additionally, we train an extra set of FAF modules specifically for the IVF and MMF tasks. In this stage, to enhance the model's ability to preserve the intensity of the source images, we use unsupervised training to fit the IVF and MMF data. However, relying solely on unlabeled data from the IVF and MMF tasks for unsupervised training is not sufficient. This is because the related images in these two types of tasks lack rich detail information, which makes models trained only on this data unable to effectively preserve the fine details of the source images. To address this issue, we incorporate supervised learning from the first stage. We use weight coefficients to balance the weights of multiple tasks during this training phase.

The total loss function for training on stage 2 is a weighted sum of one supervised loss function and one unsupervised loss function:

$$\mathcal{L}_{s2} = \mathcal{L}_{super} + \beta \mathcal{L}_{unsuper}, \tag{5}$$

where $\mathcal{L}_{super}$ is the supervised loss function, $\mathcal{L}_{unsuper}$ is the unsupervised loss function, and β is a trade-off positive parameter. We directly adopt the loss function used in the first stage as the loss function for supervised training. The loss function of unsupervised learning also includes an image fusion loss function and an image reconstruction loss function. Thus, we have:

$$\mathcal{L}_{unsuper} = \mathcal{L}_{fuse} + \beta_1 \mathcal{L}_{recon}, \tag{6}$$

where $\mathcal{L}_{recon}$ stands for the image reconstruction loss, $\mathcal{L}_{fuse}$ denotes the (unsupervised) image fusion loss, which is the loss function in

Table 1
The configuration of the involved training datasets.

Dataset	Labeled			Unlabeled				
	RealMFF	MFI-WHU	SICE	RS	TNO	CT-MRI	PET-MRI	SPECT-MRI
Num	94	96	201	190	30	184	269	337

Table 2
Average metrics of all the compared DL-based approaches on 23 samples of the Lytro dataset. The results of the multi-task training-based methods are in orange color. The best, second, and third results are in red, blue, and bold, respectively.

	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	#Param. (M)
SESF [52]$^{\text{NCA-2019}}$	7.4983	14.9967	0.9876	0.0040	0.3724	0.8945	0.8001	1.1468	\
IFCNN [15]$^{\text{IF-2019}}$	7.4779	14.9558	**0.9903**	0.0128	0.3541	0.8875	0.6845	0.8849	0.083M
CNNFuse [53]$^{\text{IF-2017}}$	7.4956	14.9912	0.9887	0.0001	**0.3717**	0.8951	0.8015	1.1364	0.2M
ZMFF [1]$^{\text{IF-2023}}$	7.4927	14.9854	0.9853	0.0296	0.3508	0.8847	0.7403	0.8520	4.67M
SwinFusion [39]$^{\text{JAS-2022}}$	7.4749	14.9499	0.9879	**0.0027**	0.3608	0.8871	0.6672	0.8482	**0.973M**
U2Fusion [49]$^{\text{TPAMI-2020}}$	7.2835	14.5670	0.9765	0.0306	0.3449	0.8784	0.6364	0.7838	2.636M
Ours (only stage 1)	**7.4933**	**14.9866**	0.9917	0.0058	0.3709	0.8935	0.7442	0.9736	1.803M
Ours (2 stages)	7.4926	14.9851	0.9908	0.0027	0.3738	**0.8938**	**0.7435**	**0.9765**	1.803M

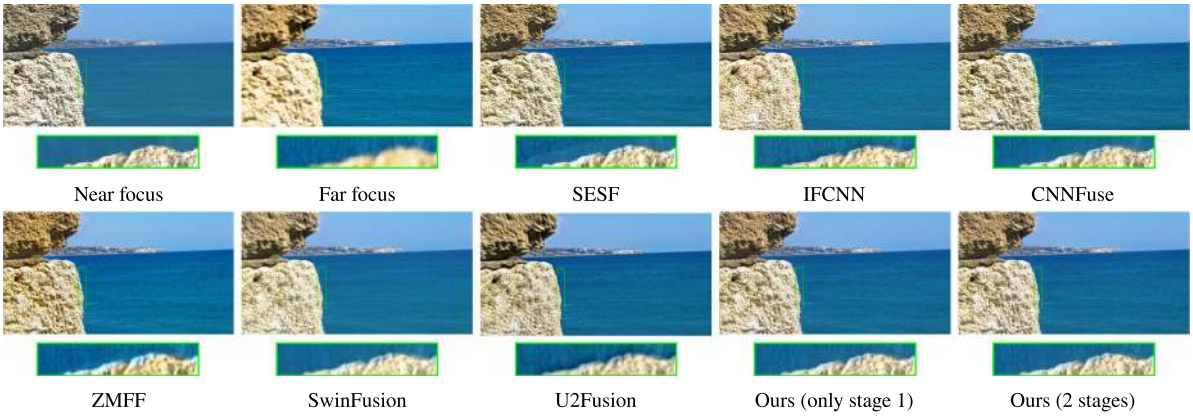

Fig. 9. Visual comparisons of all the compared approaches on the Lytro dataset. For convenience, we only show the results of the model trained in the first stage.

SwinFusion [39], and β_1 is a weight coefficient. This loss function mainly consists of three components: intensity loss function, texture loss function, and structural loss function (for more details, please refer to [39]):

$$\mathcal{L}_{fuse} = \mathcal{L}_{int}(I_1^u, I_2^u, \hat{I}^u) + \beta_2 \mathcal{L}_{text}(I_1^u, I_2^u, \hat{I}^u) + \beta_3 \mathcal{L}_{ssim}(I_1^u, I_2^u, \hat{I}^u), \tag{7}$$

where I_1^u and I_2^u are the unlabeled source images, $\hat{I}^u$ denotes the corresponding fused outcome, and β_2 is a weight coefficient. $\mathcal{L}_{int}$ denotes the intensity loss, which can be calculated as:

$$\mathcal{L}_{int}(I_1^u, I_2^u, \hat{I}^u) = \|\hat{I}^u - \mathtt{Max}(I_1^u, I_2^u)\|_1, \tag{8}$$

where $\mathtt{Max}(\cdot)$ refers to the element-wise maximum operator. Instead, $\mathcal{L}_{text}$ represents the texture loss in the gradient domain, which can be calculated as:

$$\mathcal{L}_{text}(I_1^u, I_2^u, \hat{I}^u) = \||\nabla \hat{I}^u| - \mathtt{Max}(|\nabla I_1^u|, |\nabla I_2^u|)\|_1, \tag{9}$$

where $|\cdot|$ is the absolute value function and ∇ indicates the gradient operator. $\mathcal{L}_{ssim}$ is the structural loss function which can be calculated as in Eq. (3). Meanwhile, we also minimize the ℓ_1 norm reconstruction losses so that the model can decompose the unlabeled source images into multi-frequency features (see Section 3.1):

$$\mathcal{L}_{recon} = \|I_1^u - \hat{I}_1^u\|_1 + \|I_2^u - \hat{I}_2^u\|_1, \tag{10}$$

where $\hat{I}_1^u$ and $\hat{I}_2^u$ are the reconstructed unlabeled source images.

4. Experiments

In this section, we first present the implementation details for experiments, then conduct the experiments for the MFF, MEF, IVF, and MMF tasks to verify the superiority of our method, both visually and quantitatively. Moreover, we show extensive discussions and ablation studies to corroborate the effectiveness of the proposed approach.

4.1. Implementation details

In terms of the training dataset, we chose nine datasets corresponding to four common image fusion tasks for training, as shown in Table 1. More specifically, the RealMFF [54] and MFI-WHU [55] datasets refer to the MFF task, and the SICE [56] dataset refers to the MEF task, which are used for supervised training. On the other hand, the RoadScene (RS) [57] and TNO [58] datasets refer to the IVF task, and the CT-MRI, PET-MRI, and SPECT-MRI images from the Harvard medical dataset[2] refer to the MMF task, which are used for unsupervised training. The MFI-WHU [55] dataset has 190 simulated samples; we randomly chose 96 samples for training. Besides, the RealMFF [54] has 710 real samples. Since the simulated samples from the MFI-WHU [55] dataset deliver very different distributions for the real samples from the RealMFF [54] dataset, we randomly selected 94 samples from the RealMFF [54] data for training. As a result, the sample sizes for the WHU-MFI [55] and RealMFF [54] datasets are comparable, which ensures that the model trained using these samples does not excessively overfit to a given dataset. Moreover, the SICE [56] dataset contains 229 samples; we removed 7 samples and randomly selected 195 samples for training. Since these samples cover images with multiple exposure ranges, we only select one image from each category (underexposed and overexposed) for training or testing.

[2] https://www.med.harvard.edu/AANLIB/home.htm

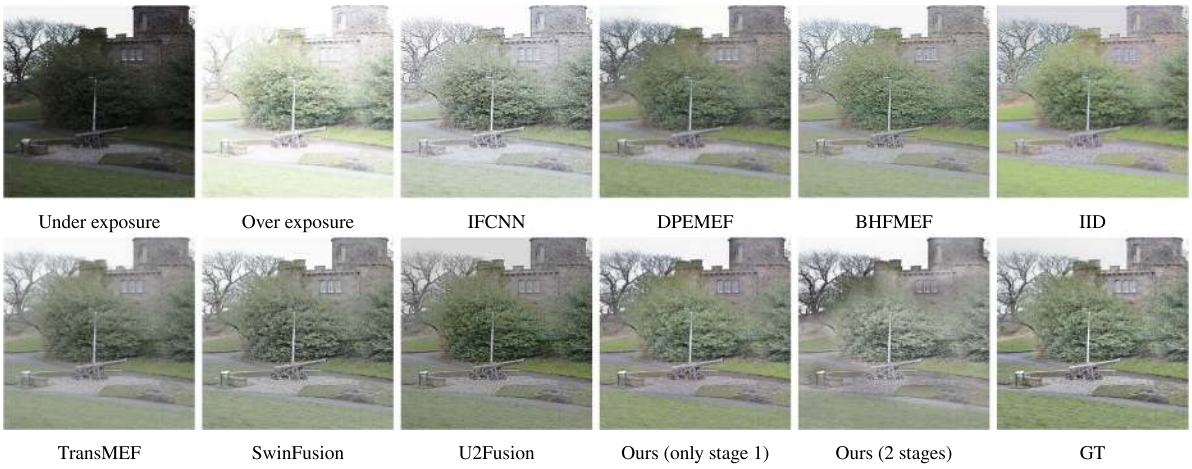

Fig. 10. Visual comparisons of all the compared approaches on the SICE dataset.

Specifically, for underexposed images, we choose the third image as the source image, and for overexposed images, we select the last image as the source image. In addition, the RS [57] dataset includes 221 samples. We randomly chose 201 samples for training. Due to the presence of both grayscale and RGB images in these datasets, for ease of training, we convert the RGB images into the YCbCr color space and only utilize the Y component for training.

The hyperparameters in the loss functions are set to $\lambda = 0.1$, $\lambda_1 = 0.5$, $\beta = 0.1$, $\beta_1 = 1.25$, $\beta_2 = 1$, and $\beta_3 = 0.5$, respectively. We chose Adam [59] as the optimizer. Due to the high resolution of the original images, during training, we randomly crop small image patches from the original images. The sizes of the image patches used for supervised training and unsupervised training are set to 128×128 and 64×64, respectively. As mentioned earlier, our model adopts a two-stage training approach. In the first training stage, the model is trained for a total of 320 epochs with an initial learning rate of 1×10^{-4}. The learning rate is halved at the 200th epoch. In the second training stage, the model is trained for a total of 20 epochs with an initial learning rate of 3×10^{-5}. Specifically, the training time for the first stage is approximately 5 h, while the training time for the second stage is approximately 30 min. We adopt ℓ_1 norm-based gradient clipping to accelerate the convergence, with the hyperparameter being set to 0.0001.

Regarding the hardware and software platform, we use an RTX 4070ti with 12 GB of memory for training, and the code is written in Pytorch 1.13.

4.2. MFF experiments

For the MFF experiments, the Lytro [60] dataset is used. The original Lytro [60] dataset has 38 image pairs. We remove the image pairs with low visual quality and use the left 23 image pairs for testing. We compare our methods with single-task training-based image fusion methods, which include IFCNN [15], CNNFuse[3] [53], SESF[4] [52], ZMFF [1], and SwinFusion [39], and a multi-task training-based image fusion method, *i.e.*, U2Fusion [49]. We select 8 popular metrics for evaluation, including EN [61], MI [62], MS-SSIM [51], N_{abf}, Q_{abf}, FMI [63], Q_{cb} [64], and NMI [65].

4.2.1. Quantitative results

Table 2 reports the quantitative comparison of experimental results. For the Lytro [60] dataset, our methods outperform U2Fusion [49] considering most metrics. Moreover, our method obtains the best Q_{abf} [66] and MS-SSIM [51]. Despite SESF [52] and CNNFuse [53] achieving very competitive performance, they rely on complex pre-process or post-process. Among the end-to-end methods, our method achieves the best performance. In general, the difference in metrics among various methods on the Lytro [60] dataset is very small. Furthermore, the fine-tuned model performs comparably to the original model. As for the parameter count, our model has a larger number of parameters compared to most competitors. However, in the future, we plan to develop models with smaller parameter counts.

4.2.2. Visual results

Fig. 9 shows the fused images on the Lytro [60] dataset. From the rectangles in the images, it can be observed that, apart from CNNFuse [53], SwinFusion [39], and our method, the fused images generated by the other compared methods exhibit noticeable blurring and distortion along the far-focus and near-focus boundaries. In comparison to our method, the fused images generated by CNNFuse [53] have fewer foreground details, while those produced by SwinFusion [39] suffer from overexposure in the foreground, resulting in a lack of fine details.

4.3. MEF experiments

We utilize the SICE [56] dataset for the MEF experiments. Because of the high resolution of the original images, we resized them to 512×512 before conducting the testing. This resizing helps in managing computational resources and ensuring consistency during the testing process. We compare our methods with single-task training-based methods, which include IFCNN [15], BHFMEF[5] [67], DPEMEF[6] [68], IID[7] [2], TransMEF[8] [69], SwinFusion [39], and a multi-task training-based method, *i.e.*, U2Fusion [49]. Since the labels are available, we chose PSNR, MS-SSIM [51], and MEF-SSIM [70] as quality metrics.

The quantitative experimental results are presented in Table 3. It can be observed that our method (only stage 1) significantly outperforms the comparison methods in all three evaluation metrics. Furthermore, it can be seen that the performance of our model with 2 training

[3] https://github.com/xingchenzhang/MFIF
[4] https://github.com/Keep-Passion/SESF-Fuse
[5] https://github.com/ZhiyingDu/BHFMEF
[6] https://github.com/dongdong4fei/DPE-MEF
[7] https://github.com/HaoZhang1018/IID-MEF
[8] https://github.com/miccaiif/TransMEF

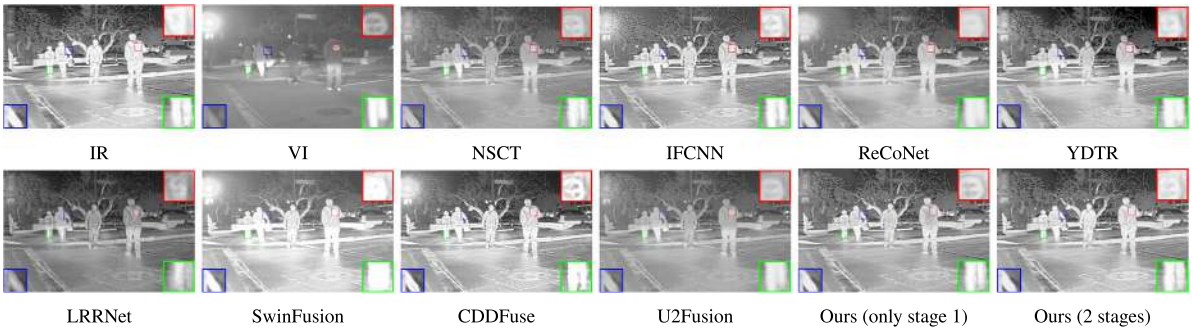

Fig. 11. Visual comparisons of all the compared approaches on the RS dataset.

Table 3
Average metrics of all compared DL-based approaches on 27 samples of the SICE [56] dataset. The results of the multi-task training-based methods are in orange color. The best, second, and third results are in red, blue, and bold, respectively.

Method	PSNR↑	MS-SSIM↑	MEF-SSIM↑	Param. (M)
IFCNN [15]	16.5957	**0.8310**	**0.7112**	0.083M
DPEMEF [68]	19.1137	0.8178	0.7008	13.603M
BHFMEF [67]	19.9319	0.8154	0.7022	0.03M
IID [2]	19.6203	0.8149	0.6966	**0.36M**
Transmef [69]	18.9058	0.8127	0.6584	19.052
SwinFusion [39]	19.5695	0.8393	0.7218	0.973M
U2Fusion [49]	17.2494	0.8267	0.6996	2.636M
Ours (only stage 1)	22.7647	0.8495	0.7375	1.803M
Ours (2 stages)	**19.7843**	0.8223	0.7104	1.803M

stages is worse than our model trained with only stage 1 in the MEF task, which validates the contrasting objectives of the MEF and IVF tasks.

In Fig. 10, we show the fused images generated by different methods. Since our method only utilizes the Y component in the YCbCr color space for training and employs a simple weighted-average rule to fuse the Cr and Cb components, there is a possibility of color distortion. To alleviate this issue, we adopt the strategy mentioned in [67] to post-process the fused images, enhancing their color. From Fig. 10, it can be observed that the generated images by our method (only stage 1) are closest to the ground truth in terms of brightness and color. However, our model exhibits noticeable overexposure in the fused images generated after the two-stage training. This is because the second training stage forces the model to preserve the intensity of the source images, resulting in overexposure in certain regions of the generated images.

4.4. IVF experiments

We use the RS [57] and TNO [58] datasets for the IVF experiments. We compare our method with a traditional method, some single-task training-based methods, and a multi-task training-based image fusion method. The traditional method is NSCT[9] [71]. The single-task training-based image fusion methods are IFCNN[10] [15], RecoNet[11] [72], YDTR[12] [37], LRRNet[13] [41], SwinFusion[14] [39], and CDDFuse[15] [3]. The multi-task training-based image fusion method is U2Fusion[16] [49].

[9] https://github.com/xingchenzhang/MFIF
[10] https://github.com/uzeful/IFCNN
[11] https://github.com/dlut-dimt/ReCoNet
[12] https://github.com/tthinking/YDTR
[13] https://github.com/hli1221/imagefusion-LRRNet
[14] https://github.com/Linfeng-Tang/SwinFusion
[15] https://github.com/Zhaozixiang1228/MMIF-CDDFuse
[16] https://github.com/hanna-xu/U2Fusion

4.4.1. Quantitative results

The results on the RS dataset are reported in Table 4. Due to the adoption of a two-stage training approach, we present the experimental results of both stages. From the table, our method demonstrates highly competitive performance. Specifically, our method outperforms the U2Fusion [49] method for most of metrics, including MS-SSIM [51], FMI [63], Q_{cb} [64], and NMI [65]. When compared to state-of-the-art (SOTA) methods like YDTR [37] and CDDFuse [3], each method has its strengths and weaknesses. Additionally, experimental results indicate that fine-tuning significantly enhances the performance of our model. Regarding the TNO dataset, the experimental results are shown in Table 5. Our method demonstrates good performance in metrics such as MS-SSIM [51], FMI [63], Q_{cb} [64], and NMI [65]. Furthermore, fine-tuning continues to improve the performance of our model. These findings highlight the competitive performance of our method on both the RS and TNO datasets.

4.4.2. Visual results

The visual comparison is shown in Figs. 11 and 12. As can be seen from the green rectangle boxes in Fig. 12, U2Fusion [49], RecoNet [72], and LRRNet [41] lose the details in the IR image. The fused image generated by YDTR [37] exhibits noticeable grid-like structural distortion. From the red rectangle in the image, it can be seen that RecoNet [72], SwinFusion [39], and CDDFuse [3] lose background details in the IR image. From the blue rectangle, it can be observed that our method well preserves the details in the VI image (see Table 7).

From the blue rectangle in Fig. 11, it can be observed that the fused image generated by CDDFuse [3] exhibits noticeable black dot-like distortions. From the red rectangle, it can be seen that the images generated by ReCoNet [72] and LRRNet [41] lose textual details in the VI image, while the textual structures in the images generated by U2Fusion [49], YDTR [37], and LRRNet [41] are not clear enough. From the green rectangle, it can be observed that the images generated by ReCoNet [72], YDTR [37], SwinFusion [39], and LRRNet [41] fail to preserve the structural information of salient objects in the IR image. The fused images generated by our method are highly competitive in terms of visual quality. In conclusion, experimental results demonstrate that the images generated by our model contain rich details. This indicates that training the model using labeled natural images from MFF and MEF can enhance the model's ability to preserve fine details in other image fusion tasks.

4.5. MMF experiments

In this section, we report the experimental results of the multi-model image fusion tasks. We selected two typical and widely studied MMF tasks, *i.e.,* the MRI-PET image fusion task and the MRI-CT image fusion task. The images used for testing are from the Harvard Medical image dataset. Specifically, we randomly chose 30 pairs of images for the MRI-PET image fusion task and 20 pairs of images for the

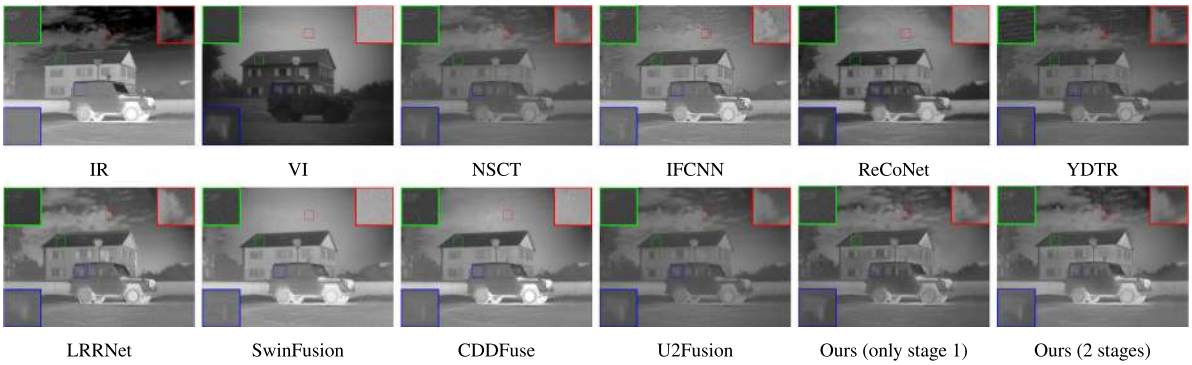

Fig. 12. Visual comparisons of all the compared approaches on the TNO dataset.

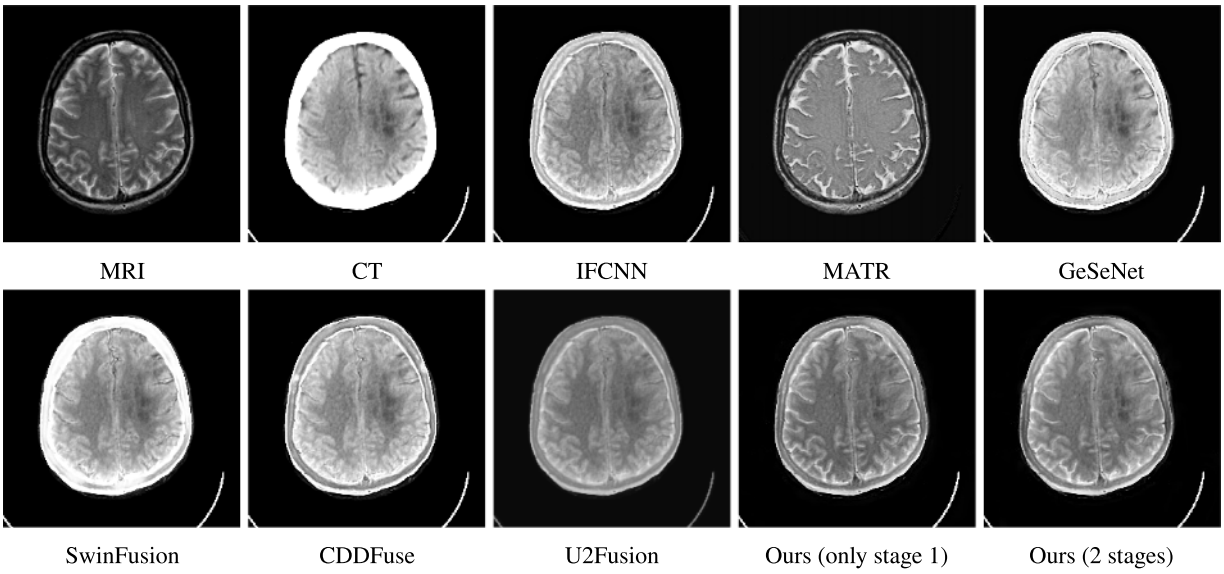

Fig. 13. Visual comparisons on the Harvard medical image dataset for the MRI-CT medical image fusion task.

Table 4
Average metrics of all the approaches on 20 samples of the RS dataset. The results of the multi-task training-based methods are in orange color. The best, second, and third results are in red, blue, and bold, respectively.

	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	#Param. (M)
NSCT [71]$^{\text{Infr.-2013}}$	7.2273	14.4546	0.8839	0.0025	0.1526	0.8601	0.4847	0.4121	\
IFCNN [15]$^{\text{IF-2019}}$	7.3933	14.7867	0.9154	0.0231	0.1572	**0.8637**	0.4873	0.4311	**0.083M**
ReCoNet [72]$^{\text{ECCV-2022}}$	7.3858	14.7716	0.8642	0.0103	0.1879	0.8557	0.5326	0.4405	0.007M
YDTR [37]$^{\text{TMM-2022}}$	7.7558	15.5116	0.9392	0.0254	0.1365	0.8533	0.5370	0.4078	0.2M
LRRNet [41]$^{\text{TPAMI-2023}}$	7.2669	14.5338	0.7791	0.0277	0.1186	0.8227	0.4885	0.3592	0.049M
SwinFusion [39]$^{\text{JAS-2022}}$	7.2072	14.4144	0.8469	0.0161	0.2107	0.8577	0.5189	**0.4708**	0.973M
CDDFuse [3]$^{\text{CVPR-2023}}$	7.6017	15.2033	0.8977	0.0533	**0.1752**	0.8527	0.4970	0.4176	1.188M
U2Fusion [49]$^{\text{TPAMI-2020}}$	7.2037	14.4074	0.8768	0.0003	0.1697	0.8608	**0.5588**	0.4283	2.636M
Ours (only stage 1)	7.4858	14.9717	**0.9355**	**0.0044**	0.1297	0.8779	0.5786	0.4926	1.533M
Ours (2 stages)	**7.5670**	**15.1339**	0.9474	0.0064	0.1415	0.8844	0.5997	0.4844	1.533M

MRI-CT image fusion task. We compare our methods with the single-task training-based methods, which include IFCNN [15], MATR[17] [38], GeSeNet[18] [4], SwinFusion [39] and CDDFuse [3], and a multi-task training-based method, *i.e.*, U2Fusion [49].

4.5.1. Quantitative results

The experimental results for MRI-CT image fusion are reported in Table 6 and Table 7, where our method outperforms U2Fusion [49] for most of the metrics. Compared to the remaining methods, our approach demonstrates competitive performance in metrics as EN [61], MI [62], MS-SSIM [51], N_{abf}, Q_{abf}, and FMI [63]. Meanwhile, our method achieves similar performance in the MRI-PET image fusion task. Furthermore, the model trained using a single-stage training approach achieves a similar performance to the model trained using both stages simultaneously. Regarding the parameter count of the models, except for MATR [38] and IFCNN [15], most models have a similar order of magnitude in terms of parameter count.

4.5.2. Visual results

Fig. 13 demonstrates the fused products for the MRI-CT image fusion task. From these images, it can be observed that except for MATR [38], the compared methods successfully preserve the structural

[17] https://github.com/tthinking/MATR
[18] https://github.com/lok-18/GeSeNet

Table 5
Average metrics of all compared DL-based approaches on 26 samples of the TNO dataset. The results of the multi-task training-based methods are in orange color. The best, second, and third results are in red, blue, and bold, respectively.

	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	#Param. (M)
NSCT [71]$^{\text{Infr.-2013}}$	6.2825	12.5650	0.8863	0.0011	0.2497	0.8971	0.4548	0.2860	\
IFCNN [15]$^{\text{IF-2019}}$	6.6554	13.3107	0.9132	0.0290	0.2811	0.8961	0.4727	**0.3632**	**0.083M**
ReCoNet [72]$^{\text{ECCV-2022}}$	6.8035	13.6070	0.8953	0.0185	0.2600	0.8920	0.4743	0.3433	0.007M
YDTR [37]$^{\text{TMM-2022}}$	6.7246	13.4493	0.9153	0.0273	0.2236	0.8803	0.4764	0.2685	0.2M
LRRNet [41]$^{\text{TPAMI-2023}}$	6.7207	13.4414	0.7716	0.0672	0.1215	0.8712	0.4071	0.2678	0.049M
SwinFusion [39]$^{\text{JAS-2022}}$	**6.7819**	**13.5639**	0.8999	0.0271	**0.3385**	**0.9007**	0.4773	0.4607	0.973M
CDDFuse [3]$^{\text{CVPR-2023}}$	6.7819	14.0701	0.9069	0.0677	0.3009	0.8983	0.4875	0.4255	1.188M
U2Fusion [49]$^{\text{TPAMI-2020}}$	6.2690	12.5380	0.8875	0.0000	0.2218	0.8952	**0.4925**	0.2925	2.636M
Ours (only stage 1)	6.4214	12.8429	0.8800	**0.0037**	0.3560	0.9020	0.4955	0.3401	1.533M
Ours (2 stages)	6.6389	13.2778	**0.9111**	0.0043	0.3413	0.9131	0.4928	0.3209	1.803M

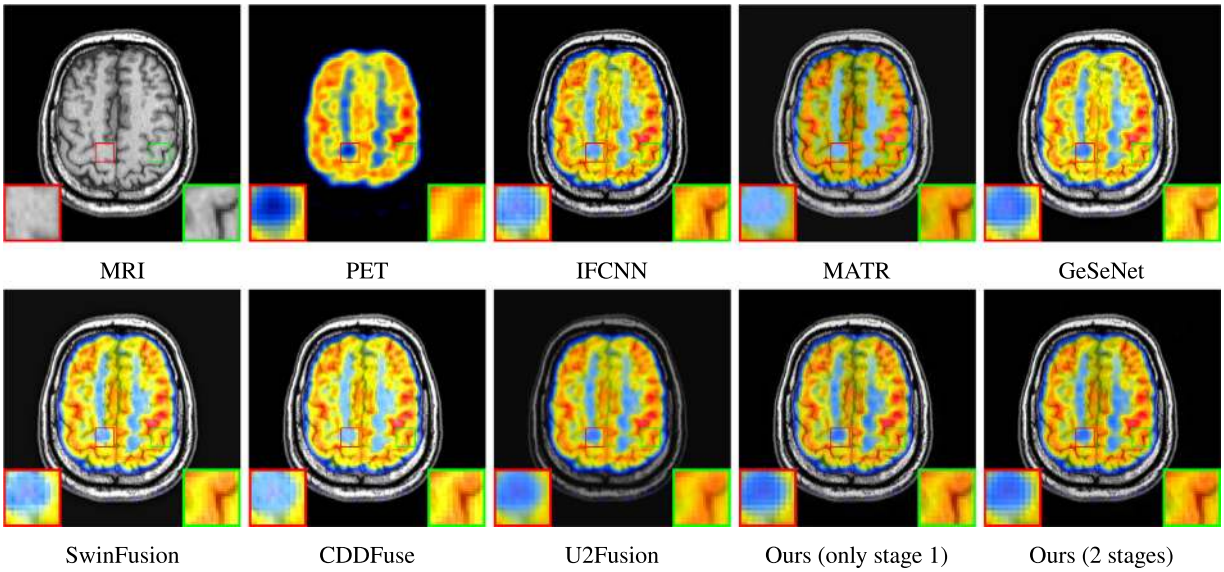

Fig. 14. Visual comparisons on the Harvard medical image dataset for the MRI-PET medical image fusion task.

information of the CT image in the fused images. However, none of them effectively extracts spatial structural information from the MRI image. In contrast, our method outperforms the competing methods regarding visual quality. Fig. 14 depicts the fused products for the MRI-PET image fusion task. From the enlarged green rectangle, it can be observed that the fused images generated by U2Fusion [49] fail to effectively preserve the details in the MRI image. From the enlarged red rectangle, it can be seen that the colors of the fused images generated by our method are closest to the colors of the PET image. This indicates that our method can better preserve the details from both source images.

4.6. Ablation experiments

4.6.1. Effects of multi-task semi-supervised learning

We proposed a multi-task semi-supervised framework to explore complementary information among different image fusion tasks and promote the learning of unlabeled data through supervised learning using labeled data. Our training framework consists of two stages: the first stage involves the pre-training of the model using multi-task supervised learning, and the second stage fine-tunes the fusion modules using multi-task semi-supervised learning. To validate the effectiveness of our proposed training framework, we conducted a series of ablation experiments

Effects of the Multi-task Supervised Learning In the first stage of our training framework, we simultaneously employed supervised learning to train the model on both the MFF and MEF tasks. From Table 8, the performance of the model trained using multi-task learning on a single image task is comparable to that of a model trained solely using single-task training. From Fig. 15, it can be observed that the fusion images generated by the model trained solely on the MFF task exhibit noticeable distortions. Similarly, the fusion images generated by the model trained solely on the MEF task lack significant details. On the other hand, the model trained using multi-task learning generates images with rich details and effectively restores details in underexposed and overexposed images. The experimental results demonstrate that the model trained using multi-task learning can indeed extract complementary information from different tasks without necessarily sacrificing performance on individual tasks.

Effects of the two-stage training From Table 9, it can be observed that without the stage 1 training, the model achieves good performance on the IVF task but experiences a significant decline in performance on the MEF task. Moreover, the semi-supervised fine-tuning in the second stage noticeably improves the performance of the model on the IVF task. From Fig. 16, it can be seen that without the stage 1 training, the model achieves good metrics, but the fusion images generated in the MEF task are noticeably overexposed, while the fusion images generated in the IVF task lose details. On the other hand, the semi-supervised fine-tuning in the second stage improves the accuracy of the intensity in the generated images.

Effects of the supervised training in stage 2 In the second stage of training, in addition to unsupervised training, we also propose to simultaneously perform the same supervised training as in the first stage. This allows us to leverage supervised training to facilitate the extraction of richer, more detailed information in unsupervised learning. From Table 9, it can be observed that the absence of supervised training does not lead to a decrease in the metrics. However, from Fig. 16, it can be seen that the fusion images generated by the model

Table 6
Average metrics of all the compared approaches on 20 samples of the Harvard medical image dataset for the task of MRI-CT image fusion. The results of the multi-task training-based methods are in orange color. The best, second-best, and third results are in red, blue, and bold, respectively.

Method	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	#Param. (M)
IFCNN[IF-2019] [15]	4.4958	8.9916	0.9410	0.0156	0.3051	0.8848	0.3419	**0.7540**	0.083M
MATR [38][TIP-2022]	4.5836	9.1673	0.6031	**0.0032**	0.0925	0.8475	0.2749	0.7116	0.013M
GeSeNet [4][TNNLS-2023]	4.7744	9.5488	0.9207	0.0178	0.3345	0.8871	0.3190	0.7530	**0.241M**
SwinFusion [39][JAS-2022]	4.0171	8.0342	0.9352	0.0105	0.3388	0.8940	0.6777	0.8592	0.973M
CDDFuse [3][CVPR-2023]	4.2861	8.5722	0.9275	0.0175	0.3049	0.8871	0.6825	0.8018	1.188M
U2Fusion [49][TPAMI-2020]	4.4808	8.9616	0.8625	0.0000	0.1232	0.8819	0.3529	0.7389	2.636M
Ours (only stage 1)	4.8504	9.7008	0.9061	0.0022	0.3121	**0.8927**	0.4164	0.7005	1.803M
Ours (2 stages)	**4.6340**	**9.2680**	**0.9303**	0.0038	**0.3258**	0.8964	**0.6069**	0.7151	1.803M

Table 7
Average metrics of all the compared approaches on 30 samples of the Harvard medical image dataset for the task of MRI-PET image fusion. The results of the multi-task training-based methods are in orange color. The best, second, and third results are in red, blue, and bold, respectively.

Method	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	#Param. (M)
IFCNN [15][IF-2019]	5.6275	11.2550	0.9346	0.0091	0.6667	0.8414	0.4169	0.5893	0.083M
MATR [38][TIP-2022]	5.6011	11.2023	0.8188	0.0009	0.7309	0.8344	0.3600	0.8272	0.013M
GeSeNet [4][TNNLS-2023]	**5.9935**	**11.9869**	**0.9314**	0.0125	0.6752	0.8525	0.4060	0.6173	**0.241M**
SwinFusion [39][JAS-2022]	6.0200	12.0400	0.9327	0.0113	0.7184	0.8702	0.3791	**0.6879**	0.973M
CDDFuse [3][CVPR-2023]	5.5631	11.1261	0.9223	0.0162	0.7210	**0.8641**	0.6286	0.7674	1.188M
U2Fusion [49][TPAMI-2020]	5.4823	10.9646	0.8293	0.0000	0.2378	0.8270	0.3959	0.6145	2.636M
Ours (only stage 1)	6.1060	12.2121	0.9151	**0.0014**	0.7127	0.8600	**0.4956**	0.5673	1.803M
Ours (2 stages)	5.7869	11.5738	0.9168	0.0016	**0.7197**	0.8654	0.5383	0.6487	1.803M

Table 8
Ablation experiments for the multi-task training of the first training stage on two image fusion tasks. The best results are in red.

Training method	MFF (Lytro)								MEF (SICE)		
	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	PSNR↑	MS-SSIM↑	MEF-SSIM↑
W/O MEF-training	7.4942	14.9885	0.9916	0.0064	0.3701	0.8936	0.7412	0.9714	19.9804	0.8187	0.7135
W/O MFF-training	7.4611	14.9221	0.9563	0.0067	0.3348	0.8742	0.6007	0.6290	22.8046	0.8491	0.7368
Ours (Multi-task)	7.4944	14.9888	0.9924	0.0054	0.3698	0.8935	0.7408	0.9657	22.7647	0.8495	0.7375

Table 9
Ablation experiments on the two image fusion tasks. The best results are in red.

Training method	IVF (TNO)								MEF (SICE)		
	EN↑	MI↑	MS-SSIM↑	N_{abf} ↓	Q_{abf} ↑	FMI↑	Q_{cb} ↑	NMI↑	PSNR↑	MS-SSIM↑	MEF-SSIM↑
W/O stage 1	6.8248	13.6495	0.8862	0.0209	0.3345	0.9124	0.4949	0.5331	12.9348	0.8177	0.7101
W/O stage 2	6.4214	12.8429	0.8800	0.0037	0.3560	0.9020	0.4955	0.3401	22.3067	0.8409	0.7254
W/O supervised (stage 2)	6.8040	13.6079	0.8729	0.0066	0.3952	0.9161	0.4983	0.6135	17.6702	0.8335	0.7187
W/O fixed parameters	6.8195	13.6389	0.8807	0.0055	0.3200	0.9148	0.4887	0.6379	\	\	\
Full	6.6389	13.2778	0.9111	0.0043	0.3413	0.9131	0.4928	0.3209	22.7647	0.8495	0.7375

without supervised training noticeably lack structural information. This indicates that supervised learning indeed helps the model extract richer information in the unsupervised task.

Effects of fixing parameters In the second stage of training, we propose to fix the parameters of the encoders and decoder, and only update the parameters of the FAF modules. This allows the complementary information extracted in the first training stage to be shared by the model in the second training stage. From Table 9, without fixing the parameters, the model still achieves competitive performance. However, as shown by Fig. 16, the fused outcome demonstrates less details. This indicates that fixing the parameters of the encoders and decoder helps share task complementary information.

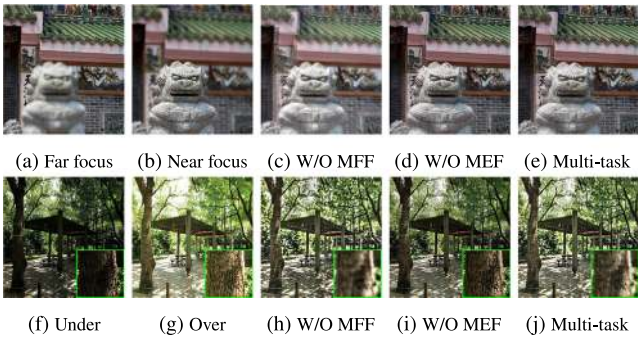

Fig. 15. Visual comparisons of single-task training and multi-task training. The first row and the second row show the fusion product on the Lytro dataset for the MFF task, and the fusion product on the SICE dataset for the MEF task, respectively.

4.6.2. Ablation experiments on the PSLPT structure

The proposed PSLPT has the following characteristics: (i) it utilizes two independent encoders to process the two source images separately, which helps extract unique features from each source image; (ii) the encoders, fusion modules, and decoders are separated, allowing PSLPT to learn both image fusion and source image reconstruction. This aids in the better decomposition of the source images; (iii) the PSLPT learns the fusion rule instead of using manually designed fusion rules; (iv) PSLPT adopts the transformer instead of CNN to extract features, enabling the learning of long-range features. To validate the effectiveness of these designs, we conducted a series of ablation experiments on the MEF task using the SICE [56] dataset. The results are shown in Table 10 and Fig. 17.

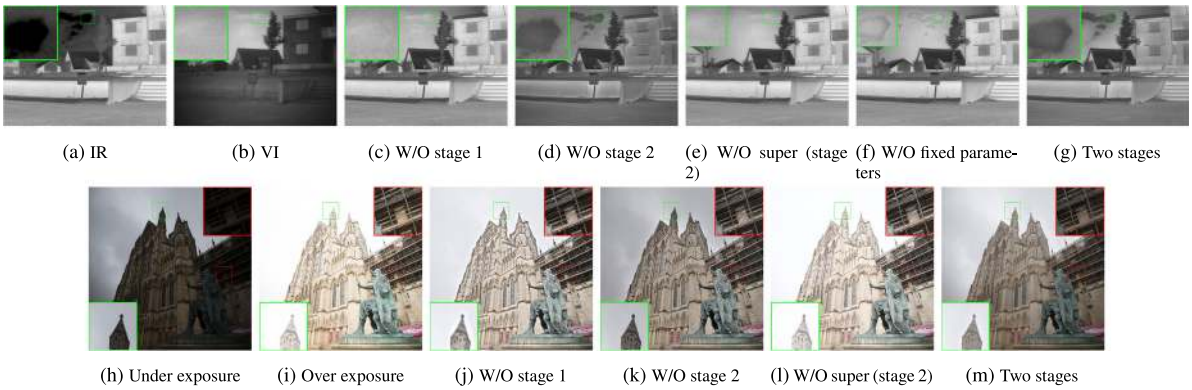

Fig. 16. Visual comparisons of different variant training strategies. The first row and the second row show the fusion product on the SICE dataset for the MEF task, and the fusion product on the TNO dataset for the IVF task, respectively.

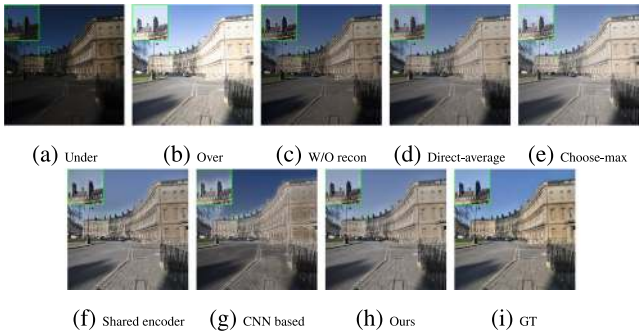

Fig. 17. Visual comparisons of PSLPT with various structures on the SICE dataset.

Table 10
Ablation experiments about the PSLPT structures on the SICE dataset. The best results are in red. "W/O Recon" means that PSLPT does not learn to reconstruct the source images.

Method	PSNR↑	MS-SSIM↑	MEF-SSIM↑
W/O Recon	15.3934	0.8166	0.7088
Direct-average	16.5957	0.8310	0.7112
Choose-max	16.2797	0.8263	0.6764
Shared encoder	21.2312	0.8402	0.7242
CNN-Based	17.2044	0.8051	0.7086
Ours (1 stage)	22.3067	0.8409	0.7254

From Table 10, it can be observed that all the variants of the PSLPT model showed a significant decrease in performance in the MEF task. Additionally, in Fig. 17, from the rectangular boxes in the images, it can be observed that the variant models using choose-max or shared encoder exhibit significant distortions in the generated fusion images. On the other hand, the variant models using the direct-average rule and the one that does not learn to reconstruct the source images produce underexposed fusion images. Furthermore, due to the limited receptive field, the CNN-based PSLPT produces fusion images with highly inaccurate exposure. Our approach, however, generates fusion images with the most satisfactory visual results.

5. Conclusion

In this work, we proposed the first multi-task semi-supervised learning framework for general image fusion. Our training framework can extract complementary information from different image tasks while leveraging the learning from labeled data to enhance the learning from unlabeled data. Our training framework consists of two stages: the first stage involves multi-task supervised pre-training, and the second stage involves multi-task semi-supervised fine-tuning. In addition, we propose the PSLPT for general image fusion, which can decompose the source image into multi-frequency features and fuse them with learned fusion rules. The PSLPT consists of two Laplacian pyramid networks with the same structure. Each of them has a multi-scale encoder and a multi-scale decoder with the parameters of the decoder being shared. The entire PSLPT uses the Swin-Transformer module to extract features. We conducted a series of experiments and the results corroborated the effectiveness of our method. In four mainstream image fusion tasks, we compare our method with the current SOTA image fusion methods. The experimental results show that our method is very competitive in both quantitative metrics and visual performance.

CRediT authorship contribution statement

Wu Wang: Writing – original draft, Software, Methodology. **Liang-Jian Deng:** Writing – review & editing, Methodology, Funding acquisition. **Gemine Vivone:** Writing – review & editing, Validation.

Declaration of competing interest

This is No Conflict of Interest.

Data availability

The data that has been used is confidential.

Acknowledgments

This research is supported by National Natural Science Foundation of China (12271083), and National Key Research and Development Program of China (Grant No. 2020YFA0714001).

References

[1] X. Hu, J. Jiang, X. Liu, J. Ma, ZMFF: Zero-shot multi-focus image fusion, Inf. Fusion 92 (2023) 127–138.
[2] H. Zhang, J. Ma, IID-MEF: A multi-exposure fusion network based on intrinsic image decomposition, Inf. Fusion 95 (2023) 326–340, http://dx.doi.org/10.1016/j.inffus.2023.02.031.
[3] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, L. Van Gool, Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 5906–5916.
[4] J. Li, J. Liu, S. Zhou, Q. Zhang, N.K. Kasabov, Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–14, http://dx.doi.org/10.1109/TNNLS.2023.3293274.
[5] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, G. Vivone, A triple-double convolutional neural network for panchromatic sharpening, IEEE Trans. Neural Netw. Learn. Syst. (2022) http://dx.doi.org/10.1109/TNNLS.2022.3155655.

[6] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D.-F. Hong, G. Vivone, Fusformer: A transformer-based fusion network for hyperspectral image super-resolution, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5, http://dx.doi.org/10.1109/LGRS.2022.3194257.

[7] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, J. Chanussot, Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks, IEEE Trans. Neural Netw. Learn. Syst. 33 (12) (2022) 7251–7265, http://dx.doi.org/10.1109/TNNLS.2021.3084682.

[8] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J.-F. Hu, G. Vivone, VO+Net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening, IEEE Trans. Geosci. Remote Sens. 60 (2021) http://dx.doi.org/10.1109/TGRS.2021.3066425.

[9] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (2017) 100–112.

[10] Q. Cao, L.-J. Deng, W. Wang, J. Hou, G. Vivone, Zero-shot semi-supervised learning for pansharpening, Inf. Fusion 101 (2024) 102001.

[11] H. Wang, H. Zhang, X. Tian, J. Ma, Zero-sharpen: A universal pansharpening method across satellites for reducing scale-variance gap via zero-shot variation, Inf. Fusion 101 (2024) 102003.

[12] R. Dian, A. Guo, S. Li, Zero-shot hyperspectral sharpening, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 12650–12666.

[13] R. Dian, T. Shan, W. He, H. Liu, Spectral super-resolution via model-guided cross-fusion network, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–12.

[14] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, G. Li, Ham-mfn: Hyperspectral and multispectral image multiscale fusion network with rap loss, IEEE Trans. Geosci. Remote Sens. 58 (7) (2020) 4618–4628.

[15] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, IFCNN: A general image fusion framework based on convolutional neural network, Inf. Fusion 54 (2020) 99–118.

[16] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.

[17] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, P. Li, Didfuse: Deep image decomposition for infrared and visible image fusion, in: International Joint Conference on Artificial Intelligence, IJCAI, 2020, pp. 970–976.

[18] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, J. Liu, Efficient and model-based infrared and visible image fusion via algorithm unrolling, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1186–1196.

[19] H. Zhang, H. Shen, Q. Yuan, X. Guan, Multispectral and SAR image fusion based on Laplacian pyramid and sparse representation, Remote Sens. 14 (4) (2022) 870.

[20] W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, Y. Li, Laplacian pyramid dense network for hyperspectral pansharpening, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–13.

[21] S. Huang, D.W. Messinger, An unsupervised Laplacian pyramid network for radiometrically accurate data fusion of hyperspectral and multispectral imagery, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–17, http://dx.doi.org/10.1109/TGRS.2022.3168511.

[22] J. Sun, Q. Han, L. Kou, L. Zhang, K. Zhang, Z. Jin, Multi-focus image fusion algorithm based on Laplacian pyramids, J. Opt. Soc. Amer. A 35 (3) (2018) 480–490, http://dx.doi.org/10.1364/JOSAA.35.000480.

[23] M. Cai, J. Yang, G. Cai, Multi-focus image fusion algorithm using LP transformation and PCNN, in: 2015 6th IEEE International Conference on Software Engineering and Service Science, ICSESS, 2015, pp. 237–241, http://dx.doi.org/10.1109/ICSESS.2015.7339045.

[24] J. Du, W. Li, B. Xiao, Q. Nawaz, Union Laplacian pyramid with multiple features for medical image fusion, Neurocomputing 194 (2016) 326–339.

[25] A. Sahu, V. Bhateja, A. Krishn, Himanshi, Medical image fusion with Laplacian pyramids, in: 2014 International Conference on Medical Imaging, M-Health and Emerging Communication Systems, MedCom, 2014, pp. 448–453, http://dx.doi.org/10.1109/MedCom.2014.7006050.

[26] H. Yin, J. Xiao, Laplacian pyramid generative adversarial network for infrared and visible image fusion, IEEE Signal Process. Lett. 29 (2022) 1988–1992.

[27] J. Shen, Y. Zhao, S. Yan, X. Li, et al., Exposure fusion using boosting Laplacian pyramid, IEEE Trans. Cybern. 44 (9) (2014) 1579–1590.

[28] X. Li, X. Guo, P. Han, X. Wang, H. Li, T. Luo, Laplacian redecomposition for multimodal medical image fusion, IEEE Trans. Instrum. Meas. 69 (9) (2020) 6880–6890.

[29] W. Wang, W. Zeng, Y. Huang, X. Ding, Deep multiscale feedback network for hyperspectral image fusion, IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.

[30] J. Yao, Y. Zhao, Y. Bu, S.G. Kong, J.C.-W. Chan, Laplacian pyramid fusion network with hierarchical guidance for infrared and visible image fusion, IEEE Trans. Circuits Syst. Video Technol. (2023).

[31] X. Luo, G. Fu, J. Yang, Y. Cao, Y. Cao, Multi-modal image fusion via deep Laplacian pyramid hybrid network, IEEE Trans. Circuits Syst. Video Technol. (2023) 1, http://dx.doi.org/10.1109/TCSVT.2023.3281462.

[32] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, G. Vivone, PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–15, http://dx.doi.org/10.1109/TGRS.2023.3244750.

[33] C. Jin, L.-J. Deng, T.-Z. Huang, G. Vivone, Laplacian pyramid networks: A new approach for multispectral pansharpening, Inf. Fusion 78 (2022) 158–170.

[34] Y. Fu, T. Xu, X. Wu, J. Kittler, Ppt fusion: Pyramid patch transformer for a case study in image fusion, 2021, arXiv preprint arXiv:2107.13967.

[35] H. Li, X.-J. Wu, T. Durrani, NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, IEEE Trans. Instrum. Meas. 69 (12) (2020) 9645–9656.

[36] R. Liu, J. Liu, Z. Jiang, X. Fan, Z. Luo, A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion, IEEE Trans. Image Process. 30 (2020) 1261–1274.

[37] W. Tang, F. He, Y. Liu, YDTR: Infrared and visible image fusion via Y-shape dynamic transformer, IEEE Trans. Multimed. (2022).

[38] W. Tang, F. He, Y. Liu, Y. Duan, MATR: Multimodal medical image fusion via multiscale adaptive transformer, IEEE Trans. Image Process. 31 (2022) 5134–5149.

[39] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217.

[40] J. Liu, S. Li, H. Liu, R. Dian, X. Wei, A lightweight pixel-level unified image fusion network, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[41] H. Li, T. Xu, X.-J. Wu, J. Lu, J. Kittler, LRRNet: A novel representation learning guided fusion network for infrared and visible images, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[42] H. Li, X.-J. Wu, DenseFuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2019) 2614–2623.

[43] H. Li, X.-J. Wu, J. Kittler, MDLatLRR: A novel decomposition method for infrared and visible image fusion, IEEE Trans. Image Process. 29 (2020) 4733–4746.

[44] H. Li, X.-J. Wu, J. Kittler, RFN-nest: An end-to-end residual fusion network for infrared and visible images, Inf. Fusion 73 (2021) 72–86.

[45] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, J. Jiang, Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion, Inf. Fusion 62 (2020) 110–120.

[46] H. Xu, J. Ma, X.-P. Zhang, MEF-GAN: Multi-exposure image fusion via generative adversarial networks, IEEE Trans. Image Process. 29 (2020) 7203–7216.

[47] H. Zhang, J. Yuan, X. Tian, J. Ma, GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators, IEEE Trans. Comput. Imaging 7 (2021) 1134–1147.

[48] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.

[49] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2022) 502–518.

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 10012–10022.

[51] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Vol. 2, 2003, pp. 1398–1402.

[52] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, M. Mukeshimana, SESF-fuse: An unsupervised deep model for multi-focus image fusion, Neural Comput. Appl. 33 (2021) 5793–5804.

[53] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.

[54] J. Zhang, Q. Liao, S. Liu, H. Ma, W. Yang, J.-H. Xue, Real-MFF: A large realistic multi-focus image dataset with ground truth, Pattern Recognit. Lett. 138 (2020) 370–377.

[55] H. Zhang, Z. Le, Z. Shao, H. Xu, J. Ma, MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion, Inf. Fusion 66 (2021) 40–53.

[56] J. Cai, S. Gu, L. Zhang, Learning a deep single image contrast enhancer from multi-exposure images, IEEE Trans. Image Process. 27 (4) (2018) 2049–2062.

[57] H. Xu, J. Ma, L. Zhuliang, J. Junjun, G. Xiaojie, FusionDN: A unified densely connected network for image fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, AAAI, 2020, pp. 12484–12491.

[58] T. Alexander, The TNO multiband image data collection, Data Brief 15 (2017) 249–251.

[59] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2014.

[60] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84, http://dx.doi.org/10.1016/j.inffus.2014.10.004.

[61] R. Wesley, van Aardt Jan, A. Fethi, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens. 2 (2008) 1–28.

[62] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (2002) 1–7.

[63] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Comput. Electr. Eng. 37 (5) (2011) 744–756.

[64] Y. Chen, R.S. Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (10) (2009) 1421–1432.

[65] M. Hossny, S. Nahavandi, D. Creighton, Comments on'Information measure for performance of image fusion', 2008.
[66] C.S. Xydeas, P.V.V., Objective image fusion performance measure, Mil. Techn. Cour. 56 (4) (2000) 181–193.
[67] P. Mu, Z. Du, J. Liu, 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.
[68] D. Han, L. Li, X. Guo, J. Ma, Multi-exposure image fusion via deep perceptual enhancement, Inf. Fusion 79 (2022) 248–262, http://dx.doi.org/10.1016/j.inffus.2021.10.006.
[69] L. Qu, S. Liu, M. Wang, 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, Vol. 36, AAAI, 2022, pp. 2126–2134.
[70] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Trans. Image Process. 24 (11) (2015) 3345–3356.
[71] J. Adu, J. Gan, Y. Wang, J. Huang, Image fusion based on non-subsampled contourlet transform for infrared and visible light image, Infrared Phys. Technol. 61 (2013) 94–100.
[72] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, Z. Luo, ReCoNet: Recurrent correction network for fast and efficient multi-modality image fusion, in: European Conference on Computer Vision, ECCV, 2022, pp. 539–555.