Interactive Feature Embedding for Infrared and Visible Image Fusion

Fan Zhao, Wenda Zhao, Member, IEEE, and Huchuan Lu, Senior Member, IEEE

Abstract—General deep learning-based methods for infrared and visible image fusion rely on the unsupervised mechanism for vital information retention by utilizing elaborately designed loss functions. However, the unsupervised mechanism depends on a well-designed loss function, which cannot guarantee that all vital information of source images is sufficiently extracted. In this work, we propose a novel interactive feature embedding in a self-supervised learning framework for infrared and visible image fusion, attempting to overcome the issue of vital information degradation. With the help of a self-supervised learning framework, hierarchical representations of source images can be efficiently extracted. In particular, interactive feature embedding models are tactfully designed to build a bridge between self-supervised learning and infrared and visible image fusion learning, achieving vital information retention. Qualitative and quantitative evaluations exhibit that the proposed method performs favorably against state-of-the-art methods.

Index Terms—Hierarchical representations, infrared and visible image fusion, interactive feature embedding, self-supervised learning.

I. INTRODUCTION

VISIBLE image, captured by the reflected light of a scenario, contains abundant texture details. Complementarily, the infrared image exhibits strong anti-interference ability (e.g., smoke and night) by capturing thermal radiation. However, the detailed structure information is insufficient in the infrared image. Infrared and visible image fusion aims to construct a high-quality image by integrating the vital information from source images, which is more conducive to subsequent applications, such as security, monitoring, target tracking, and object recognition [1], [2], [3], [4].

Vital feature extraction and fusion are keys associated with infrared and visible image fusion. On the one hand, visible image mainly represents the reflected light information with detailed textures, while infrared image depicts the thermal radiation information with high contrast pixel intensities [as shown in Fig. 1(a) and (b)]. These two types of features have domain discrepancy, which needs special attention. On the other hand, both visible and thermal images contain some common vital features, such as brightness and object semantics. Thus, how to comprehensively extract and fuse the aforementioned features, including universal features and domain discrepancy features, remains a major stumbling block.

To alleviate the above problems, previous methods can be mainly divided into three categories.

1) Handcrafted feature-based methods implement image transform (e.g., multiscale transform (MST) [5], [6], [7] and hybrid models [8], [9], [10]) to extract some specific information, such as contrast and textures.

2) Convolutional neural network (CNN)-based methods (e.g., DenseFuse [11], IFCNN [12], and U2Fusion [13]) learn to extract multilevel features, thereby fusing vital and domain discrepancy information.

3) Adversarial learning-based methods [14], [15], [16] intend to fuse thermal radiation information and textural detail information through adversarial training along with designed loss.

Although infrared and visible image fusion has made progress, previous methods generally have the following limitations. On the one hand, the same image transform or normal stacked convolutions can hardly extract comprehensive features. On the other hand, single-stage feature fusion may not make the fused image retain all vital features of source images. As shown in Fig. 1, the intensity information in the visible image (e.g., the pavilion) and the texture information in the infrared image (e.g., the man) are lost in Fig. 1(c); the intensity information of source images is not well preserved in Fig. 1(d) and (e); and low-contrast dilemmas exist in Fig. 1(f) and (g). Addressing the above problems, we propose a self-supervised hierarchical feature extraction and stage-interactive feature fusion framework (IFESNet) for infrared and visible image fusion. Especially, the self-supervised hierarchical feature extraction is built to obtain comprehensive features by reconstructing source images in a self-supervised manner. Moreover, the stage-interactive feature embedding is implemented through a bidirectional hierarchical representation interaction to fuse these features.

Specifically, we first conceive the self-supervised strategy for jointly source image reconstruction and fusion. As an auxiliary task, the image reconstruction task is trained by regarding the source image as ground truth in a self-supervised
way. Therefore, richer and more comprehensive features of source images can be learned. Compared with an unsupervised mechanism, our self-supervised strategy can capture more informative representations. Furthermore, we propose an interactive feature embedding model (IFEM) to build a bridge between the self-supervised learning and the infrared and visible image fusion learning, achieving vital information retention. IFEM is formulated by interacting hierarchical representations between fusion and reconstruction tasks. Specifically, the interaction process is recursively conducted in the corresponding hierarchical layers between reconstruction and fusion tasks. Note that the hierarchical representation interaction process is bidirectional. Therefore, we utilize the internal relationship between different tasks to efficiently extract and fuse these feature representations, improving the performance of the fusion task.

As illustrated in Fig. 1(h), our IFESNet can retain a larger amount of thermal radiation information (e.g., person) from Fig. 1(b) and textural detail information (e.g., pavilions and trees) from Fig. 1(a). In addition, as shown in the locally enlarged areas, our fusion result can also effectively preserve the texture and edge information of Fig. 1(b) (e.g., the dorsal area of the person in the green box), and high-contrast intensity information of Fig. 1(a) (e.g., street lamp in red box). On the contrary, GTF, FusionGAN, and DDcGAN lose partial texture details in the infrared image, while DeepFuse and DenseFuse lose some contrast information in the visible image. Concretely, our contributions are given as follows.

1) We make an attempt to develop a self-supervised strategy to solve the vital information missing dilemmas in infrared and visible image fusion. Compared with the most widely used nonadversarial and adversarial fusion methods, our framework is simple yet effective, better improving the performance of fused images.

2) The interactive feature embedding model is designed across fusion and reconstruction tasks, gradually extracting vital information representations and promoting fusion tasks.

3) Compared with the state of the arts, our proposal can retain more vital features, including universal features and domain discrepancy features.

The remainder of this article is organized as follows. Section II introduces the related work. Section III presents our proposed framework. In Section IV, qualitatively and quantitatively results are evaluated and analyzed, and ablation studies are conducted. Finally, Section V concludes this article.

II. RELATED WORK

A. Handcrafted Feature-Based Methods

Varied attempts based on handcrafted feature extraction have been proposed. For instance, MST [5], [6], [7], sparse representation (SR) [17], [18], [19], [20], [21], subspace [22], [23], [24], hybrid models [8], [9], [10], and gradient [25] are widely used. For improving feature extraction ability, Laplace pyramid [5], contourlet [6], gradient [26], SR [21], independent component analysis (ICA) [23], principal component analysis (PCA) [22], and nonnegative matrix factorization (NMF)-based methods [24] are developed, attempting to transform the source images with the same operation. As illustrated in Fig. 1, visible image mainly represents the reflected light information, while infrared image depicts the thermal radiation information. Inherently, these two types of features are specific to the source image with domain discrepancy, which can hardly be represented in the same manner. In addition, handcrafted features cannot comprehensively represent all features of source images, resulting in limited performance.

B. Nonadversarial Fusion Methods

Deep learning models [11], [13], [27], [28], [29], [30], [31], [32], [33], [34], [35] for image fusion have been put forward, whereas feature extraction and fusion still remain an
A. Motivation

We are committed to alleviating this vital feature loss dilemma. The infrared and visible images are characterized by thermal radiation and visible gradient information, respectively. Their domain discrepancy needs special attention. On the other hand, they contain some common attribute features, such as gradient variation, intensity, contrast, and saturability. The majority fusion methods adopt the normal stacked convolutions for feature extraction, resulting in a limited performance for dealing with domain discrepancy features. Most importantly, existing fusion methods involve unsupervised strategies with elaborate loss functions for vital feature fusion. Such a mechanism is not adequate for retaining all vital information. Since it is infeasible to design a comprehensive and adaptive loss function that covers all vital features, ignoring any information (e.g., texture in infrared or intensity in visible) will result in a vital feature missing (as shown in Fig. 1).

In this section, we develop a novel IFESNet for infrared and visible image fusion. Different from the widely used unsupervised mechanism, we attempt to develop the self-supervised strategy in cooperation with stage-interactive feature embedding learning to solve the issue of vital information missing. The pipeline is illustrated in Fig. 3. Several concepts have been considered to conceive such architecture, including: 1) self-supervised hierarchical feature extraction for joint reconstruction and fusion tasks in Section III-B and 2) stage-interactive feature embedding learning in Section III-C. Please note that, since the pooling operation will reduce the spatial resolution of features, our IFESNet consists of convolutional layers to retain spatial details of the fused image effectively.

B. Self-Supervised Hierarchical Feature Extraction

By the self-supervised mechanism, we aim to achieve hierarchical feature extraction containing more informative representations, thereby boosting fusion performance. As shown
in Fig. 3, IFESNet includes the self-supervised hierarchical feature extraction network of infrared image (SHFENet-ir), the self-supervised hierarchical feature extraction network of visible image (SHFENet-vis), and the infrared and visible image fusion network (IVIFNet). In Section III-B1 and III-B2, we will detail the self-supervised strategy for joint reconstruction and fusion tasks.

1) Self-Supervised Feature Extraction: Vital feature extraction is the premise of boosting fusion performance. Thus, we aim to extract hierarchical features $F'_n$ and $F''_n$, which contain comprehensive features of infrared image $I_1$ and visible image $I_2$. Essentially, this is achieved through reconstructing images $I'_1$ and $I'_2$ using the source image as ground truth in a self-supervised way. Since hierarchical features $F'_n, F''_n$ can be reconstructed back to source images, it ensures that the corresponding hierarchical layers are competent to extract vital features of source images.

Take the SHFENet-ir as an example (see Fig. 3). From the perspective of information flow, the SHFENet-ir is constructed by a hierarchical feature interaction process between each layer of SHFENet-ir and IVIFNet. In particular, the hierarchical features $F'_n$ and $F''_n$ are totally obtained from the corresponding hierarchical features $F_n$ of IVIFNet to possess the main features of IVIFNet. Thus, $F'_n, F''_n$ in turn, constrain the hierarchical feature $F_n$ of IVIFNet to possess the main features of IVIFNet can be formulated as

$$F'_n, F''_n = C^2(Cat(F'_{n-1}, F''_{n-1}))$$

where $F'_n$ and $F''_n$ denote the hierarchical features of layer $n$ in SHFENet-ir and SHFENet-vis, respectively. $F_n$ is hierarchical feature of IVIFNet. $C$ and $C^2$ represent conducting convolution operations once and twice, respectively. Cat denotes the concatenation operation. For the last layer of SHFENet, the outputs are the reconstructed results, which can be formulated as

$$I'_1 = C^3(F'_n), \quad I'_2 = C^3(F''_n)$$

where $I'_1$ and $I'_2$ denote reconstructed results of infrared and visible images, respectively. $C^3$ represents conducting the convolution operation three times.

The self-supervised strategy ensures that hierarchical features $F'_n, F''_n$ contain the main features of source image $I_1, I_2$ during the reconstruction task. It is worth noting that the hierarchical features $F'_n, F''_n$ of the reconstruction task are totally obtained from the corresponding hierarchical features $F_n$ of the fusion task. Thus, $F'_n, F''_n$, in turn, constrain the hierarchical feature $F_n$ of IVIFNet to possess the main features of $I_1, I_2$. In other words, the self-supervised strategy promotes the fusion task. Specifically, SHFENet is constructed by six convolution layers with 3*3 kernels and 64, 128, 256, 128, 64, and one channels, respectively. Note that downsampling and upsampling structures are not adopted in SHFENet, which can avoid the loss of effective information.

2) Hierarchical Feature Fusion: In IVIFNet, we aim to generate the fusion result utilizing hierarchical features obtained by the self-supervised mechanism. Considering the extracted hierarchical features $F'_n, F''_n$ cover sufficient information of source images, it is the potential for encouraging fusion tasks. Thus, how to utilize these vital features for fusion tasks remains a problem to be solved.

To achieve this, the hierarchical feature interaction process between the reconstruction and fusion tasks is designed to gradually promote the fusion network. As shown in Fig. 3, we first concatenate the source images $I_1$ and $I_2$, and then, the hierarchical feature $F_1$ can be obtained by conducting convolution operation $Conv$ twice. Based on the idea of merging and then separating features [36], we fuse the layerwise hierarchical features $F''_n$ and $F'_{n-1}$ from SHFENet, and the hierarchical feature $F_n$ of IVIFNet can be formulated as

$$F_n = C^2(Cat \left( \frac{C(F_{n-1}), C(F''_{n-1})}{F_{n-1}} \right))$$

At this point, the hierarchical feature $F_n$ of IVIFNet is derived from hierarchical features $F'_{n-1}, F''_{n-1}$, which heuristically shares low-, mid-, and high-level features for fusion. Thus, the extracted hierarchical features $F'_n, F''_n$ via self-supervised strategy can be fully utilized for fusion task, which, in turn, avoids the vital features missing in fusion result. Besides, hierarchical connections between the two stages are
implemented by the concat operation to integrate the low-level and high-level features. Specifically, IVIFNet is constructed by ten convolution layers with 3*3 kernels and 64, 64, 128, 128, 256, 256, 256, 128, 64, and one channels, respectively. The outputs of the last convolution layer of IVIFNet are weight maps for infrared and visible images. Thus, the fusion result can be generated by using channelwise multiplication with source images $I_1, I_2$, which can be written as

$$I_f = \sum_{i=1}^{2} C^i(F_n) \otimes I_i$$

where $I_f$ denotes the fusion result. $W_i$ is the $i$th weight map, which is calculated by conducting four convolutions on $F_n$. $I_i$ represents the $i$th source image.

C. Stage-Interactive Feature Embedding Model

As shown in Fig. 3, vertically, our framework is consisted of multistage IFEMs. IFEMs conduct hierarchical feature interaction between SHFENet and IVIFNet. This allows us to jointly learn correlative representations for alleviating vital feature missing in fusion results. We argue that the layers of SHFENet and IVIFNet can be treated as different feature descriptors, and features learned from different tasks can be treated as different representations of source images. Thus, the feature representations related to reconstruction can provide additional vital features for fusion. Specifically, hierarchical feature interaction is conducted as a bridge between reconstruction and fusion tasks, which can utilize the internal relationship between different tasks to improve the feature representations, thereby boosting the performance of the fusion task. The $n$-stage IFEM for hierarchical feature interaction between $F_n$ and $F_n'$, $F_n''$, $F_n'''$, $F_n''''$ can be expressed as

$$F_n' = Bi(F_n), \quad F_n = INT(F_n', F_n''', F_n''''')$$

where $Bi$ denotes the hierarchical feature delivering process from IVIFNet to SHFENet and INT represents the hierarchical feature delivering process from SHFENet to IVIFNet, which are equivalent to formulas (1) and (3), respectively. $n$ denotes the stage number. While the interaction process is bidirectional, $F_n$ and $F_n'$, $F_n''$ can be represented by each other.

To be specific, the INT process is designed to ensure that the vital hierarchical features $F_n'$, $F_n''$ are concatenated and shared to the corresponding hierarchical layer of fusion network, thereby promoting the fusion task. The stage-interactive INT process can greatly reduce the loss of intermediate information by leveraging all the hierarchical features for fusion. The $Bi$ process aims to deliver the fused feature $F_n$ to the reconstruction task, which, in turn, ensures that the fused hierarchical feature $F_n$ contains important information of source images. Therefore, the stage-interactive feature embedding learning for hierarchical feature interaction between the reconstruction and fusion tasks can improve fusion performance.

D. Model Training

We aim to design a loss function to achieve vital feature extraction and fusion via interactive feature embedding in the self-supervised learning framework. Specifically, we jointly train the reconstruction and fusion tasks. Thus, the designed loss can be expressed as

$$L = L_I + L_V + L_F + L_M$$

where $L_I$ and $L_V$ are self-supervised reconstruction loss functions for SHFENet-ir and SHFENet-vis, respectively. $L_F$ denotes the loss function for IVIFNet. $L_M$ stands for weight map constraint.

Previous consistent loss functions (e.g., energy-based contrast constraint [37] and perceptual constraint [38]) generally fuse some specific information, such as luminance and edge. In contrast, we introduce a self-supervision constraint by designing an image reconstruction task, which is trained regarding the source image as ground truth. Thus, more comprehensive features of source images can be learned. Here, we adopt the standard mean square error (mse) to train the self-supervised hierarchical feature extraction network

$$L_I = \text{mse}(I_1, I_1), \quad L_V = \text{mse}(I_2, I_2)$$

where $I_1$ and $I_2$ represent the visible image and the infrared image, and $I_1$ and $I_2$ are the reconstructed results. The above loss functions ensure that the hierarchical layers of the reconstruction network own the ability to extract vital features of the source images.

For further fusing the vital features, we adopt a loss function based on the structural similarity index metric (SSIM) [39], [40], [41] for IVIFNet. Specifically, input images $I_n = I_n|n = 1, 2$ are represented by the components of contrast $C$, structure $S$, and luminance $I$ in the SSIM framework

$$I_n = C_n * S_n + l_n.$$  

Contrast $C_n$ and structure $S_n$ are

$$C_n = \|I_n - \mu_n\|, \quad S_n = \frac{I_n - \mu_n}{\|I_n - \mu_n\|}$$

where $\mu_n$ is the average value of $I_n$. For an expected result $\hat{T} = \xi \hat{C} \ast \hat{S}$, it should contain high contrast and the main structure of the source images. Thus, the corresponding contrast $\hat{C}$ and structure $\hat{S}$ can be expressed as

$$\hat{C} = \max_{n=1,2} C_n, \quad \hat{S} = \frac{\sum_{n=1}^{2} S_n}{\sum_{n=1}^{2} C_n}.$$  

Finally, the SSIM between the fusion result $I_f$ and the expected result $\hat{T}$ can be calculated by the following function:

$$\text{SSIM} = \frac{2\sigma_{Tf} + C}{\sigma_T^2 + \sigma_f^2 + C}$$

where $\sigma_T^2$ and $\sigma_f^2$ represent variances of $\hat{T}$ and $I_f$. $\sigma_{Tf}$ is the covariance of $\hat{T}$ and $I_f$. Thus, the loss function for IVIFNet is calculated as

$$L_F = 1 - \text{SSIM}.$$  

The weight map constraint $L_M$ aims to adjust weight map values globally, which can be written as

$$L_M = |\tau - (W_1 + W_2)|$$
where $\tau$ is a hyperparameter. Generally, a larger $\tau$ will make the network generate bigger $W_1 + W_2$, and then, the fused image will be enhanced. Thus, (13) can adjust the enhancement performance of the fused image.

**IV. EXPERIMENTS**

**A. Implementation**

The proposed model is implemented with TensorFlow on a GTX 2080TI GPU. The Adam [42] optimizer with the learning rate of 1e-4 is adopted. The batch size is 1, and the momentum value is 0.9. The weight decay is 5e-3. $\tau$ and $\xi$ are taken to 1.0 and 1.7, respectively. We train our model on the TNO database with 110 groups of infrared and visible images.\(^1\)

**B. Evaluation Criteria**

We adopt five widely used objective metrics for evaluating the performance of our method and the competitors, e.g., average gradient (AG), the gray level difference (GLD) [43], mutual information (MI) [44], spatial frequency (SF) [45], and visual information fidelity for fusion (VIFF) [46], which are consistent with subjective visual evaluation.

1) **Average Gradient:** AG reflects the clarity of the fusion image, which represents the contrast of small details and local texture changes in the image. The larger the AG, the more the structure information is retained in the fusion result

$$AG = \frac{1}{(M - 1)(N - 1)} \sum_{i=1}^{M-1} \sum_{j=1}^{N-1} \sqrt{\frac{(I_f(i + 1, j) - I_f(i, j))^2 + (I_f(i, j + 1) - I_f(i, j))^2}{2}}$$  \hspace{1cm} (14)

where $M$ and $N$ are the width and height of fusion image $I_f$ and $(i, j)$ denotes the location of each pixel.

2) **Mutual Information:** MI can be used to measure the correlation between fused images and source images in the fusion field. Larger MI represents more vital information contained in the fusion result

$$MI = MI(I_1, I_f) + MI(I_2, I_f)$$  \hspace{1cm} (15)

where $MI(I_1, I_f)$ and $MI(I_2, I_f)$ denote the correlation among the fusion result, the infrared image, and the visible image, respectively. $MI(I_n, I_f)$ is defined as

$$MI(I_n, I_f) = E(I_n) + E(I_f) - E(I_n, I_f)$$  \hspace{1cm} (16)

where $E(I_n)$ and $E(I_f)$ denote the information entropy of image $I_n$ and $I_f$, respectively. $E(I_n, I_f)$ is the joint information entropy.

3) **Gray Level Difference:** GLD denotes the amount of gradient information in the fused image. A larger GLD stands for more texture information contained in the fused result

$$GLD = \frac{1}{(M - 1)(N - 1)} \sum_{i=1}^{M-1} \sum_{j=1}^{N-1} | I_f(i, j) - I_f(i + 1, j) | + | I_f(i, j) - I_f(i, j + 1) |$$  \hspace{1cm} (17)

4) **Spatial Frequency:** SF reflects the change of the image gray level, which is consisted of horizontal and vertical gradients. A larger SF denotes a clearer fusion result

$$SF = \sqrt{H^2 + V^2}$$  \hspace{1cm} (18)

where $H$ and $V$ are

$$H = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=2}^{N} | I_f(i, j) - I_f(i, j - 1) |^2}$$  \hspace{1cm} (19)

$$V = \sqrt{\frac{1}{MN} \sum_{i=2}^{M} \sum_{j=1}^{N} | I_f(i, j) - I_f(i - 1, j) |^2}$$  \hspace{1cm} (20)

5) **Visual Information Fidelity for Fusion:** Since we aim to mitigate the vital information loss in the fused result, we adopt another evaluation matrix, named VIFF [46], VIFF measures the amount of visible information retained in the fused image, which is consistent with the human visual system. Thus, a larger VIFF represents that more visible information is fused into the fusion image and less distortion between the fused result and source images.

**C. Results and Analysis**

In this section, we conduct qualitative and quantitative evaluations. The experiments are conducted on two widely used datasets for infrared and visible image fusion, named INFV-20 dataset\(^2\) and INFV-41 dataset\(^3\) respectively. For each dataset, we compare our model with the state-of-the-art methods, including two handcrafted feature-based methods, six nonadversarial CNN-based fusion methods, and one adversarial CNN-based fusion method, i.e., gradient transfer fusion (GTF) [26], DTCWT [6], DeepFuse [39], DenseFuse [11], IFCNN [12], U2Fusion [13], FusionGAN [14], YDTR [47], and Swinfusion [48].

1) **Quantitative Evaluation:** We first conduct a quantitative comparison between our results and the results generated by the competitors using AG, MI, GLD, SF, and VIFF metrics on INFV-20 and INFV-41. Table I summarizes the average quality metrics of different methods on the INFV-41 dataset. It is obvious that our method achieves the best performance in terms of AG, MI, GLD, and SF metrics, and subprime performance on VIFF metrics. The largest values of AG, GLD, and SF indicate that larger gradients, richer texture, and higher contrast information are retained in the results. In addition, the satisfied values of MI and VIFF denote higher similarity between the result and source image, and meanwhile, more

\(^1\)https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029

\(^2\)https://github.com/hli1221/imagefusion_densefuse

\(^3\)https://github.com/jiayi-ma/FusionGAN
visible information is retained in the fusion image. Precisely, this proves the starting point of our approach that aims to fuse more vital information. The same conclusion can be drawn from Table II. Especially, DenseFuse and U2Fusion focus on extracting multilevel features contained in both visible and thermal images. Thus, they achieve better results of VIFF and MI since the fused image retains the universal features of source images. In contrast, our fused image not only retains the universal features of source images but also fuses discrepancy features of source images, thereby obtaining better results of AG, GLD, and SF (where they comprehensively evaluate the performance of fused results, such as universal feature and discrepancy feature retention).

2) Qualitative Evaluation: We then conduct a qualitative comparison with competitors. Visual comparisons on five pairs of representative images are provided in Figs. 4–7. In this section, we compare the retention of vital features of source images including both universal features and domain discrepancy features.

a) Retaining of thermal radiation and texture information: As shown in Fig. 4(a) and (b), the visible image is represented as gradient details, while the infrared image is represented as gradient details.
mainly characterized as thermal radiation information. Thus, we evaluate the fusion performance from the perspective of the retention degree of such information.

As illustrated in Fig. 4, on the whole, all methods can fuse the main features of source images to some extent. More concretely, GTF can greatly fuse the thermal radiation
information with high pixel intensity, while the structure features in the visible image are lost. This can be explained that GTF aims to preserve the main intensity information of the infrared image and the gradient variation of the visible image. As shown in the rectangular box of Fig. 4(c), the fusion result has strong intensity similarity with the infrared image, but the visible information (e.g., the roof of the tile) is lost. As shown in the rectangular box of Fig. 4(d) and (e), both NSCT-SR and DTCWT cannot well retain the high-intensity thermal information of the infrared image and loss partial of texture information of the visible image. Overall, those methods based on handcrafted features cannot well handle domain discrepancy between source images since low-level features cannot sufficiently represent thermal radiation information and visible appearance information, resulting in vital features missing.

Fig. 4(f) and (h) shows fusion results generated from nonadversarial-based fusion methods. Intuitively, DeepFuse, DenseFuse, and IFCNN cannot highlight the thermal information well, and partial gradient details of the visible image are lost. We explain that those fusion methods focus on extracting and preserving vital features utilizing CNN structures via an unsupervised strategy. On the one hand, utilizing the same convolution operator for generating weight maps or feature extraction consequently leads to the loss of vital features that are specific to source images with domain discrepancy. On the other hand, unsupervised training is achieved by optimizing the loss function to constrain the fusion results containing the main contents of source images. Similarly, as shown in Fig. 4(i), FusionGAN based on adversarial training also loses vital information, accompanied by ambiguity and distortion. Thus, the unsupervised mechanism is unable to extract features adequately, which subsequently cannot guarantee that all vital information of source images can be retained. Transformer-based methods also lose some contrast information, as shown in Fig. 4(j) and (k). In contrast, as shown in Fig. 4(l), the vital information, including thermal information of the infrared image and gradient details of the visible image, is well retained in our fused result. Fig. 5 presents the fusion results on two representative image pairs. It is obvious that our results also appear better fusion performance containing more vital information.

b) Retaining of other vital features: As discussed previously, not only thermal information with high pixel intensity is presented in the infrared image but also other characteristics, such as gradient variation, texture, and edge information, are also included. In the same way, visible images also contain other vital features, such as intensity, contrast, and saturability. In this section, we further evaluate the retaining of other vital features of source images.

Fig. 6 presents the comparisons. It is worth noting that, here, we mainly focus on the retention of vital information in the visible image despite that our results can also maximize the fusion of thermal radiation information in infrared images with higher contrast (as shown in the red rectangle box). More concretely, Fig. 6(a) presents an abundant appearance, including high intensity, contrast, and saturability information (as shown in the green rectangle box). As illustrated in Fig. 6(c), GTF can preserve the gradient variations of the visible image, whereas the other vital information, such as intensity, contrast, and saturability, is lost. The reason is that these features are not taken into account when modeling. The same issue appears in Fig. 6(d) and (e), which are generated by NSCT-SR and DTCWT, respectively. IFCNN can retain much more vital information from visible images compared with DeepFuse, YDTR, Swinfusion, and DenseFuse. However, compared with the visible image and our result, IFCNN still encounters a certain degree of loss in brightness, contrast, and saturation.

In addition, we further evaluate the retention of other vital features from the infrared image, such as edge, gradient, and texture information. A visual comparison is provided in Fig. 7. As shown in the highlighted regions of Fig. 7(b), the infrared image presents some texture information, which is almost
infrared in Fig. 7(a). Overall, as shown in the red boxes of Fig. 7(j), our result exhibits more texture appearance than the other methods. As illustrated in Fig. 7(c)–(k), the competitors lose some details, e.g., the texture of legs and edge of the wall. This limits the application in a scenario where gradient information is unavailable in the visible image whereas abundance in the infrared image. Thus, our method can also preserve the detail information of infrared images more completely, which is ignored by other methods, since they mainly focus on the thermal radiation information. Nevertheless, we attribute the excellent performance of our method to the following: 1) self-supervised strategy can generate more and comprehensive features of the source image and 2) multistage interactive feature embedding learning can gradually integrate all vital information into fusion results and, thus, solve the vital information missing problem.

D. Ablation Study

1) Effect of Interactive Feature Embedding Learning: As described in Section III-B, IFEM is designed for promoting hierarchical feature extraction and fusion in a bidirectional interactive way. To analyze the contribution of this mechanism, we implement a variant named IFESNet w/o IFEM for comparison. The basic units of IFESNet and IFESNet w/o IFEM are provided in Fig. 8(a) and (b). As shown in Fig. 8(b), the stage of IFESNet w/o IFEM is designed with the same architecture compared with Fig. 8(a), but the data flow direction is different. To be specific, instead of using an interactive feature embedding mechanism between fusion and reconstruction tasks in IFESNet, IFESNet w/o IFEM only adopts hierarchical feature delivering from two reconstruction networks without a feature reverse delivering process from the fusion network. Hence, hierarchical feature delivering is conducted in a nonreciprocal way. For a fair comparison, IFESNet and IFESNet w/o IFEM for self-supervised hierarchical feature extraction network training. Here, the MAE-based perceptual loss in the self-supervised mechanism is used for comparison. Note that both MAE- and mse-based methods use SSIM loss to train IVIFNet. As shown in Table IV, our mse-based loss can achieve better fusion performance.

2) Effect of Self-Supervised Reconstruction Loss: In order to retain the source image’s structural details, in Section III-D, we adopt SSIM-based loss to train IVIFNet. Here, the visible perception (VP) loss [37] is adopted for comparison. Note that both VP- and SSIM-based methods use mse to train the hierarchical feature extraction branch. As shown in Table V, our SSIM-based loss obtains better fusion performance since the fused image preserves the source image’s structural details.

3) Effect of Structural Similarity Index Loss: In order to retain the source image’s structural details, in Section III-D, we adopt SSIM-based loss to train IVIFNet. Here, the visible perception (VP) loss [37] is adopted for comparison. Note that both VP- and SSIM-based methods use mse to train the hierarchical feature extraction branch. As shown in Table V, our SSIM-based loss obtains better fusion performance since the fused image preserves the source image’s structural details.

4) Effect of Varied IFEM Stages With Hierarchical Connections: As shown in Fig. 8(a), we regard each layer of IFESNet and its corresponding layer of IVIFNet with one interactive feature embedding learning process as an IFEM stage. Besides, the hierarchical connections between the two
stages are implemented, as shown in Fig. 3. Here, we aim to analyze the performance of varied IFEM stages with hierarchical connections. To be specific, we compare IFESNet with two stages (named IFESNet-HC2), three stages (named IFESNet-HC3), and four stages (named IFESNet-HC4). Quantitative evaluation on the INFV-20 dataset is presented in Table VI. With the increase in stages, the evaluation values tend to be larger. Since more low- and high-level features are integrated, the fusion performance gets better. IFESNet-HC4 is adopted for the experiment.

5) Influence of the Pooling Layer: Since the spatial pooling operation of the middle layer will reduce the spatial resolution of the features, the proposed IFESNet consists of convolutional layers only without pooling layers. We build experiments to judge the influence of the pooling layer, where two pooling layers are added to the first and second IFEM stages, respectively, and then, two upsampling layers are added to the third and fourth IFEM stages (denoted as IFESNet w pooling). Visual comparison of the proposed IFESNet without pooling (IFESNet w/o pooling) and IFESNet w pooling is shown in Fig. 9. The weight maps of visual and infrared images generated by IFESNet with pooling have blurred edges, and thus, the details are missing in the fused image. The quantitative comparison is given in Table VII. IFESNet w/o pooling achieves better results.

E. Noise Suppression

Since image fusion focuses on generating a new image that retains the source images’ details, the fused image will contain noise if the source image (e.g., infrared image) contains noise, as shown in Figs. 4–7. We adopt bilateral filtering [49] to smooth weight maps \( W_1 \) and \( W_2 \) [see (4)] while preserving edge details. Thus, the fused image \( I_f \) can be given as

\[
I_f = I_1 \times BF(W_1) + I_2 \times BF(W_2)
\]

where \( BF(.) \) stands for the bilateral filtering with a sigmagray of 10 and a sigmaspace of 7. The principle of noise suppression (NS) is that noises contained in the weight maps are suppressed, and then, the corresponding noises in source images have a low weight that cannot be fused. As shown in Fig. 10, the noises are suppressed, and the contrast of the fused results is preserved. Table VIII shows objective evaluation of NS. Objective values of IFESNet with NS (IFESNet w NS) slightly descend, which can be explained that noise is suppressed that cannot contribute its high-frequency information to improve objective values.

V. Conclusion

In this article, the interactive feature embedding in a self-supervised learning framework for infrared and visible image fusion is proposed for improving vital information retention in fusion results. In particular, the self-supervised strategy is designed for capturing more informative representations of source images, which are adequate for joint source image
reconstruction and fusion. Moreover, the stage-interactive feature embedding learning mechanism between a fusion network and two reconstruction networks is designed for embedding the vital information through stagewise hierarchical feature interaction, which essentially is implemented by leveraging all the hierarchical features from different tasks. Qualitatively and quantitatively comparisons with the state of the arts indicate that our method cannot only better fuse the thermal radiation information of the infrared image and the structural information of the visible image but also can retain the other vital information in the infrared image (e.g., texture and edge) and visible image (e.g., intensity, contrast, and saturation).

REFERENCES

[1] X. Fan, P. Shi, J. Ni, and M. Li, "A thermal infrared and visible images fusion based approach for multitarget detection under complex environment," Math. Problems Eng., vol. 2015, pp. 1774–1783, Jan. 2015.

[2] R. Raghavendra, B. Dorizzi, A. Rao, and G. H. Kumar, "Particle swarm optimization based fusion of near infrared and visible images for improved face verification," Pattern Recognit., vol. 44, no. 2, pp. 401–411, Feb. 2011.

[3] I. Ulusoy and H. Yuruk, "New method for the fusion of complementary information from infrared and visual images for object detection," IET Image Process., vol. 5, no. 1, pp. 36–48, 2011.

[4] Y. Wang et al., "Discriminative multi-view dynamic image fusion for cross-view 3-D action recognition," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 10, pp. 5332–5345, Oct. 2022.

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

[6] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," Inf. Fusion, vol. 12, pp. 74–84, Apr. 2011.

[7] X. Jin et al., "Infrared and visual image fusion method based on discrete cosine transform and local spatial frequency in discrete stationary wavelet transform domain," Inf. Phys. Technol., vol. 88, pp. 1–12, Jan. 2018.

[8] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," Inf. Fusion, vol. 24, pp. 147–164, Jul. 2015.

[9] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," Inf. Fusion, vol. 30, pp. 15–26, Jun. 2016.

[10] X. Jin, P. Du, W. Liu, and X. Liang, "A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation," Neurocomputing, vol. 226, pp. 182–191, Feb. 2017.

[11] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2614–2623, May 2019.

[12] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and Z. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," Inf. Fusion, vol. 54, pp. 99–118, Feb. 2020.

[13] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 502–518, Jan. 2022.

[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," Inf. Fusion, vol. 50, pp. 11–26, Aug. 2019.

[15] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDeGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," IEEE Trans. Image Process., vol. 29, pp. 4980–4995, 2020.

[16] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, and J. Wu, "Infrared and visible image fusion via detail preserving adversarial learning," Inf. Fusion, vol. 51, pp. 85–96, Aug. 2020.

[17] Q. Zhang and M. D. Levine, "Robust multi-focus image fusion using multi-task sparse representation and spatial context," IEEE Trans. Image Process., vol. 25, no. 5, pp. 2045–2058, May 2016.
[41] H. Li, K. Ma, H. Yong, and L. Zhang, “Fast multi-scale structural patch decomposition for multi-exposure image fusion,” IEEE Trans. Image Process., vol. 29, pp. 5805–5816, 2020.
[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.
[43] W. Zhao, H. Lu, and D. Wang, “Multisensor image fusion and enhancement in spectral total variation domain,” IEEE Trans. Multimedia, vol. 20, no. 4, pp. 866–879, Apr. 2018.
[44] R. Nava, G. Cristobal, and B. Escalante-Ramrez, “Mutual information improves image fusion quality assessments,” SPIE News Room, vol. 34, pp. 94–109, Sep. 2007.
[45] Y. Zheng, E. A. Essock, B. C. Hansen, and A. M. Haun, “A new metric based on extended spatial frequency and its application to DWT based fusion algorithms,” Inf. Fusion, vol. 8, no. 2, pp. 177–192, 2007.
[46] Y. Han, Y. Cai, Y. Cao, and X. Xu, “A new image fusion performance metric based on visual information fidelity,” Inf. Fusion, vol. 14, pp. 127–135, Apr. 2008.
[47] W. Tang, F. He, and Y. Liu, “YDTR: Infrared and visible image fusion via Y-shape dynamic transformer,” IEEE Trans. Multimedia, early access, Jul. 20, 2022, doi: 10.1109/TMM.2022.3192661.
[48] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, “SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer,” IEEE/CAA J. Autom. Sinica, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.
[49] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in Proc. 6th Int. Conf. Comput. Vis., Jan. 1998, pp. 839–846.

Fan Zhao received the B.S. degree from the Changchun University of Science and Technology, Changchun, China, in 2011, and the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2014 and 2017, respectively. She is currently a Faculty Member with the School of Physics and Electronic Technology, Liaoning Normal University, Dalian, China. Her research interests are low- and high-level vision tasks.

Wenda Zhao (Member, IEEE) received the B.S. degree from Jilin University, Changchun, China, in 2011, and the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2016. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology (DUT), Dalian, China. His research interests are low-, middle-, and high-level vision tasks.

Huchuan Lu (Senior Member, IEEE) received the M.S. degree in signal and information processing and the Ph.D. degree in system engineering from the Dalian University of Technology (DUT), Dalian, China, in 1998 and 2008, respectively. He has been a Faculty Member and a Professor with the School of Information and Communication Engineering, DUT, since 1998 and 2012, respectively. His research interests are in the areas of computer vision and pattern recognition. Dr. Lu serves as an Associate Editor for the IEEE Transactions on Cybernetics and IEEE Transactions on Circuits and Systems for Video Technology.