Heterogeneous Knowledge Distillation for Simultaneous Infrared-Visible Image Fusion and Super-Resolution

Wanxin Xiao, Yafei Zhang, Hongbin Wang, Fan Li, and Hua Jin

***Abstract*— Recently, infrared–visible image fusion has attracted more and more attention, and numerous excellent methods in this field have emerged. However, when the low-resolution images are being fused, most fusion results are of low resolution, limiting the practical application of the fusion results. Although some methods can simultaneously realize the fusion and super-resolution of low-resolution images, the improvement of fusion performance is limited due to the lack of guidance of high-resolution fusion results. To address this issue, we propose a heterogeneous knowledge distillation network (HKDnet) with multilayer attention embedding to jointly implement the fusion and super-resolution of infrared and visible images. Precisely, the proposed method consists of a high-resolution image fusion network (teacher network) and a low-resolution image fusion and super-resolution network (student network). The teacher network mainly fuses the high-resolution input images and guides the student network to obtain the ability of joint implementation of fusion and super-resolution. In order to make the student network pay more attention to the texture details of the visible input image, we designed a corner embedding attention mechanism. The mechanism integrates channel attention, position attention, and corner attention to highlight the visible image's edge, texture, and structure. For the input infrared image, the dual-frequency attention (DFA) is constructed by mining the relationship of interlayer features to highlight the role of salient targets of the infrared image in the fusion result. The experimental results show that compared with the existing methods, the proposed method preserves the image information of both visible and infrared modalities, achieves sound visual effects, and displays accurate and natural texture details. The code of the proposed method can be available at https://github.com/firewaterfire/HKDnet.**

***Index Terms*— Attention mechanism, image fusion, infrared and visible images, knowledge distillation, super-resolution.**

I. Introduction

THE visual sensor can detect the spatial texture details, and the infrared sensor can capture the thermal target in poor lighting conditions. Therefore, the images from the same scene obtained by both sensors have complementary information. If this complementary information can be integrated into an image, the comprehensiveness and accuracy of the image description for the scene can be improved, facilitating the subsequent computer vision tasks. Thus, the infrared–visible image fusion has received widespread attention [1]–[6].

Some studies on infrared–visible image fusion have achieved good fusion results when the source images are high resolution [5]–[7]. When the input images are low resolution, the fusion results will not be satisfactory. The stepped processing is commonly used to improve the fusion results' resolution. However, this kind of method is inefficient, and the artificial effect produced in the first step may be spread to the second step, which will affect the outcome. This issue has been taken into consideration in the literature [8] to jointly implement low-resolution image fusion and super-resolution. However, these methods lack the guidance of high-resolution fusion results, limiting their fusion performance.

This article proposes a heterogeneous knowledge distillation network (HKDnet) with multilayer attention embedding to jointly realize the fusion and super-resolution of infrared and visible images. The proposed approach consists of a high-resolution image fusion network and a low-resolution image fusion and super-resolution network. The former network fuses high-resolution input images and guides the low-resolution image fusion and super-resolution network to realize the combination of fusion and super-resolution. It is noted as the teacher network. The latter network is trained under the guidance of the teacher network, and it is noted as the student network. The knowledge required for heterogeneous knowledge distillation is the teacher network's fused features and fusion images. During the training process of the entire network, the knowledge in the teacher network will continue to transfer to the student network. In order to solve the problem of the lack of high-resolution fusion image labels during simultaneous image fusion and super-resolution, we use the high-resolution fusion image generated by the teacher network as the label to supervise the training of the student network.

In the existing knowledge distillation works, the teacher and the student networks usually share the same task. The teacher network is more complex with higher performance, while the student network is a lightweight network in terms

Manuscript received November 2, 2021; revised January 20, 2022; accepted January 23, 2022. Date of publication February 7, 2022; date of current version March 2, 2022. This work was supported by the National Natural Science Foundation of China under Grant 62161015. The Associate Editor coordinating the review process was Dr. Wenqiang Liu. *(Wanxin Xiao and Yafei Zhang contributed equally to this work.) (Corresponding author: Hongbin Wang.)*

Wanxin Xiao, Yafei Zhang, Hongbin Wang, and Fan Li are with the Faculty of Information Engineering and Automation and the Yunnan Provincial Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China (e-mail: 1806010730@qq.com; zyfeimail@163.com; whbin2007@126.com; 478263823@qq.com).

Hua Jin is with the Department of Anesthesiology, Affiliated Hospital of Kunming University of Science and Technology, Kunming 650031, China (e-mail: jinhuakm@163.com).

Digital Object Identifier 10.1109/TIM.2022.3149101

of the network structure. In this work, the task of the teacher network is different from that of the student network. The teacher network is only responsible for fusing high-resolution images, and the student network is responsible for the fusion and super-resolution of low-resolution images. The teacher network comprises the encoder, the fusion layer, and the decoder. The student network includes the modules to realize the fusion and the super-resolution, which is more complex than the teacher network. Therefore, the proposed network is an HKDnet.

We design different attention mechanisms for different modal images in the student network to highlight their information differences. The visible image carries a large amount of information, such as edge details and texture structure. The fusion performance could be improved by integrating the information into the fusion result. For the visible image, the corner embedding attention (CEA) mechanism is designed to enable the network to focus on the texture details of the input visible image. The CEA mechanism integrates channel attention, position attention, and corner attention to highlight the visible image's edge, texture, and structure. Compared with the visible image, the infrared image carries a large amount of brightness information of the target, which reflects the saliency of the object. If the brightness information can be transferred to the fusion result, it is helpful to improve the visual effect of the fusion result and highlight the saliency of the object. Therefore, for the input infrared image, the dual-frequency attention (DFA) is constructed by mining the relationship among the layer features to highlight the role of the salient object of the infrared image in the fusion result.

The main contributions of this work are summarized as follows.

1) A heterogeneous knowledge distillation model is proposed, which first introduces the idea of knowledge distillation into the joint implementation of infrared–visible image fusion and super-resolution to solve the problem of poor fusion quality due to lack of label guidance. This model uses a heterogeneous teacher network to generate pseudo-labels and applies the label and the fused features to train the student network for simultaneous image fusion and super-resolution.
2) The different feature extraction branches are designed for the infrared and visible images. For the feature extraction of the visible image, the CEA mechanism is proposed, which integrates corner attention, position attention, and channel attention to highlight the role of texture, edge, and structure of the visible image in the fusion and super-resolution.
3) For the feature extraction of the infrared image, the DFA mechanism is proposed, which can retain the saliency information in the infrared image. The mechanism gives greater weight to the salient areas by mining the relationship between the features of different layers of the network to preserve the saliency brightness information of the infrared image.

The remainder of this article is organized as follows. Section II presents the related work, and the proposed method is described in Section III. The experimental results are discussed in Section IV before a conclusion in Section V.

II. Related Work

A. Infrared and Visible Images Fusion

The traditional image fusion methods mainly include multiscale transform-based methods [9], [10] and sparse representation-based methods [11]–[13]. However, due to the limited feature expression ability of multiscale transform and sparse representation, the performance of these fusion methods still has room to be improved. With the assistance of deep learning, the salient features of the source image can be efficiently extracted. With the development of deep learning, image fusion methods based on deep learning have been proposed in recent years. Because the salient features of the source images can be efficiently extracted by deep learning, the image fusion method based on deep learning has achieved good performance. In 2017, Liu *et al.* [14] proposed a multifocus image fusion method based on convolutional neural networks (CNNs). The source images with different blurred parts are used to train the network. Then, a weight map is generated by the network. The fused image is obtained using the source images and the weight map. Li *et al.* [15] proposed to decompose the infrared image and the visual image to be fused into the basic part and the high-frequency part. Then, the two parts are fused with different strategies to reconstruct the fused image. Ma *et al.* [16] proposed FusionGAN for infrared–visible image fusion. A generator is used to build a fused image, and the discriminator is used to make the fused image have more details in the visible image. Li *et al.* [17] proposed a convolutional architecture based on dense connections for infrared–visible image fusion. The encoder composed of convolution layers and dense blocks are used to extract image features, and then, the fused image is generated by fusion strategy and decoder. Ma *et al.* [18] recently proposed a multiresolution infrared and visible image fusion method based on the dual discriminator GAN model, assuming that the resolutions of the two source images differ by a factor of 4. Zhang *et al.* [19] presented a general image fusion framework based on CNN (IFCNN) to handle different kinds of image fusion tasks. Lahoud and Süsstrunk [20] proposed to decompose the source images into a base layer and a detail layer for image fusion. Prabhakar *et al.* [21] performed a CNN-based approach for the exposure fusion problem. They proposed a simple CNN-based architecture that contains two CNN layers in the encoding network and three CNN layers in the decoding network.

However, the methods above assume that the source images are high resolution, and satisfactory fusion results can be achieved only under this assumption. To this end, the super-resolution method is usually used to increase the resolution after the fusion of source images. However, the artifacts generated in the first step may be spread to the next step, thereby reducing the visual quality of the final result. In order to make image fusion more feasible, a unified framework is required to realize image fusion and super-resolution simultaneously.

B. Joint Implementation of Image Fusion and Super-Resolution

Yin *et al.* [8] proposed a method for image fusion and super-resolution simultaneously based on sparse representation. It is supposed that the high-frequency components of the amplified low-resolution source image have the same sparse coding coefficients as the reconstruction result. The coefficients are used to reconstruct the high-resolution fusion image. Iqbal and Chen [22] realized simultaneous image fusion and super-resolution by assuming that the representation coefficients of the high-resolution image block and its corresponding low-resolution version are the same. According to the principle that the fused image has the same or similar structure tensor with the source images, Li *et al.* [23] proposed a joint framework for image fusion and super-resolution based on fractional differential and variational methods. Xie *et al.* [24] proposed a residual compensation method for simultaneous image fusion and super-resolution. Those methods performed well for low-scale factors. The visual quality is severely reduced for high-scale (e.g., $4\times$) factors.

C. Teacher–Student Network

The teacher–student network is a type of transfer learning, where the teacher network is a more complex network with better performance and generalization ability. The teacher network can guide the student network with fewer parameters to achieve similar performance as it. Li *et al.* [25] proposed to use a teacher network and targeted its soft output on unlabeled data as a small-sized student network. Hinton *et al.* [26] found that the additional information encapsulated in the soft output of the teacher network is helpful to train the student network. Based on this, Romero *et al.* [27] proposed a staged training strategy for the teacher–student network. The student network first learns the intermediate feature of the teacher network and then uses the actual soft output of the teacher network for training. Based on the intermediate feature representation, Yim *et al.* [28] proposed applying knowledge distillation from a DNN to another DNN of identical architecture and reported that the student model trains faster and achieves greater accuracy than the teacher. Zagoruyko and Komodakis [29] forced the student to match the attention map of the teacher (norm across the channels dimension in each spatial location) at the end of each residual stage.

The challenge in the joint implementation of image fusion and super-resolution is the lack of high-resolution fused images as labels. At present, the simultaneous image fusion and super-resolution includes traditional methods [23], [24] and deep learning-based methods [30]. The traditional methods have low efficiency and are not convenient for practical application. Due to the lack of supervision of high-resolution fusion images, the existing deep learning-based method [30] only has a good effect on the low magnification factor, and the visual effect of fusion and super-resolution is not good.

Inspired by the knowledge distillation, we propose a heterogeneous network with multilayer attention embedding for the simultaneous fusion and super-resolution of infrared and visible images. This method comprises a high-resolution image fusion network (teacher network) and low-resolution image fusion and super-resolution network (student network). Unlike the existing teacher–student networks, the proposed teacher network in this work has fewer parameters and less computation. It provides high-resolution fusion image labels for student network training. The student network needs to realize image fusion and super-resolution. It has more network layers and requires heavier computation. After the student network receives low-resolution infrared and visible images, it can directly generate the high-resolution fused image. Compared with other simultaneous image fusion and super-resolution methods, the proposed method has more obvious advantages under high magnification factors. The image fusion and super-resolution are completed on a unified framework guided by the high-resolution fused features and image. In addition, we adopt two novel attention mechanisms (DFA group (DFAG) and CEA) to assist the network to retain better the pixel intensity information of the infrared image and the texture detail information of the visible image, which are helpful to enhance the fusion performance of the network.

III. METHODS

A heterogeneous knowledge distillation framework with multilayer attention embedding, as shown in Fig. 1, is proposed to realize infrared–visible image fusion and super-resolution simultaneously. The framework is mainly composed of five parts: 1) teacher network; 2) feature extraction module; 3) feature fusion module; 4) super-resolution module; and 5) loss and training strategy. The teacher network is composed of an encoder, a fusion layer, and a decoder. The high-definition fusion label image is obtained by inputting the high-resolution source images into the pretrained teacher network. The student network is composed of the feature extraction module, the feature fusion layer, and the super-resolution module. The high-resolution fused image can be obtained by feeding low-resolution source images into the student network. The fused features and fusion results of high-resolution source images in the teacher network are used to guide the training of the student network.

Through the feedforward network, the student network directly learns the mapping function S_{θ_s} with the parameter θ_s

$$\mathbf{I}_f^s = S_{\theta_s}\left(\mathbf{I}_{\text{ir}}^l, \mathbf{I}_{\text{vis}}^l\right) \tag{1}$$

where $\mathbf{I}_{\text{ir}}^l$ and $\mathbf{I}_{\text{vis}}^l$ are low-resolution infrared and visible images, respectively, and $\mathbf{I}_f^s$ is the high-resolution fused image generated by the student network. The student network is optimized by

$$\hat{\theta}_s = \arg\min_{\theta_s} L_s\left(\mathbf{I}_f^t, \mathbf{I}_f^s\right) \tag{2}$$

where L_s is the loss function of the student network and $\mathbf{I}_f^t$ is the label generated by the teacher network. Specifically, $\mathbf{I}_f^t$ is generated by the mapping function T_{θ_t} of the teacher network

$$\mathbf{I}_f^t = T_{\theta_t}\left(\mathbf{I}_{\text{ir}}^h, \mathbf{I}_{\text{vis}}^h\right) \tag{3}$$

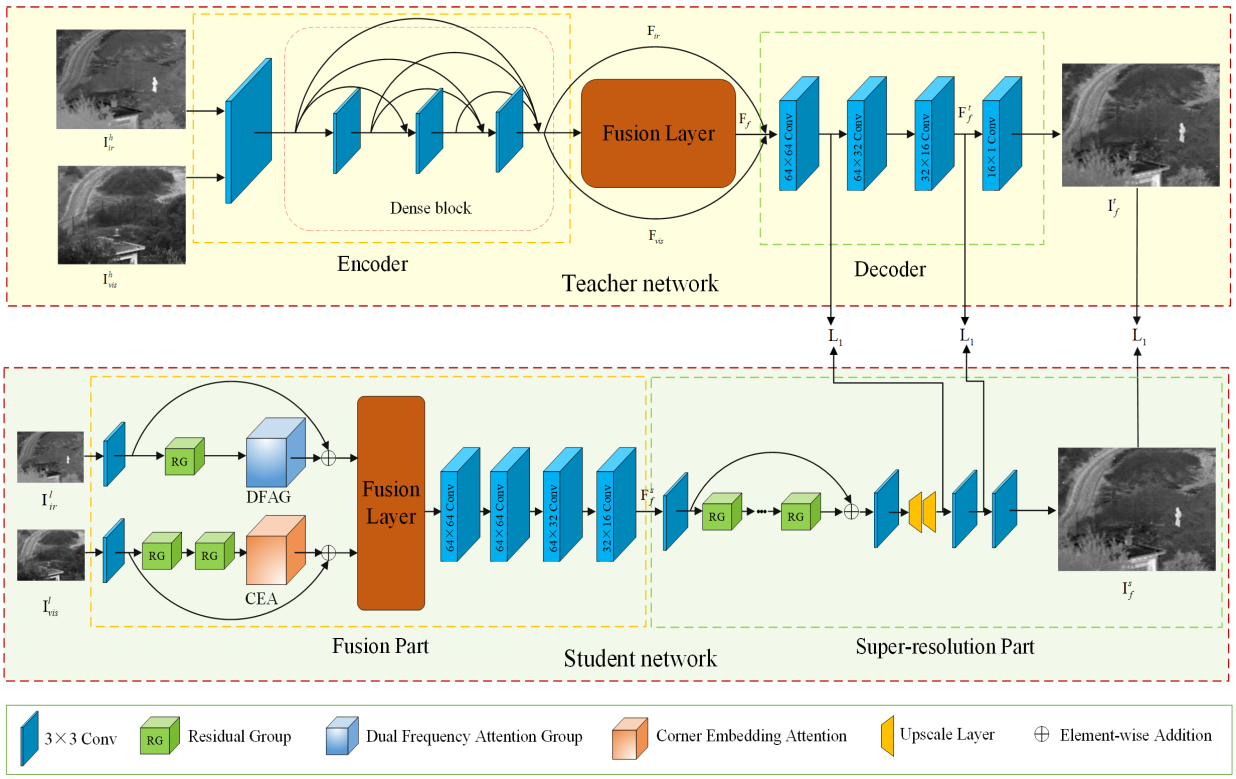

Fig. 1. Heterogeneous knowledge distillation framework with multilayer attention embedding. It consists of a teacher network and a student network. The teacher network is composed of an encoder, a fusion layer, and a decoder. It provides the labels and the intermediate layer fusion features for the training of the student network. Two feature extraction branches are used to extract image features of infrared and visible source images. The DFAG and CEA are used to assist the student network in retaining the thermal radiation information of the infrared image and the texture detail of the visible image. The super-resolution module is used to generate high-resolution fused images.

where $\mathbf{I}_{\text{ir}}^{h}$, $\mathbf{I}_{\text{vis}}^{h}$, and $\mathbf{I}_{f}^{t}$ are high-resolution infrared, visible image, and fused image, respectively. T_{θ_t} is optimized by

$$\hat{\theta}_t = \arg\min_{\theta_t} L_t(\mathbf{O}, \mathbf{I}) \tag{4}$$

where $\mathbf{O}$ and $\mathbf{I}$ are the output image and input image of the teacher network during training, respectively, and L_t is the loss function of the teacher network.

A. Teacher Network

The teacher network comprises the encoder, the fusion layer, and the decoder. The encoder consists of a 3×3 convolutional layer and a densely connected block [31]. The convolutional layer is used to extract shallow features. The densely connected block contains three convolutional layers. The output of each layer is cascaded as the input of the next layer to avoid information loss and further enhance high-frequency information. The decoder consists of four convolutional layers. The structure of the teacher network is similar to the DenseFuse network [17]. However, the fusion results of the DenseFuse suffer from severe brightness loss, so we use a residual connection to input the source image features to the decoder as well. In addition to the fused features, the features of the infrared and visible source images are also fed into the decoder. Let $\mathbf{F}_{\text{ir}}$, $\mathbf{F}_{\text{vis}}$, and $\mathbf{F}_f$ represent the infrared source image feature, visible source image feature, and fused feature, respectively. The output $\mathbf{I}_f^t$ of the decoder can be expressed as

$$\mathbf{I}_f^t = D\left(\left(\mathbf{F}_{\text{ir}} + \alpha\mathbf{F}_{\text{vis}} + \mathbf{F}_f\right), W_2\right) \tag{5}$$

where W_2 is the weight set of decoder D and α is a balance parameter that controls the input ratio of the infrared and visible image features. In order to make the fusion result supplement more pixel intensity information of source image, the value of α is maintained between 0 and 1.

B. Feature Extraction Module

In image fusion, the saliency information of the source images plays a vital role in improving the quality of the fused image. Since the information in the infrared and visible images is very different, two different extraction branches are used to form a feature extraction module. DFA is embedded into the infrared image feature extraction branch to help the network extract pixel intensity information in the infrared image. CEA is introduced into the visible image feature extraction branch to support the network extracting the texture detail information and context information in the visible image. In the infrared image feature extraction branch, a 3×3 convolution layer is used to extract the shallow features $\mathbf{F}_{\text{ir}}^{s}$ of the low-resolution infrared source image $\mathbf{I}_{\text{ir}}^{l}$. The in-depth features $\mathbf{F}_{\text{ir}}^{d}$ are obtained through a residual group (RG) [32]. Finally, the enhanced features $\mathbf{F}_{\text{ir}}^{e}$ are obtained by the DFAG composed of six DFA blocks (DFABs). In the visible image feature extraction branch, a 3×3 convolution layer is used to

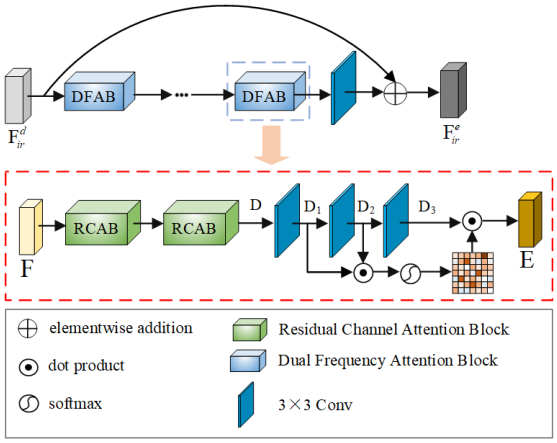

Fig. 2. Architecture of the DFAG.

extract the shallow features $\mathbf{F}^s_{\text{vis}}$ of the low-resolution visible source image $\mathbf{I}^l_{\text{vis}}$, and the in-depth features $\mathbf{F}^d_{\text{vis}}$ are obtained through two cascaded RGs. Finally, the enhanced features $\mathbf{F}^e_{\text{vis}}$ are obtained by the CEA module.

The frequency of image features varies with the network depth [33]. If the features of adjacent network layers are combined to extract the correlation between different frequency features, the correlation can enhance the image features. In the infrared image, the intensity represents the thermal radiation information [34]. Inspired by the knowledge that the different frequency features have complementary information, the features of adjacent network layers are fused by the Hadamard product. Then, the DFA map is obtained through softmax. As the network is trained, the correlation of the features from adjacent layers will be adaptively encoded into the DFA map. By weighting the features of the infrared image with a DFA map, the region with higher intensity will be assigned a more significant weight to enhance the thermal radiation information.

Based on the above idea, we proposed the DFAG, which contains six DFABs. The specific structure of DFAB is shown in Fig. 2. The feature $\mathbf{F}$ is input into two cascaded RCAB [32] to obtain the feature $\mathbf{D} \in \mathbb{R}^{C\times H\times W}$ with stronger representation capacity. Then, $\mathbf{D}$ is fed into three convolutional layers to obtain $\mathbf{D}_1 \in \mathbb{R}^{C\times H\times W}$, $\mathbf{D}_2 \in \mathbb{R}^{C\times H\times W}$, and $\mathbf{D}_3 \in \mathbb{R}^{C\times H\times W}$. After the Hadamard product of $\mathbf{D}_1$ and $\mathbf{D}_2$ passes through a softmax layer, the DFA map $\mathbf{M}_{\text{DFA}} \in \mathbb{R}^{C\times H\times W}$ can be obtained

$$\mathbf{M}_{\text{DFA}} = \text{softmax}(\mathbf{D}_1 \odot \mathbf{D}_2) \tag{6}$$

where the element in $\mathbf{M}_{\text{DFA}}$ represents the saliency of the feature of the object, which increases with the object brightness. The weighted feature $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$ is obtained

$$\mathbf{E} = \delta \mathbf{M}_{\text{DFA}} \odot \mathbf{D}_3 \tag{7}$$

where δ is the weight parameter learned from 0. The saliency of the brighter object has been enhanced in $\mathbf{E}$. As can be seen from Fig. 2, the enhanced feature $\mathbf{F}^e_{\text{ir}}$ is obtained after the infrared image feature $\mathbf{F}^d_{\text{ir}}$ passes through the DFAG.

For visible images, it will be beneficial to preserve the saliency information of source images if more attention is given to the features corresponding to the texture and edges. A better feature representation can be obtained by capturing

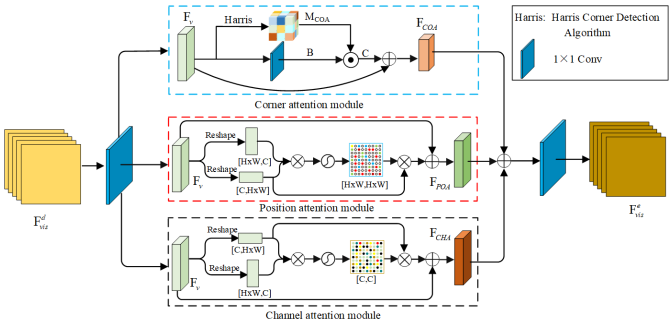

Fig. 3. Architecture of the proposed CEA.

the contextual relationships between objects in the visible image. Therefore, we design the CEA mechanism to predict the weights of visible image features in different spatial locations and capture the feature relationship in spatial and channel dimensions. As shown in Fig. 3, the CEA mechanism integrates the channel attention, the position attention, and the corner attention. $\mathbf{F}^d_{\text{vis}}$ is the input of CEA. $\mathbf{F}_v \in \mathbb{R}^{C\times H\times W}$ is obtained when $\mathbf{F}^d_{\text{vis}}$ passes through a 1×1 convolution. $\mathbf{F}_v$ are fed into the corner, position, and channel attention module, respectively.

1) Corner Attention Module: According to the gradient change in different directions of each region, the image can be divided into smooth, edge, and corner regions. The corner contains key directional cues and controls the appearance of edges and textures. In areas with dense corners, there are also lots of edges and textures [35]. Therefore, paying attention to the corner region is equivalent to paying attention to the texture edge details of the image. Inspired by the above ideas, we propose a corner attention module. The feature $\mathbf{F}_v$ is input into the corner attention module, which is processed by the Harris algorithm [35] to obtain the corner point attention map $\mathbf{M}_{\text{COA}} \in \mathbb{R}^{C\times H\times W}$

$$\mathbf{M}^{cij}_{\text{COA}} = \begin{cases} 1, & \text{Harris}\left(\mathbf{F}^{cij}_v\right) > \gamma\,\text{Max}\left(\text{Harris}\left(\mathbf{F}^c_v\right)\right) \\ 0, & \text{Harris}\left(\mathbf{F}^{cij}_v\right) \le \gamma\,\text{Max}\left(\text{Harris}\left(\mathbf{F}^c_v\right)\right) \end{cases} \tag{8}$$

where $c = 1, 2, \ldots, C$, $i = 1, 2, \ldots, W$, and $j = 1, 2, \ldots, H$. $\mathbf{M}^{cij}_{\text{COA}} = 1$ means that the position (i, j) of the cth channel of $\mathbf{F}_v$ belongs to the corner point region and, vice versa, does not belong to the corner point region. $\mathbf{F}^c_v$ denotes the cth channel of $\mathbf{F}_v$. Harris(x) represents that x is processed by the Harris function. The larger the value, the closer x is to the corner area. γ is a coefficient between 0 and 1.

Meanwhile, the feature $\mathbf{F}_v \in \mathbb{R}^{C\times H\times W}$ is input into a 1×1 convolution to obtain a new feature $\mathbf{B} \in \mathbb{R}^{C\times H\times W}$. Then, the dot product of $\mathbf{M}_{\text{COA}}$ and $\mathbf{B}$ is applied to obtain the enhanced visible image feature $\mathbf{C} \in \mathbb{R}^{C\times H\times W}$. The final output $\mathbf{F}_{\text{COA}}$ is obtained by adding the elements of $\mathbf{F}_v$ and $\mathbf{C}$. The element $\mathbf{F}^{cij}_{\text{COA}}$ in $\mathbf{F}_{\text{COA}}$ can be expressed as

$$\mathbf{F}^{cij}_{\text{COA}} = \beta \mathbf{B}^{cij} \mathbf{M}^{cij}_{\text{COA}} + \mathbf{F}^{cij}_v \tag{9}$$

where β is initialized to 0, which will be assigned a larger value as the network's ability to extract features increases. As can be seen from (9), all the locations of the feature $\mathbf{F}_v$ are weighted by the corner attention map $\mathbf{M}_{\text{COA}}$. Therefore,

the edge corner areas of the visible image have received more attention.

2) Position Attention Module: Discriminant features are important for image processing, obtained by capturing long-distance context information [36]. Based on this, the self-attention mechanism is used to capture the context in the visible image. The structure of the position attention module is shown in Fig. 3. $\mathbf{F}_v$ is reshaped into $\mathbf{F}_{v,r} \in \mathbb{R}^{C\times N}$, where $N = H \times W$. The matrix multiplication is used between the transpose of $\mathbf{F}_{v,r}$ and $\mathbf{F}_{v,r}$, and then, the pixel-level softmax function is adopted to calculate the correlation between the position $i\,(i = 1, 2, \ldots, N)$ and $j\,(j = 1, 2, \ldots, N)$

$$\omega_{i,j} = \frac{\exp\left[\left(\mathbf{F}_{v,r}^T \times \mathbf{F}_{v,r}\right)(i,j)\right]}{\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left[\left(\mathbf{F}_{v,r}^T \times \mathbf{F}_{v,r}\right)(i,j)\right]} \tag{10}$$

where $\mathbf{F}_{v,r}(i, j)$ represents the feature of position j on the ith channel of $\mathbf{F}_{v,r}$. Finally, the enhanced visible image feature $\mathbf{F}_{\mathrm{POA}} \in \mathbb{R}^{C\times H\times W}$ can be obtained by

$$\mathbf{F}_{\mathrm{POA}} = \mathbf{F}_v \times \Omega + \mathbf{F}_v \tag{11}$$

where $\Omega \in \mathbb{R}^{N\times N}$ is the matrix composed of $\omega_{i,j}$. It can be seen from (11) that $\mathbf{F}_{\mathrm{POA}}$ is the weighted sum between the original feature $\mathbf{F}_v$ and the enhanced feature by position attention. Therefore, similar semantic features in $\mathbf{F}_v$ can enhance each other, which improves the discrimination of features.

3) Channel Attention Module: For visible images, the feature maps of each channel can be regarded as a mapping of the image, and the mapping in the different channels is usually different [30], [37]. By mining the dependencies between different channels and assigning greater weights to those channels with larger mapping values, we can obtain the feature map with a stronger representation ability. Based on this idea, the channel attention module is used to capture the dependence between different channels, which is shown in Fig. 3. Similar to the position attention module, $\mathbf{F}_v$ is also reshaped into $\mathbf{F}_{v,r} \in \mathbb{R}^{C\times N}$. The matrix multiplication is used to model the dependence among different channels for $\mathbf{F}_{v,r}$ and its transpose $\mathbf{F}_{v,r}^T$ and a pixel-level softmax function is used to calculate the influence of the ith$(i = 1, 2, \ldots, C)$ channel on the jth$(i = 1, 2, \ldots, C)$ channel

$$\psi_{i,j} = \frac{\exp\left[\left(\mathbf{F}_{v,r} \times \mathbf{F}_{v,r}^T\right)(i,j)\right]}{\sum_{i=1}^{C}\sum_{j=1}^{C}\exp\left[\left(\mathbf{F}_{v,r} \times \mathbf{F}_{v,r}^T\right)(i,j)\right]}. \tag{12}$$

Finally, the feature of channel attention enhancement can be formulated as

$$\mathbf{F}_{\mathrm{CHA}} = \Psi \times \mathbf{F}_v + \mathbf{F}_v \tag{13}$$

where $\Psi \in \mathbb{R}^{C\times C}$ is the matrix composed of $\psi_{i,j}$.

The features $\mathbf{F}_{\mathrm{COA}}$, $\mathbf{F}_{\mathrm{POA}}$, and $\mathbf{F}_{\mathrm{CHA}}$ are fused to obtain better feature representation. Then, a 1×1 convolution is used to obtain the enhanced feature $\mathbf{F}_{\mathrm{vis}}^e$.

C. Feature Fusion Strategy and Super-Resolution Network

It is assumed that $\mathbf{F}_{\mathrm{ir}} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_{\mathrm{vis}} \in \mathbb{R}^{C\times H\times W}$ represents the feature of the infrared and visible images. C is the number of channels. H and W are the numbers of rows and columns, respectively. After obtaining $\mathbf{F}_{\mathrm{ir}}$ and $\mathbf{F}_{\mathrm{vis}}$, a feature fusion strategy is needed to fuse them. As investigated in [17], l_1-norm and softmax are used to fuse $\mathbf{F}_{\mathrm{ir}}$ and $\mathbf{F}_{\mathrm{vis}}$. The same fusion strategy is adopted for the fusion layer in teacher and student networks. Specifically, the l_1-norm of all channel features at each position of $\mathbf{F}_{\mathrm{ir}}$ and $\mathbf{F}_{\mathrm{vis}}$ is calculated, respectively, to obtain the activity level map $\mathbf{M}_{\mathrm{ir}} \in \mathbb{R}^{H\times W}$ and $\mathbf{M}_{\mathrm{vis}} \in \mathbb{R}^{H\times W}$

$$\mathbf{M}_{\mathrm{ir}}(i,j) = \left\|\mathbf{F}_{\mathrm{ir}}^{1:C}(i,j)\right\|_1, \quad \mathbf{M}_{\mathrm{vis}}(i,j) = \left\|\mathbf{F}_{\mathrm{vis}}^{1:C}(i,j)\right\|_1 \tag{14}$$

where $i = 1, 2, \ldots, H$ and $j = 1, 2, \ldots, W$.

Then, a 3×3 neighborhood averaging of $\mathbf{M}_{\mathrm{ir}}$ and $\mathbf{M}_{\mathrm{vis}}$ is performed to obtain $\bar{\mathbf{M}}_{\mathrm{ir}}$ and $\bar{\mathbf{M}}_{\mathrm{vis}}$. The weighting coefficients $\mathbf{W}_{\mathrm{ir}}$ and $\mathbf{W}_{\mathrm{vis}}$ for the infrared and visible features are obtained by performing softmax on $\bar{\mathbf{M}}_{\mathrm{ir}}$ and $\bar{\mathbf{M}}_{\mathrm{vis}}$

$$\mathbf{W}_{\mathrm{ir}} = \frac{\bar{\mathbf{M}}_{\mathrm{ir}}}{\bar{\mathbf{M}}_{\mathrm{ir}} + \bar{\mathbf{M}}_{\mathrm{vis}}}, \quad \mathbf{W}_{\mathrm{vis}} = \frac{\bar{\mathbf{M}}_{\mathrm{vis}}}{\bar{\mathbf{M}}_{\mathrm{ir}} + \bar{\mathbf{M}}_{\mathrm{vis}}}. \tag{15}$$

Finally, the features of each channel of $\mathbf{F}_{\mathrm{ir}}$ and $\mathbf{F}_{\mathrm{vis}}$ are weighted and summed by $\mathbf{W}_{\mathrm{ir}}$ and $\mathbf{W}_{\mathrm{vis}}$ to obtain the fused feature

$$\mathbf{F}_f^c = \mathbf{W}_{\mathrm{ir}} \odot \mathbf{F}_{\mathrm{ir}}^c + \mathbf{W}_{\mathrm{vis}} \odot \mathbf{F}_{\mathrm{vis}}^c \tag{16}$$

where $c = 1, 2, \ldots, C$. $\mathbf{F}_f^c$ denotes the cth channel of the fused feature $\mathbf{F}_f$. $\odot$ denotes dot product.

In the student network, the high-resolution fusion image can be reconstructed by feeding the fused feature into the super-resolution module. In order to ensure the effectiveness of the super-resolution and to control the parameters of the total network, we select RCAN [32] as the basic network of super-resolution. Different from RCAN, only five RGs are used in our super-resolution module, while there are ten RGs in RCAN. In addition, the high-resolution feature extraction and image reconstruction in our super-resolution module is guided by the teacher network.

D. Loss Function

The original infrared and visible images are real images, and their low-resolution versions are used as input for model training. In order to realize better performance and feature learning capabilities, the loss function enables student networks to perform fusion and super-resolution simultaneously using multitask learning. Due to the lack of real data as labels for simultaneous fusion and super-resolution, the teacher network generates the corresponding pseudo-labels to train the student network.

In the teacher network training, both the pixel loss and structural similarity loss are used to make the input image and the reconstructed image have similar pixel intensity and structure. Since texture details in the visible images are mainly represented by gradient changes [34], gradient loss is used to constrain the reconstructed image, making its gradient changes similar to the input image. In addition, perceptual loss [38] is used to realize better visual effects of reconstructed images. The total loss of teacher network can be expressed as

$$L_t = L_{\mathrm{pixel}} + \mu L_{\mathrm{ssim}} + L_{\mathrm{grad}} + L_{\mathrm{percep}} \tag{17}$$

where L_t, L_{pixel}, L_{ssim}, L_{grad}, and L_{percep} represent the total loss, pixel loss, structural similarity loss, gradient loss, and perceptual loss, respectively. μ is the weight parameter, which balances the contribution of structural similarity loss to the total loss. Pixel loss L_{pixel}, structural similarity loss L_{ssim}, gradient loss L_{grad}, and perceptual loss L_{percep} are defined as

$$L_{\text{pixel}} = \|\mathbf{O} - \mathbf{I}\|_2 \tag{18}$$

$$L_{\text{ssim}} = 1 - \text{SSIM}(\mathbf{O}, \mathbf{I}) \tag{19}$$

$$L_{\text{grad}} = \|\nabla\mathbf{O} - \nabla\mathbf{I}\|_2 \tag{20}$$

$$L_{\text{percep}} = \|\phi_{i,j}(\mathbf{O}) - \phi_{i,j}(\mathbf{I})\|_2 \tag{21}$$

where $\|\cdot\|_2$ represents the l_2-norm. SSIM$(\cdot)$ [39] represents the structure similarity function. For gradient loss, first, calculate the gradient of the input image and the output image, and then, the Euclidean distance between them is used to obtain the gradient loss value. ∇ represents the gradient operator. For perceptual loss, the input and output images are fed into the pretrained VGG16 network to obtain their respective feature representations. The Euclidean distance between them is defined as the perceptual loss. $\phi_{i,j}$ represents the feature map obtained by the jth convolution (activated) before the ith maximun pooling layer in the VGG network.

After the pretraining of the teacher network is completed, we start the training of the student network. Since the student network is complicated in structure, only the fusion part is trained in the first 60 epochs to make it have preliminary fusion ability to provide the fusion features with the better representational knowledge for the subsequent super-resolution. At this stage, the low-resolution infrared and visible light source images are fed into the student network and the teacher network. The output of the penultimate layer of the teacher network decoder is $\mathbf{F}^t_f$, which is used as a label for low-resolution feature fusion of the student network. The output of the fusion part of the student network is the low-resolution fused feature $\mathbf{F}^s_f$. Between $\mathbf{F}^s_f$ and $\mathbf{F}^t_f$, the l_1 loss is used to train the fusion part of the student network

$$L_m\left(\mathbf{F}^t_f, \mathbf{F}^s_f\right) = \left\|\mathbf{F}^t_f - \mathbf{F}^s_f\right\|_1.. \tag{22}$$

Since the loss of mean square error will cause the output image blurred [40], l_1 loss is used to provide better convergence for the student network. To transfer the rich high-resolution fusion feature knowledge from the teacher network to the student network, l_1 loss is also used in the middle layer features of the two networks. The total loss of the student network is calculated by

$$L_s\left(\mathbf{I}^t_f, \mathbf{I}^s_f\right) = \left\|\mathbf{I}^t_f - \mathbf{I}^s_f\right\|_1 + \varphi \sum_{(m,n)\in H} \left\|\mathbf{T}^m - \mathbf{S}^n\right\|_1 \tag{23}$$

where $\mathbf{I}^t_f$ and $\mathbf{I}^s_f$ are the labels generated by the teacher network and the high-resolution fusion image generated by the student network, respectively, φ is a balance parameters, $\mathbf{T}^m$ is the feature of the mth layer in the teacher network, and $\mathbf{S}^n$ is the feature of nth layer in student network. H is a set of candidate feature pairs. We use two feature pairs in the proposed method. The first feature pair is the features of the second-to-last layer in the teacher network and the student network, and the second feature pair is the feature of the fourth-to-last layer in the teacher network and the feature of third-to-last layer in the student network.

Algorithm 1 Training Procedure of HKDnet

Input: Given data: $\mathbf{I}^h_{ir}$, $\mathbf{I}^h_{vis}$, $\mathbf{I}^l_{ir}$, $\mathbf{I}^l_{vis}$;Parameters: α, γ, μ, φ; The total training epochs of the teacher network $N_t = 100$, The total training epochs of the student network $N_s = 250$, the epochs that the fusion part of the student network is trained first $N_f = 60$, batch size $m = 64$, the number of the samples for the teacher network training $M_t = 12000$, the number of the samples for the teacher network training $M_s = 12000$, the step number of each epoch k.

Output: High-resolution fusion image $\mathbf{I}^s_f$

```
Training teacher network:
for t1 = 1 : Nt do
    for k = 1 : Ms/m step do
        Select m source images {I^(1), ..., I^(m)} from training set;
        Input the selected source images into the teacher network to generate I^t_f; Update teacher network by AdamOptimizer: ∇θt(Lt)
Training student network:
for t2 = 1 : Ns do
    for k = 1 : Mt/m steps do
        Select m infrared image {I^(1)_ir, ..., I^(m)_ir};
        Select m visible image {I^(1)_vis, ..., I^(m)_vis};
        if t2 ≤ Nf then
            Update fusion part by AdamOptimizer: ∇θm(Lm)
        Update the student network by AdamOptimizer: ∇θs(Ls)
end
```

E. Training Strategy

The training is of two steps: 1) the teacher network training and 2) the student network training. At the training stage of the teacher network, we only train the encoder and decoder of the network, where the fusion layer is not trained. After the teacher network is trained, the generated labels are used to supervise and train the student network. During the training of the student network, we use a two-stage: 1) training the fusion part in the first 60 epochs and 2) training the entire student network with L_s, as described in the HKDnet training process shown in Algorithm 1.

IV. Experiment

A. Datasets and Training Details

The FLIR[1] and KAIST[2] datasets are commonly used for training models in infrared–visible image fusion. FLIR contains 14 452 infrared and visible image pairs, and KAIST

[1]https://soonminhwang.github.io/rgbt-ped-detection/

[2]https://www.flir.ca/oem/adas/adas-dataset-form/

Fig. 4. Six pairs of source images. Top row: infrared images. Bottom row: visible images.

Fig. 5. Six high-resolution fusion images generated by the teacher network.

contains 95 000 pairs. A total of 12 000 image pairs are randomly selected from the two datasets as the training set. Another 400 image pairs are selected as the validation set to verify the fusion and reconstruction capability of the network in each iteration. The visible RGB image in the dataset is converted into a grayscale image. In the test phase, the test set is composed of 15 pairs of infrared and visible images from the TNO[3] dataset and 15 pairs of infrared and visible images from the VLIRVDIF[4] dataset. Six samples of these images are shown in Fig. 4.

The high-resolution infrared and visible images are cropped in the training phase to obtain a high-resolution source pair of 128×128. Then, it is downsampled by the bilinear interpolation algorithm to obtain low-resolution images of various sizes. The teacher network is first trained. We only train the encoder and decoder to reconstruct the input image [17]. When the weights of the encoder and decoder are fixed, the fusion layer is used to fuse the infrared and visible image features extracted by the encoder, and then, the fused image is reconstructed through the decoder. The trained teacher network serves as a pretraining model to provide training labels for the student network. The high-resolution fused images generated by the teacher network are shown in Fig. 5.

The training of the student network is completed after 250 epochs. The super-resolution part is frozen in the first 60 epochs, and the loss L_m is used to train the fusion part. In the latter 190 epochs, the fusion part and the super-resolution part are trained together under the constraint of loss L_s. The learning rate is set to 1×10^{-4}, the parameter α is set to 0.5, γ is set to 0.7, μ is set to 1×10^3, and φ is set to 1. The model is trained by using the ADAM optimizer [41], and the batch size is set to 64. The PyTorch1.7 framework is selected to implement the proposed network, which is trained by using an NVIDIA RXT3090 GPU. The convergence curves of the training loss during the training of the total student network for scale 4 magnification are shown in Fig. 6(a).

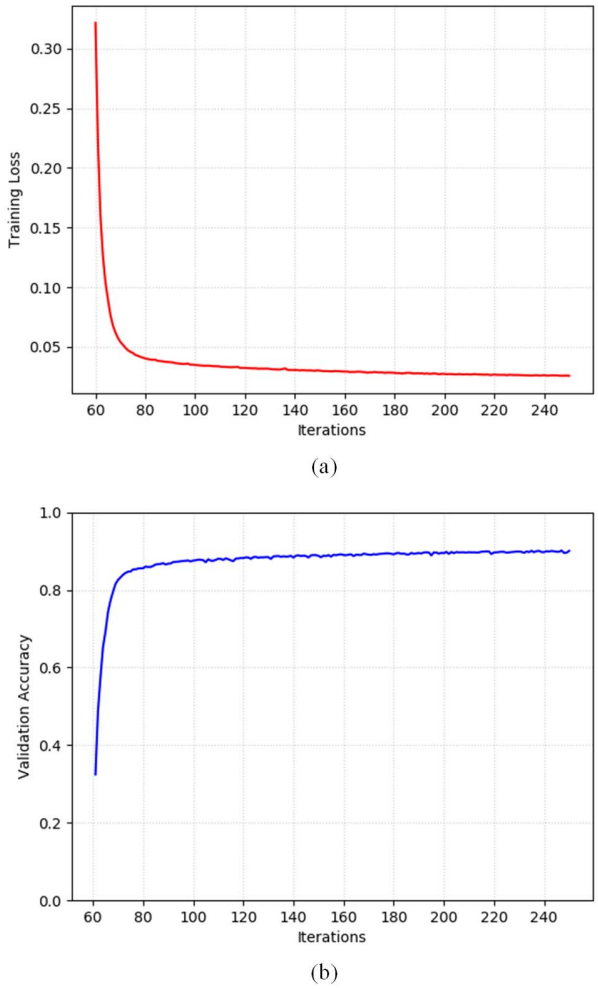

Fig. 6. (a) Training loss curve and (b) validation accuracy curve during the training of the total student network for scale 4 magnification.

As can be seen from Fig. 6(a), the training loss gradually decreases as the number of training increases. When reaching 200 epochs, the loss curve converges. The validation accuracy

[3]https://figshare.com/articles/TNO Image Fusion Dataset/1008029
[4]http://www02.smt.ufrj.br/ fusion/

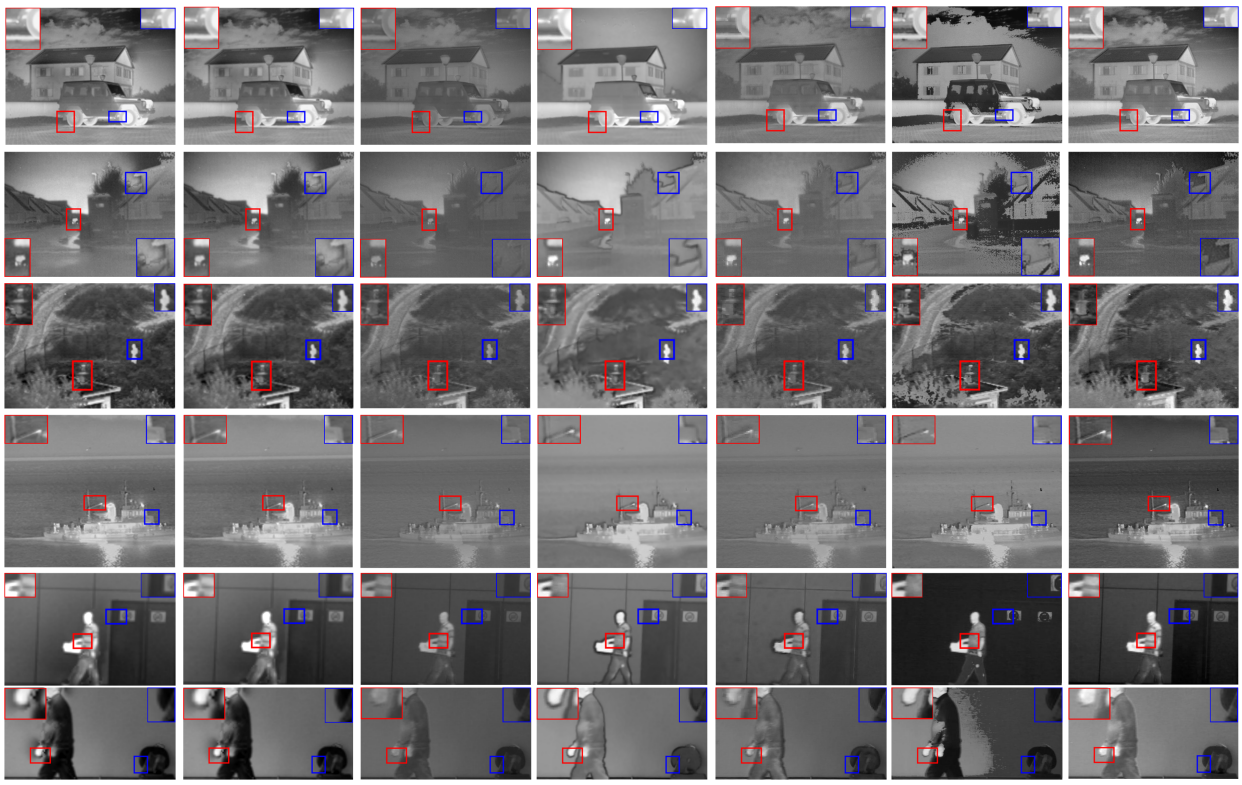

Fig. 7. Fusion results of different methods on the six examples with scale ×2. From top to bottom: results on the "Marne," "Movie," "Camp," "Vessel," "man1," and "man2" examples. From left to right: results generated by CNN+EDSR, CNN+RCAN, DenseFuse+RCAN, Yin's method, Li's method, Cen's method, and our method.

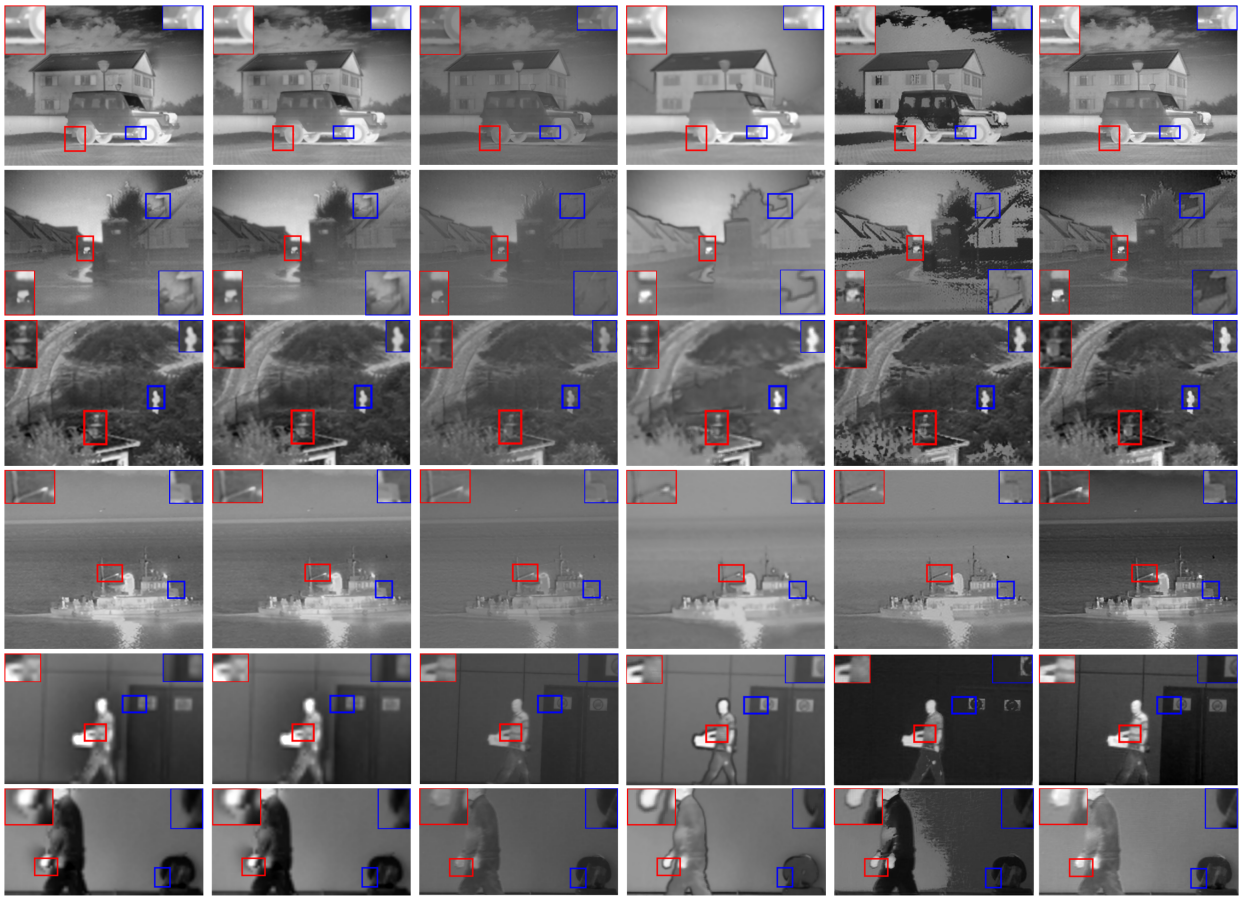

Fig. 8. Fusion results of different methods on the six examples with scale ×3. From top to bottom: results on the "Marne," "Movie," "Camp," "Vessel," "man1," and "man2" examples. From left to right: results generated by CNN+EDSR, CNN+RCAN, DenseFuse+RCAN, Yin's method, Cen's method, and our method.

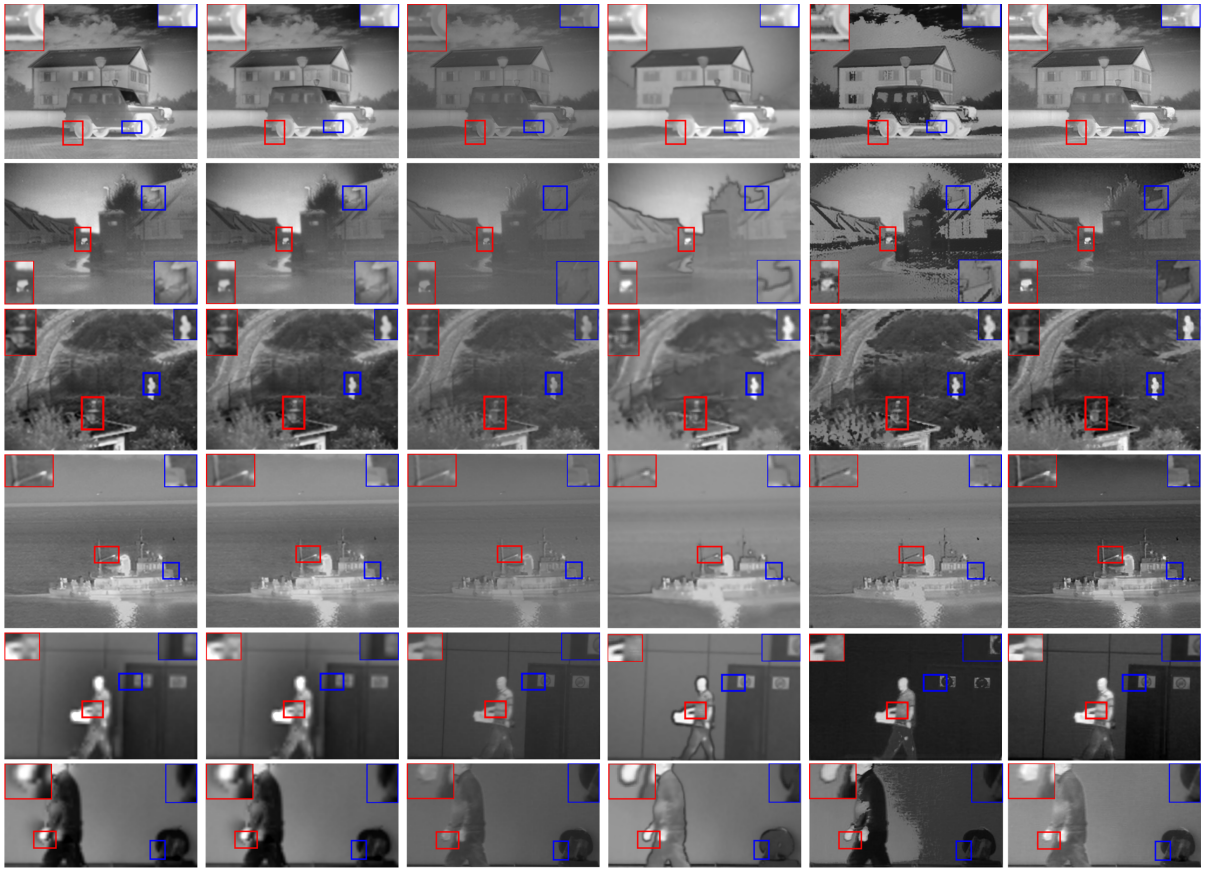

Fig. 9. Fusion results of different methods on the six examples with scale ×4. From top to bottom: results on the "Marne," "Movie," "Camp," "Vessel," "man1," and "man2" examples. From left to right: results generated by CNN+EDSR, CNN+RCAN, DenseFuse+RCAN, Yin's method, Cen's method, and our method.

curve for scale 4 magnification on validation dataset is shown in Fig. 6(b). As can be seen in Fig. 6(b), the validation accuracy improves as the number of training increases, which indicates that the difference between the fused images of the student network and the teacher network becomes smaller with training.

B. Objective Evaluation Metrics

In this work, six objective evaluation metrics are used: FMI_{dct}, FMI_{w}, SCD, MS_SSIM, AG, and CC. FMI_{dct} [42] and FMI_{w} [42] calculate the mutual information of discrete cosine and wavelet features, respectively. SCD [43] evaluates the effect of image fusion by calculating the sum of the correlation difference between the source image and the fused image. MS_SSIM [44] measures the retaining of local structure on a fine-scale and captures the brightness consistency on a coarser scale. They are used to evaluate the image perception quality of the fusion and super-resolution results. AG [45] quantifies the gradient information of the fused image and represents its detail and texture. The larger the AG metric, the more gradient information the fused image contains, and the better the performance of the fusion algorithm. CC [46] measures the degree of linear correlation of the fused image and source images. The larger CC value indicates more similarity of the fused image to the source images. For all mentioned indicators, the higher the objective evaluation value, the better the fusion and super-resolution quality.

C. Comparative Analysis of Fusion and Super-Resolution Methods

In order to verify the superiority of the proposed method, it is compared not only with the simultaneous fusion and super-resolution methods but also with the stepped methods for fusion and super-resolution. In terms of the simultaneous fusion and super-resolution method, the proposed method is compared with three representative methods: 1) the method based on fractional differential variation (Li's method [23]); 2) the method based on sparse representation (Yin's method [8]); and 3) the method based on deep learning (Cen's method [30]). For the stepped methods for fusion and super-resolution, the low-resolution fusion images are obtained first through excellent fusion algorithms such as DenseFuse [17] and CNN [14]. Then, powerful super-resolution algorithms are used, such as EDSR [40] and RCAN [32] to obtain high-resolution fused images.

The results of 2×, 3×, and 4× fusion and super-resolution obtained by the three existing simultaneous fusion and super-resolution methods, three stepped methods for fusion and super-resolution, and the proposed method are shown in Figs. 7–9. Since only 2× fusion and super-resolution can be realized by Li's method [23], the results of 3× and 4× fusion and super-resolution obtained by the other six methods are displayed. Due to the space limitation, we use six image pairs to evaluate the relevant performance on each scale.

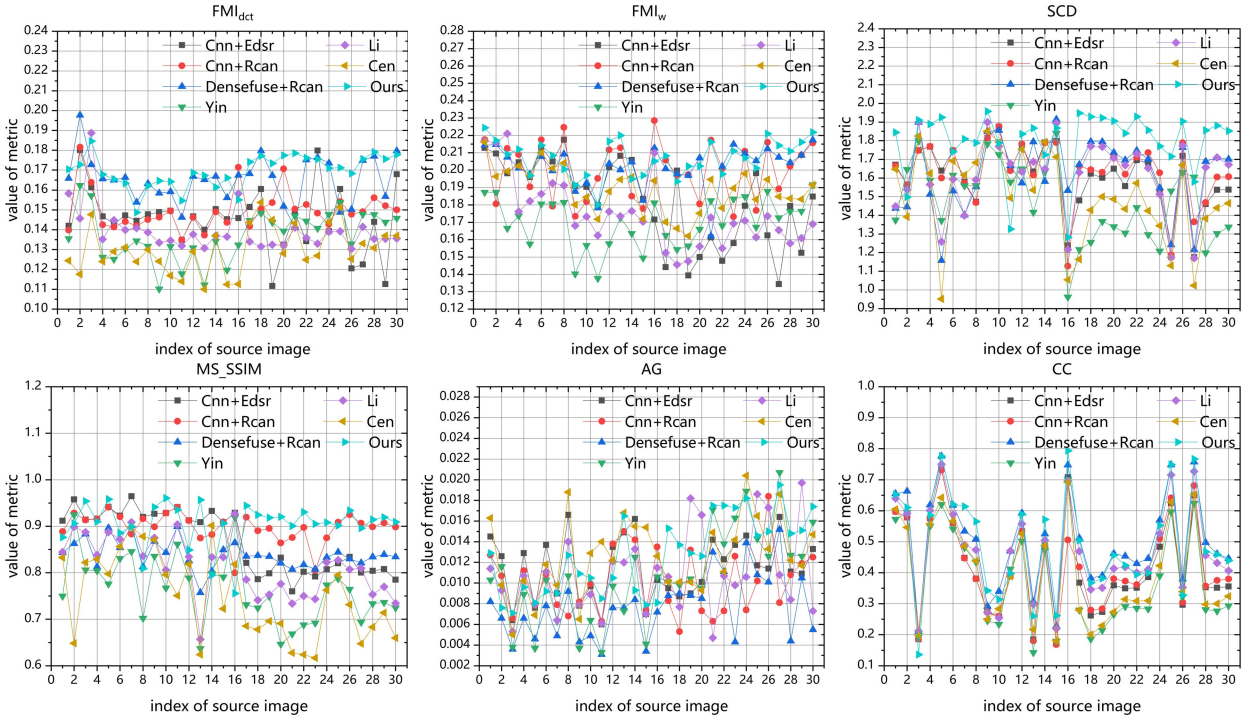

Fig. 10. Quantitative evaluation of the results of all the comparative methods on each pair of test images.

TABLE I

QUANTITATIVE EVALUATION RESULTS OF ALL COMPARISON METHODS. THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE MARKED IN BLUE

Model	Scale	FMI_{dct}	FMI_w	SCD	MS_SSIM	AG	CC
CNN+EDSR [14], [40]	2	0.1535	0.2011	1.7023	0.8903	0.0112	0.4243
CNN+RCAN [14], [32]	2	0.1458	0.1939	1.6972	0.8973	0.0099	0.4232
DenseFuse+RCAN [17], [32]	2	0.1609	0.1986	1.6225	0.8633	0.0073	0.4907
Yin [8]	2	0.1373	0.1707	1.4010	0.7573	0.0106	0.3856
Li [23]	2	0.1396	0.1750	1.5839	0.8150	0.0107	0.4715
Cen [30]	2	0.1303	0.1888	1.4961	0.7412	**0.0127**	0.4014
Ours	2	**0.1684**	**0.2016**	**1.7941**	**0.9060**	0.0121	**0.4932**
CNN+EDSR [14], [40]	3	0.1310	0.1716	1.6946	0.8668	0.0104	0.4231
CNN+RCAN [14], [32]	3	0.1298	0.1693	1.6908	0.8602	0.0093	0.4220
DenseFuse+RCAN [17], [32]	3	0.1279	0.1677	1.6128	0.8460	0.0061	0.4780
Yin [8]	3	0.1185	0.1615	1.3868	0.7254	0.0087	0.3819
Cen [30]	3	0.1146	0.1663	1.4610	0.7165	**0.0116**	0.3961
Ours	3	**0.1329**	**0.1735**	**1.7745**	**0.8713**	0.0115	**0.4894**
CNN+EDSR [14], [40]	4	0.1124	0.1582	1.6767	0.8434	0.0081	0.4197
CNN+RCAN [14], [32]	4	0.1105	0.1583	1.6773	0.8437	0.0081	0.4197
DenseFuse+RCAN [17], [32]	4	0.1006	0.1565	1.6037	0.8329	0.0053	0.4656
Yin [8]	4	0.1017	0.1557	1.3762	0.6960	0.0072	0.3791
Cen [30]	4	0.1076	0.1548	1.3937	0.6770	**0.0105**	0.3847
Ours	4	**0.1186**	**0.1617**	**1.7554**	**0.8548**	0.0101	**0.4840**

The fusion and super-resolution results obtained by Yin's method and Li's method contain many artifacts, and many salient objects are unclear, such as people, houses, roads, and branches. On the contrary, the result obtained by the proposed method has fewer artifacts and noises and contains more texture details. Compared with the fusion and super-resolution result obtained by Cen's method, the contour of the object in the result obtained by the proposed method is more apparent and more natural in visual appearance, as shown in the red box in Fig. 7. The fusion and super-resolution results obtained by CNN+EDSR and CNN+RCAN have almost no visual difference compared with the proposed method. However, in the red and blue boxes in Fig. 7, we can find that our method produces clearer results, implying the superiority of the proposed method. The proposed method retains more pixel intensity information than the fusion and super-resolution result obtained by DenseFuse+RCAN. Figs. 8 and 9 show that even in large-scale fusion and super-resolution, the proposed method can still retain the salient and detailed information in the source images. The proposed method performs better than other methods, indicating that the method has more significant advantages in large-scale fusion and super-resolution.

TABLE II
EFFECTIVENESS ANALYSIS OF THE DFAG AND CEA. THE OPTIMUM VALUES ARE BOLDED AND THE SECOND BEST RESULTS ARE MARKED IN BLUE

Model	Scale	FMI_{dct}	FMI_w	SCD	MS_SSIM	AG	CC
Baseline	2	0.1676	0.1986	1.7515	0.8839	0.0110	0.4668
Baseline+DFAG	2	**0.1713**	0.2009	1.7678	0.8841	0.0119	0.4873
Baseline+CEA	2	0.1696	0.2001	1.7532	0.8935	**0.0127**	0.4883
Baseline+DFAG+CEA	2	0.1684	**0.2016**	**1.7941**	**0.9060**	0.0121	**0.4932**
Baseline	3	0.1305	0.1728	1.7496	0.8655	0.0107	0.4617
Baseline+DFAG	3	0.1311	0.1730	1.7435	0.8761	0.0114	0.4843
Baseline+CEA	3	**0.1332**	0.1725	1.7407	**0.8794**	0.0113	0.4851
Baseline+DFAG+CEA	3	0.1329	**0.1735**	**1.7745**	0.8713	**0.0115**	**0.4894**
Baseline	4	0.1149	0.1592	1.7169	0.8474	0.0095	0.4589
Baseline+DFAG	4	**0.1192**	0.1606	1.7239	0.8467	0.0091	0.4803
Baseline+CEA	4	0.1177	0.1609	1.7191	0.8513	0.0093	0.4793
Baseline+DFAG+CEA	4	0.1186	**0.1617**	**1.7554**	**0.8548**	**0.0101**	**0.4840**

The average values of the six metric values of 30 fused and super-resolution images of various magnifications obtained by all methods are shown in Table I. The best values are highlighted in bold, and the second bests are highlighted in blue. The proposed method has five best averages and one second-best average value. It shows that our results are effective and can retain the structural information in the source image. This is mainly because the model is trained in multitask learning (fusion and super-resolution), making the network perform feature learning better and reducing error propagation in the network. Fig. 10 shows a more recognizable comparison of the objective performance of different fusion and super-resolution methods. In this article, we only show the 2× fusion and super-resolution results. The scores obtained by the same method on 30 source image pairs are plotted as a curve for each indicator FMI_{dct}, FMI_{w}, SCD, MS_SSIM, AG, and CC. It can be seen from Fig. 10 that the proposed method outperforms the others on the entire testing set in terms of stability and is convenient for practical application.

D. Ablation Experiment

In this section, the influences of the DFAG and CEA on the model are further verified. The model without those two modules is taken as the Baseline. The Baseline with DFAG is noted as Baseline+DFAG, the Baseline with CEA is noted as Baseline+CEA, and the Baseline with DFAG and CEA is noted as Baseline+DFAG+CEA. Table II shows the average value of the objective evaluation results of each model in the entire test set. The visual effects of Baseline, Baseline+DFAG, Baseline+CEA, and Baseline+DFAG+CEA are shown in Fig. 11.

1) Effectiveness of DFA: In order to prove the effectiveness of DFAG, the fusion and super-resolution results obtained by the Baseline model are compared with those of the Baseline+DFAG model. Table II shows that the Baseline+DFAG model has advantages in most evaluation indicators compared with the Baseline model. Among them, most evaluation indicators of Baseline+DFAG are higher than Baseline, which indicates that DFAG successfully improves the fusion effect of the model. DFAG assigns a higher weight to areas with high intensity to better retain the pixel intensity information in the infrared source image. As shown in Fig. 11, the brightness of the fusion result of Baseline+DFAG is higher than that of Baseline. The above results prove the effectiveness of DFAG in the fusion framework.

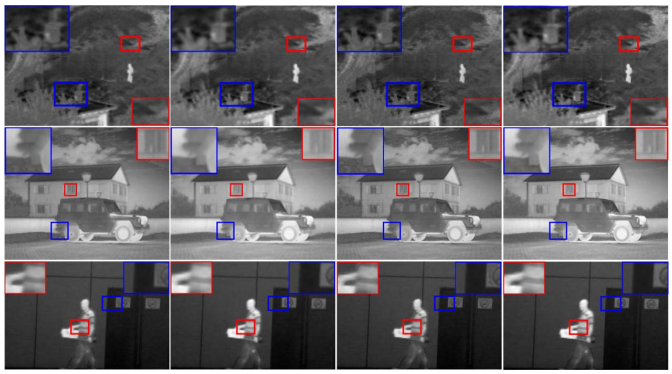

Fig. 11. Effectiveness validation of DFAG and CEA. From left to right: the results of "Baseline," the results of "Baseline+DFAG," the results of "Baseline+CEA," and "Baseline+DFAG+CEA."

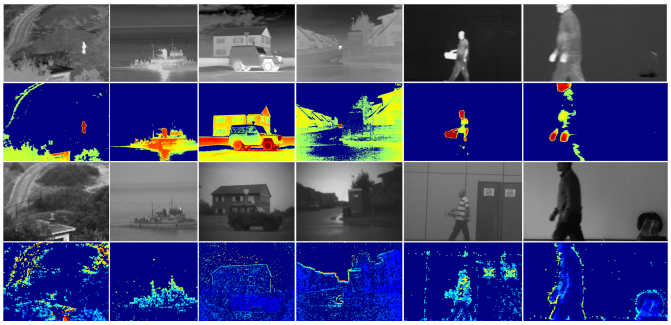

Fig. 12. Heat maps of the DFAG and CEA. The first row is the infrared images. The second row is the heat map of the DFAG. The third row is the visible images. The fourth row is the heat map of the CEA.

2) Effectiveness of CEA: As an important module in the fusion framework, CEA encodes context information into local features, thereby improving the representation ability of local features. It also detects the texture detail areas of visible images to enhance these areas' features further. The baseline model is compared with the Baseline+CEA model to verify

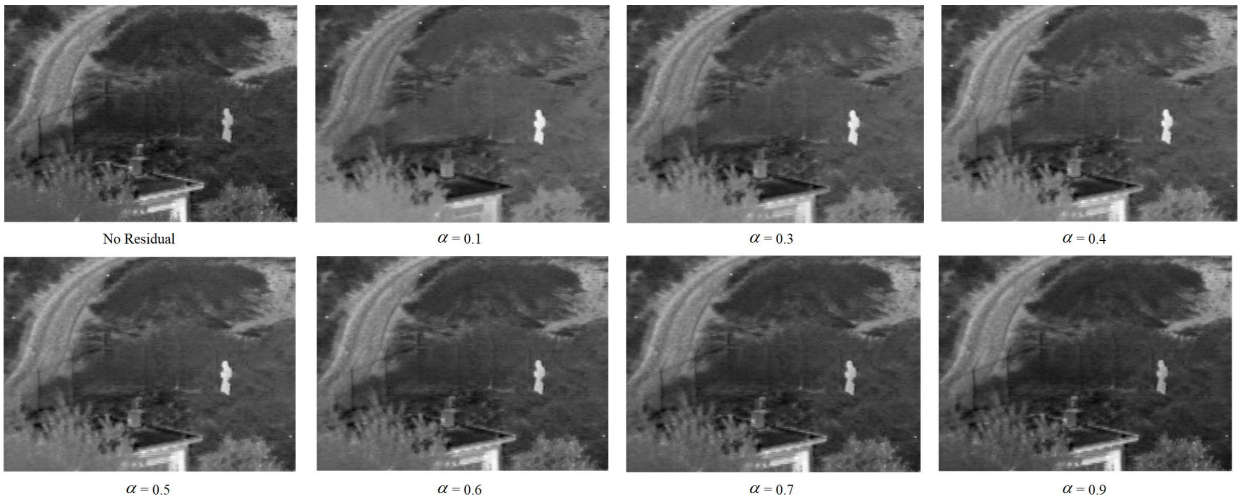

Fig. 13. Fusion results of different α values on the teacher network.

TABLE III

QUANTITATIVE COMPARISON OF FUSION RESULTS WITH DIFFERENT μ VALUES. BEST VALUES AND SECOND BEST VALUES ARE HIGHLIGHTED IN BOLD AND IN BLUE, RESPECTIVELY

μ	FMI_{dct}	FMI_w	SCD	MS_SSIM	AG	CC
1	0.3270	0.3569	1.8517	0.9304	0.0142	0.4982
10	0.3810	0.4027	1.8528	0.9321	**0.0142**	0.4983
100	0.3943	0.4112	1.8568	0.9331	0.0141	0.4984
1000	**0.4015**	**0.4195**	**1.8582**	**0.9335**	0.0141	**0.4984**

TABLE IV

QUANTITATIVE COMPARISON OF FUSION RESULTS WITH DIFFERENT γ VALUES. BEST VALUES AND SECOND BEST VALUES ARE HIGHLIGHTED IN BOLD AND IN BLUE, RESPECTIVELY

γ	FMI_{dct}	FMI_w	SCD	MS_SSIM	AG	CC
0.3	0.1185	**0.1621**	1.7532	0.8535	0.0101	0.4832
0.5	**0.1188**	0.1602	1.7578	0.8541	0.0105	0.4830
0.7	0.1186	0.1611	**1.7613**	**0.8549**	**0.0105**	**0.4847**
0.9	0.1186	0.1617	1.7560	0.8538	0.0103	0.4831

TABLE V

QUANTITATIVE COMPARISON OF FUSION RESULTS WITH DIFFERENT φ VALUES. BEST VALUES AND SECOND BEST VALUES ARE HIGHLIGHTED IN BOLD AND IN BLUE, RESPECTIVELY

φ	FMI_{dct}	FMI_w	SCD	MS_SSIM	AG	CC
0.1	0.1164	0.1607	1.7580	0.8536	0.0103	0.4830
0.5	0.1170	0.1608	1.7587	**0.8539**	0.0101	0.4832
1	**0.1179**	**0.1619**	**1.7638**	0.8531	0.0101	**0.4837**
5	0.1162	0.1597	1.7604	0.8424	**0.0103**	0.4834

the effectiveness of CEA. Table II shows that the average value of most fusion and super-resolution evaluation indicators increases when CEA is used. Among them, most evaluation indicators of Baseline+CEA are higher than Baseline. Moreover, as shown in Fig. 11, the edge of the Baseline+CEA fused result is more evident than that of the Baseline. It proves that CEA can improve the local feature representation capability and strengthen the characteristics of texture detail area.

Moreover, the DFAG and CEA heat maps are shown in Fig. 12. The infrared images, the heat map of the DFAG, the visible images, and the heat map of the CEA are in the first, second, third, and fourth rows, respectively. The red indicates that the area receives more attention in the heat map, and the blue indicates that the area receives less attention. As shown in Fig. 13, the regions with greater brightness in the infrared images and the edges in the visible images have attracted more attention. It further proves the effectiveness of DFAG and CEA.

E. Parameter Sensitivity Analysis

This section analyzes the sensitivities of α and μ in HKDnet. It is found that satisfactory performance can be achieved by selecting parameters in a particular range. In the experiment, we change the value of one parameter within a certain range with the other parameters fixed.

The fusion layer of the teacher network will inevitably cause information loss in the feature fusion of infrared and visible images. The subfigure in Fig. 13 with no residual is obtained when the effects of $\mathbf{F}_{\text{ir}}$ and $\mathbf{F}_{\text{vis}}$ are removed in (5). As shown in Fig. 13, the brightness information in the fused image with no residual label is significantly reduced. Therefore, $\mathbf{F}_{\text{ir}}$ and $\mathbf{F}_{\text{vis}}$ are directly into the decoder of the teacher network to enhance the brightness information of the fused image, and their input ratio is controlled by the parameter α. In order to test the influence of the α on the fusion result, the value of α is set to 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.9. The fusion results obtained with different values of α are shown in Fig. 13. It shows that with α increase, the pixel intensity in the fused image decreases, and the texture details increase. The value of α is set to 0.5 to balance the pixel intensity information and texture detail information in the fused image. μ is the weight coefficient of the loss function of the teacher network. Because the dataset used by the teacher network is different from that in [17], we analyze the influence of μ value on

the fusion performance of the teacher network. The value of μ is set to 1, 10, 100, and 1000. Since the change of μ does not significantly impact the visual effect of fusion results, the average values of six metrics of fusion results in the whole test set obtained with different values of μ are shown in Table III. Table III shows that when μ is 1000, the teacher network performs better, so μ is set to 1000.

γ is a threshold between 0 and 1 used for the corner attention module in CEA. We set it to 0.3, 0.5, 0.7, and 0.9 in the experiment to select the appropriate value of γ. The average values of quality metrics with the different values of γ are listed in Table IV. It can be seen from Table IV that when γ is set to 0.7, the proposed network has the optimal performance, so we set γ to 0.7. φ is the balance parameter in the total loss function of the student network. In order to analyze the influence of different φ values on the network performance, we set φ to 0.1, 0.5, 1, and 5. When φ takes different values, the average values of quality metrics are shown in Table V. It can be seen that when $\varphi = 1$, most of the quality metrics values have the highest value, so we set φ to 1.

V. Conclusion

A heterogeneous knowledge distillation with multilayer attention embedding is proposed to jointly implement the fusion and super-resolution of infrared and visible images. The proposed method generates a pseudo-label through the teacher network to supervise the training of the student network. The student network consists of the fusion part and the super-resolution part. The input of the fusion part is low-resolution infrared and visible images. The two source images extract features through two branches. Then, they get the fusion feature through a fusion layer. Finally, the low-resolution fused feature is obtained by using a decoder. The super-resolution part based on the RCAN takes the features obtained from the fusion part as input and generates a high-resolution fused image. The experimental results show that the proposed method is optimal in infrared–visible image fusion and super-resolution.

References

[1] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

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

[3] H. F. Li, X. G. He, Z. T. Yu, and J. B. Luo, "Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior," *Inf. Sci.*, vol. 523, pp. 14–37, Jun. 2020.

[4] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, 2018, Art. no. 1850018.

[5] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[6] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[7] 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.

[8] H. Yin, S. Li, and L. Fang, "Simultaneous image fusion and super-resolution using sparse representation," *Inf. Fusion*, vol. 14, no. 3, pp. 229–240, 2013.

[9] Z. Zhu, M. Zheng, G. Qi, D. Wang, and Y. Xiang, "A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain," *IEEE Access*, vol. 7, pp. 20811–20824, 2019.

[10] Y. Yang, "Multimodal medical image fusion through a new DWT based technique," in *Proc. 4th Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2010, pp. 1–4.

[11] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, and D. Tao, "Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1082–1102, Apr. 2020.

[12] H. Li, X. He, D. Tao, Y. Tang, and R. Wang, "Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning," *Pattern Recognit.*, vol. 79, pp. 130–146, Jul. 2018.

[13] Y. Zhang, M. Yang, N. Li, and Z. Yu, "Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion," *Signal Process.*, vol. 167, Feb. 2020, Art. no. 107327.

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

[15] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710.

[16] 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. 48, pp. 11–26, Aug. 2019.

[17] 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.

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

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

[20] F. Lahoud and S. Süsstrunk, "Fast and efficient zero-learning image fusion," 2019, *arXiv:1905.03590*.

[21] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4724–4732.

[22] M. Iqbal and J. Chen, "Unification of image fusion and super-resolution using jointly trained dictionaries and local information contents," *IET Image Process.*, vol. 6, no. 9, pp. 1299–1310, Dec. 2012.

[23] H. Li, Z. Yu, and C. Mao, "Fractional differential and variational method for image fusion and super-resolution," *Neurocomputing*, vol. 171, pp. 138–148, Jan. 2016.

[24] M. Xie, Z. Zhou, and Y. Zhang, "Joint framework for image fusion and super-resolution via multicomponent analysis and residual compensation," *IEEE Access*, vol. 7, pp. 174092–174107, 2019.

[25] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. Interspeech*, Sep. 2014, pp. 1910–1914.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 38–39, 2015.

[27] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[28] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7130–7138.

[29] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[30] H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 4070–4083, 2021.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[32] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 294–310.
[33] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7982–7991.
[34] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[35] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–151.
[36] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.
[37] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[38] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[40] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[42] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE 8th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2014, pp. 1–3.
[43] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, 2015.
[44] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
[45] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Opt. Commun.*, vol. 341, pp. 199–209, Apr. 2015.
[46] M. Deshmukh and U. Bhosale, "Image fusion and image quality assessment of fused images," *Int. J. Image Process.*, vol. 4, no. 5, pp. 484–508, 2010.

Wanxin Xiao received the B.E. degree in computer science and technology from Wuchang Shouyi University, Wuhan, China, in 2019. He is currently pursuing the M.E. degree in computer technology with the Kunming University of Science and Technology, Kunming, China.

His research interests include image processing and computer vision.

Yafei Zhang received the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2008.

She is currently an Associate Professor with the College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her main research interests include image processing and pattern recognition.

Hongbin Wang was born in 1983. He received the Ph.D. degree in computer science from Jilin University, Changchun, China, in 2013.

He is currently an Associate Professor with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include intelligent information systems, natural language processing, and computer vision.

Fan Li received the Ph.D. degree from the School of Information Science and Technology, Beijing Forestry University, Beijing, China, in 2015.

He is currently an Associate Professor with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His research interests include image identification and identification of insect species.

Hua Jin received the M.D. and Ph.D. degrees from Kunming Medical University, Kunming, China, in 2004 and 2011, respectively.

She is currently a Professor with the Kunming University of Science and Technology, Kunming. She has authored or coauthored more than 50 scientific articles. Over the past years, her research interests include clinical big data processing, computer vision, and perioperative analgesia.