Different Input Resolutions and Arbitrary Output Resolution: A Meta Learning-Based Deep Framework for Infrared and Visible Image Fusion

Huafeng Li, Yueliang Cen, Yu Liu, *Member, IEEE*, Xun Chen, *Senior Member, IEEE*, and Zhengtao Yu

***Abstract*— Infrared and visible image fusion has gained ever-increasing attention in recent years due to its great significance in a variety of vision-based applications. However, existing fusion methods suffer from some limitations in terms of the spatial resolutions of both input source images and output fused image, which prevents their practical usage to a great extent. In this paper, we propose a meta learning-based deep framework for the fusion of infrared and visible images. Unlike most existing methods, the proposed framework can accept the source images of different resolutions and generate the fused image of arbitrary resolution just with a single learned model. In the proposed framework, the features of each source image are first extracted by a convolutional network and upscaled by a meta-upscale module with an arbitrary appropriate factor according to practical requirements. Then, a dual attention mechanism-based feature fusion module is developed to combine features from different source images. Finally, a residual compensation module, which can be iteratively adopted in the proposed framework, is designed to enhance the capability of our method in detail extraction. In addition, the loss function is formulated in a multi-task learning manner via simultaneous fusion and super-resolution, aiming to improve the effect of feature learning. And, a new contrast loss inspired by a perceptual contrast enhancement approach is proposed to further improve the contrast of the fused image. Extensive experiments on widely-used fusion datasets demonstrate the effectiveness and superiority of the proposed method. The code of the proposed method is publicly available at https://github.com/yuliu316316/MetaLearning-Fusion.**

Manuscript received January 28, 2021; accepted March 22, 2021. Date of publication April 2, 2021; date of current version April 8, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61701160, Grant 61966021, Grant 61562053, and Grant 61922075; in part by the Fundamental Research Funds for the Central Universities under Grant JZ2020HGPA0111; and in part by the National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Huafeng Li and Yueliang Cen contributed equally to this work.) (Corresponding author: Yu Liu.)*

Huafeng Li, Yueliang Cen, and Zhengtao Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China, and also with the Yunnan Provincial Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China (e-mail: lhfchina99@kust.edu.cn; ceny11996@163.com; ztyu@hotmail.com).

Yu Liu is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: yuliu@hfut.edu.cn).

Xun Chen is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China (e-mail: xunchen@ustc.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2021.3069339, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3069339

***Index Terms*— Image fusion, convolutional neural network, infrared image, meta learning, super-resolution.**

I. Introduction

DUE to the different spectrums adopted, infrared and visible sensors have different imaging characteristics. Infrared sensors describe the thermal radiation of a scene and they are good at capturing thermal targets even under poor lighting conditions, but infrared imagery is often not preferred by the human visual system due to the loss of most spatial details. On the contrary, visible sensors can capture abundant spatial details but tend to lose their effectiveness under poor lighting conditions such as low illumination and thick fog. Infrared and visible image fusion aims to combine the advantage of these two imaging sensors by generating a composite image, which can provide a more comprehensive description of the scene. Many vision-based applications including objection detection [1], face recognition [2], video surveillance [3] have benefited from this image fusion technique.

In recent years, although great progress has been achieved in the field of infrared and visible image fusion with the arising of various fusion methods [4]–[6], several considerable limitations still exist in the current study. First, almost all the existing fusion methods require the input source images to be of the same spatial resolution. However, in real-world scenarios, images obtained by different categories of imaging sensors usually have different spatial resolutions. Second, due to the restrictions of the factors like cost and power, the source images captured by infrared or visible sensors may sometimes suffer from low spatial resolution, whereas fused images with high resolution are mostly expected.

In the above two situations, image super-resolution technique can be applied to help the fusion issue. A routine way is to perform super-resolution prior to fusion, but this two-phase separating manner has obvious drawbacks. To address this issue, a few methods that simultaneously conduct fusion and super-resolution using an integrated model have been proposed [7]–[9]. However, the spatial resolutions of different source images still need to be the same in these methods, i.e., the first problem is not solved. In addition, these methods can only improve the resolution of the fused image by a few integer scale factors (e.g., $\times 2$, $\times 3$, $\times 4$) and the super-resolution models of different scale factors

are learned separately (i.e., a specific model is trained for each scale factor), which greatly reduces their usefulness in practical applications. Thus, there still exist two urgent problems to be solved: 1) the source images need to have the same spatial resolution and 2) only a few integer scale factors are available for super-resolution.

To solve the above problems, this paper proposes a meta learning-based deep framework for the fusion of infrared and visible images. The most distinctive feature of the proposed method is that it can simultaneously tackle the source images (the source images are assumed to be spatially aligned, which is a general assumption in the study of image fusion) of different resolutions and generate the fused image of arbitrary resolution just with a single learned model. This is mainly achieved by adopting a meta-upscale module [10], which can dynamically predict the weights of the upscale filters by taking the scale factor as input. Furthermore, the proposed method can obtain the super-resolution results of two source images as by-products at the same time. Fig. 1 shows the framework of the proposed meta learning-based infrared and visible image fusion method. It is mainly composed of two feature extraction modules (FEMs) that is designed to extract features from the source images, one fusion module (FM) that aims to fuse the salient features, and a series of residual compensation modules (RCMs) to make up for the loss of details. In a FEM, features of each source image are first extracted by a convolutional network and then upscaled by a meta-upscale module (MUM) with an appropriate factor (the value can be arbitrary) which is set according to the requirement of a specific fusion problem. The upscaled features are merged by the FM via a dual attention mechanism to generate the fused features. Then RCMs further extract and compensate the loss information during the up-sampling process of feature maps, and it can be iteratively employed for several times in our fusion framework. It can be seen from Fig. 1 that this framework can simultaneously achieve fusion and super-resolution in a multi-task learning manner (i.e., one fusion branch and two super-resolution branches), leading to more powerful ability in feature learning, which is helpful to promote the quality of fusion results. Extensive experiments demonstrate the effectiveness and superiority of the proposed method. The main contributions of this work are summarized into the following four folds.

- We propose a meta learning-based deep framework for infrared and visible image fusion. Unlike most existing image fusion methods, the proposed framework can accept the source images of different resolutions and generate the fused image of arbitrary resolution just with a single learned model, which has great significance in practical usage.
- We develop a dual attention mechanism-based feature fusion module, in which position attention and channel attention are simultaneously taken into account to fuse features from different source images.
- We present a residual compensation module that can be iteratively adopted in the proposed fusion framework to enhance the capability of our method in detail extraction.
- We formulate the loss function in a multi-task learning manner via simultaneous fusion and super-resolution, which is helpful to learn more effective features and improve the quality of final fusion results. In addition, a new contrast loss based on the theory of perceptual color correction [11] is proposed in this work.

The rest of this paper is organized as follows. Section II briefly reviews some related works. Section III describes the details of the proposed method. Experimental validations are provided in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Infrared and Visible Image Fusion

In the past decades, a large number of infrared and visible image fusion methods have been proposed [4], [5], while most of them are based on multi-scale transform [12]–[14] and sparse representation [15]–[18]. In the last few years, the deep learning-based study has been gaining in popularity in the field of image fusion because of its impressive performance in extracting crucial information from the source images [19]–[22]. Liu *et al.* [20] first introduced a convolutional neural network (CNN)-based approach to fuse multi-focus images. Since then, deep learning-based study has emerged as an active branch in various image fusion issues including infrared and visible image fusion. Liu *et al.* [23] extended the classification CNN model presented in [20] to infrared and visible image fusion by adopting a Laplacian pyramid-based fusion framework. Li *et al.* [22] developed a dense connection-based convolutional architecture named DenseFuse for infrared and visible image fusion. Lahoud *et al.* [24] proposed to decompose the source images into a base layer and a detail layer. The base layers are fused under the guidance of a visual saliency map, while the detail layers are merged based on the weight map generated by a CNN model. Zhang *et al.* [25] proposed an end-to-end image fusion architecture consisting of feature extraction, fusion and reconstruction stages. Jian *et al.* [26] presented an infrared and visible image fusion method based on a symmetric encoder and decoder architecture via residual blocks. Jung *et al.* [27] proposed an unsupervised deep image fusion method by applying the structure tensor representations, which measure the difference between the fused image and the source images in the gradient domain, to design a loss function. Ma *et al.* [28] first introduced the generative adversarial network (GAN) into the field of image fusion for combining infrared and visible images. In their method (known as FusionGAN), a generator network is used to create the fused image from source images, while a discriminator network is adopted to further extract spatial details from the visible image.

All the methods mentioned above require that the images to be fused have the same resolution. Ma *et al.* [29] recently proposed a multi-resolution infrared and visible image fusion method based on a dual-discriminator GAN model, while it makes an assumption that the resolution of the visible image is 4×4 times that of the infrared image. A straightforward approach to tackle the above problem is to unify the resolution of the source images before fusion with a super-resolution approach. However, this two-phase separating manner has

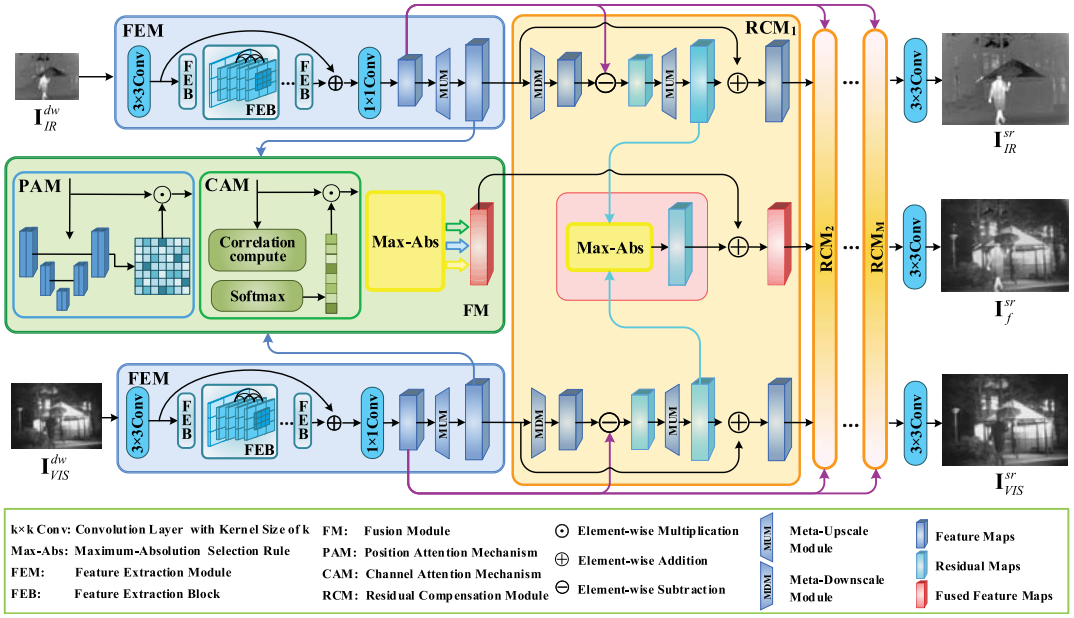


Fig. 1. The framework of the proposed meta learning-based infrared and visible image fusion method. It is mainly composed of two feature extraction modules (FEMs), one fusion module (FM) and a series of residual compensation modules (RCMs). The FEM is used to extract features for MUM and FM, and it consists of multiple FEBs. The FM is used to merge the upscaled features constructed by MUM. The RCMs is used to extract and compensate the information lost in the up-sampling process of feature maps.

its inherent drawbacks, such as repeated feature extraction procedures in super-resolution and fusion tasks may cause more computational cost and the quality of fusion results rely heavily on the effect of the super-resolution method adopted. In practical applications, a unified processing framework is always preferred. Moreover, in most existing fusion methods, the spatial resolution of the output fused image remains the same as that of the source images, which arises another restriction in practice when the fused images in higher resolution are required.

B. Simultaneous Image Fusion and Super-Resolution

To obtain the fused images with higher spatial resolution, some simultaneous image fusion and super-resolution approaches have been proposed in the literature [7]–[9], [30]. Yin *et al.* [7] developed a sparse representation-based approach for simultaneous image fusion and super-resolution by assuming that the high-frequency components of the upscaled low-resolution source images and the reconstructed result share the same sparse coding coefficients. Based on low-rank sparse representation theory, Xie *et al.* [9] proposed a residual compensation method for simultaneous image fusion and super-resolution. Iqbal *et al.* [30] proposed to learn a set of multi-scale dictionaries from high-resolution images to construct the high-resolution fused results. Li *et al.* [8] presented a variational and fractional differential model for image fusion and super-resolution, according to the fact that the fused image should contain the basic geometry of the source images.

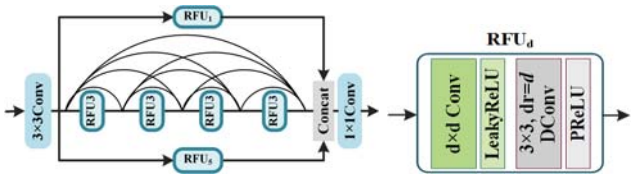
However, all of these methods still require the input source images to be of the same resolution and can only increase the resolution of the fused image by a few integer scale factors. In addition, the applied models of different scale factors are learned separately, namely, the model should be re-trained when the scale factor changes. These shortcomings greatly limit the potential of these methods used in real-world scenarios.

C. Meta Leaning-Based Super-Resolution

Meta learning is one of the new technologies in few-shot and zero-shot learning, which aims to let neural networks learn to learn, and it can predict the weights of filters dynamically. Benefiting from this advantage, Jo *et al.* [31] introduced a completely different framework for super-resolution of video, in which an end-to-end deep neural network was constructed based on the meta-learning for generating dynamic up-sampling filters and residual image. To delete the explicit motion compensation, these filters and residual image were produced locally and dynamically, and with these produced up-sampling filters, the HR frame was constructed. In single image super-resolution, Hu *et al.* [10] proposed a meta-learning method to predict the weights of filters to reconstruct a super-resolution image. Compared with the traditional super-resolution methods such as [32] and [33], this approach can construct a super-resolution image with arbitrary scale factor. Soh *et al.* [34] presented a meta-transfer learning method for zero-shot super-resolution. This method can exploit both external and internal information, and only need one single gradient update.

Inspired by the meta-learning and these works on super-resolution, we develop a novel deep learning framework for infrared-visible image fusion. The meta-upscale module introduced in [10] is adopted to achieve the target that the fusion framework can accept the source images of different

(a) The architecture of feature extraction network



(b) The architecture of FEB (c) The architecture of RFU_d

Fig. 2. The architecture of feature extraction network. RFU_d denotes the receptive field unit that contains a $d \times d$ convolution and a 3×3 dilated convolution with a dilation rate (dr) of d.

resolutions and generate the fused image of arbitrary resolution just with a single trained model. To the best of our knowledge, this is the first work that attempts to achieve the above target in this field.

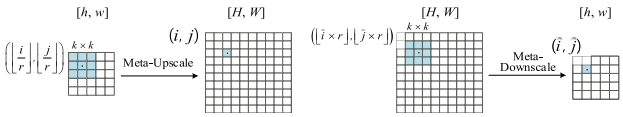
III. Methodology

A. Feature Extraction Module

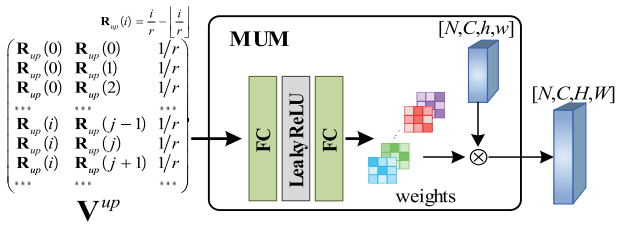
The feature extraction module (FEM) consists of a feature extraction network and a meta-upscale module. The feature extraction network is used to extract features from each source image, and the meta-upscale module is applied to upscale the obtained feature maps to the target resolution of the fused image.

1) Feature Extraction Network: As shown in Fig. 2(a), our feature extraction network contains a 3×3 convolutional layer, a series of feature extraction blocks (FEBs) (see Fig. 2(b)) and a 1×1 convolutional layer. Each FEB is composed of three branches, which are designed based on the receptive field units (RFUs) (see Fig. 2(c)) with different receptive fields. Inspired by the Receptive Field Block (RFB) in [35], RFU_d $(d = 1, 3, 5)$ contains a $d \times d$ convolution and a 3×3 dilated convolution with a dilation rate of d, where the dilated convolution is exploited to simulate the impact of the eccentricities of pRFs in the human visual cortex as in [35]. The values of d in the first branch and the third branch are 1 and 5, respectively. The second branch consists of four RFU_3 and the dense skip connection manner [36] is adopted to further improve the capacity of feature extraction. The outputs of these three branches are concatenated and a 1×1 convolution is finally employed as a bottleneck layer. To avoid the loss of information caused by pooling and strided convolution, the down-sampling layer (e.g., pooling, strided convolution) is not involved in our FEB, so that the spatial size of the output feature map remains the same as the input.

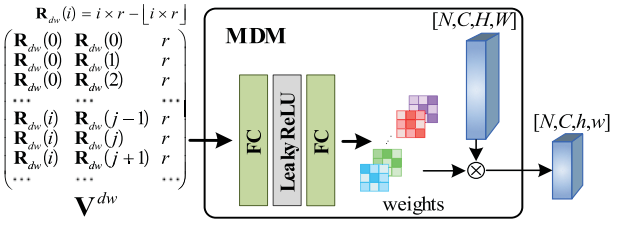
In our feature extraction network, the first 3×3 convolutional layer and the last 1×1 convolutional layer contain 64 and 8 filters, respectively. The last 1×1 convolutional layer in each FEB contains 64 filters. Each of the rest



(a) Location projection of LR to HR(left) and HR to LR(right).



(b) The structure of the meta-upscale module.



(c) The structure of the meta-downscale module.

Fig. 3. The structures of the meta-learning-based upscale and downscale modules. The symbol ⊗ represents the matrix multiplication operation.

convolutional layers contains 32 filters. Moreover, to avoid zero gradients, each convolutional layer is followed by a LeakyReLU layer with a negative scope of 0.2 because it does not introduce extra parameters that need to be learned, while each dilated convolutional layer is followed by a PReLU layer for more specialized activations [37].

2) Meta-Upscale: To achieve the goal that the method can accept the source images of different resolutions and generate the fused image of arbitrary resolution just with a single model, the meta-upscale module (MUM) presented in the meta-learning super-resolution work [10] is adopted as the up-sampling layer. Specifically, the MUM consists of three steps including position correspondence, weight prediction and feature mapping. The positional correspondence of pixels between a low-resolution image and its corresponding high-resolution version with a scale factor r is shown in Fig. 3 (a). Given a LR image $\mathbf{I}^{dw}$ of resolution $h \times w$, we use MUM to obtain its corresponding HR version $\mathbf{I}^{up}$ of resolution $\lfloor rh \times rw \rfloor$, $(rh = H, rw = W)$. The position (i, j) in $\mathbf{I}^{up}$ is corresponding to $\left(\left\lfloor \frac{i}{r} \right\rfloor, \left\lfloor \frac{j}{r} \right\rfloor\right)$ in $\mathbf{I}^{dw}$, where $r \geq 1$ denotes the scale factor and $\lfloor\rfloor$ the floor function. The batchsize and number of channel are denoted as N and C, respectively. As shown in Fig. 3 (b), the position matrix $\mathbf{V}^{up}$ constructed by coordinate vector $\mathbf{v}_{i,j} = \left(\frac{i}{r} - \left\lfloor \frac{i}{r} \right\rfloor, \frac{j}{r} - \left\lfloor \frac{j}{r} \right\rfloor, \frac{1}{r}\right)$ is fed to a simple fully-connected network to predict the weights of the upscale filters. When the scale factor is changed, the weights can be accordingly adjusted independent to the previously extracted features. Thus, the MUM can arbitrarily increase the resolution of feature maps without repeatedly training the whole model.

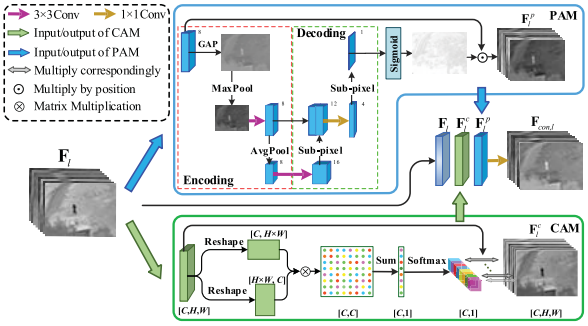


Fig. 4. The architecture of the presented dual attention mechanism.

B. Fusion Module

1) Dual Attention Mechanism: In deep learning-based image fusion methods, the fusion of feature maps plays an important role in improving the quality of the fused image. Conventional rules for feature fusion mainly include maximum selection [26], weighted averaging [22], [26] and feature concatenation [27], [29], [38]. However, in most methods, these rules are generally performed on the original features without considering the correlation and difference among the features at different positions and different channels. The method in [26] made a good attempt to alleviate this problem by applying a channel attention mechanism, but it ignores thecorrelation of features at different spatial locations. In addition, the channel attention in [26] is simply designed as a pixel-wise softmax operation over all input channels. In this work, a dual attention mechanism (DAM) is developed to address the above issue, as shown in Fig. 4. It consists of two components: position attention mechanism (PAM) and channel attention mechanism (CAM).

a) Position attention mechanism: For the extracted feature maps, it will be beneficial to the preservation of the salient source information if more attention is assigned to the coefficients corresponding to the salient objects or contours. To this end, we develop a U-Net [39] like PAM architecture to predict the weights for the features at different spatial positions, as shown in Fig. 4. The PAM architecture mainly contains two stages: encoding and decoding.

In the encoding stage, given the upscaled feature maps $\mathbf{F}_l \in \mathbb{R}^{C\times H\times W}$ $(l \in \{IR, VIS\})$, where 'IR' and 'VIS' denote the infrared and visible images, respectively. C, H and W represent the number of channels, the height and the width of the feature maps, respectively. Global Average Pooling (GAP) is employed to aggregate them and generate $\tilde{\mathbf{F}}_l \in \mathbb{R}^{H\times W}$. In the first down-sampling level, we use the max-pooling operation to preserve the most significant information in one neighborhood, while reduce the resolution of $\tilde{\mathbf{F}}_l$ to $1 \times \frac{1}{2}H \times \frac{1}{2}W$. Then a convolutional layer is used to further extract features. The above process can be formulated as:

$$\hat{\mathbf{F}}_l = \mathrm{Conv}\left(\mathrm{MP}\left(\mathrm{GAP}\left(\mathbf{F}_l\right)\right), k = 3\right), \tag{1}$$

where Conv, MP and GAP denote the convolution, max-pooling and GAP operators, respectively. k is the size of the convolution kernel. The resolution of $\hat{\mathbf{F}}_l$ is $8 \times \frac{1}{2}H \times \frac{1}{2}W$. In the second down-sampling level, we use the average pooling operation to down-sample the feature maps $\hat{\mathbf{F}}_l$, and the produced results are then fed to another convolution layer to achieve the feature maps $\mathbf{F}_l^e$ of size $16 \times \frac{1}{4}H \times \frac{1}{4}W$.

In the decoding stage, a sub-pixel convolutional layer is used to up-sample the feature maps $\mathbf{F}_l^e$ to $4 \times \frac{1}{2}H \times \frac{1}{2}W$. The obtained results and $\hat{\mathbf{F}}_l$ are concatenated, and then fed to a 1×1 convolutional layer. Finally, a sub-pixel convolutional layer is adopted to produce the position weight map $\mathbf{F}_l^d \in \mathbb{R}^{1\times H\times W}$ of the source image l. The above process can be formulated as

$$\mathbf{F}_l^d = \mathrm{SPC}(\mathrm{Conv}([\mathbf{F}_l^e, \mathrm{SPC}(\mathbf{F}_l^e)], k = 1)), \tag{2}$$

where SPC represents the sub-pixel convolution operator. With $\mathbf{F}_{l,i}^d$, we calculate the weighted feature maps $\mathbf{F}_{l,i}^p$ by

$$\mathbf{F}_{l,i}^p(m,n) = \frac{\mathbf{F}_{l,i}(m,n)}{1 + \exp(-\mathbf{F}_l^d(m,n))}, \tag{3}$$

where (m, n) denotes the position of a coefficient in the feature map and $i \in \{1, 2, \ldots, C\}$ denotes the channel index. The sigmoid function $1/(1+\exp(-\mathbf{F}_l^d(m,n))$ is used to normalize the value of $\mathbf{F}_l^d$ to $[0, 1]$.

b) Channel attention mechanism: In infrared and visible image fusion, each feature map (i.e., channel) can be regarded as the response of significant targets in the source images. For a target, its responses in different channels are always different and should be associated with each other. To emphasize the target, it is better to assign a larger weight to the feature map with stronger response. Based on this consideration, we develop a novel channel attention mechanism (CAM) to generate the weight of each channel by exploiting the interdependence among different channels.

The architecture of the CAM is also shown in Fig. 4. Let $\mathbf{F}_l \in \mathbb{R}^{C\times H\times W}$ $(l \in \{IR, VIS\})$ be the input feature maps. Here, we first reshape $\mathbf{F}_l$ to $\mathbf{F}_{r,l} \in \mathbb{R}^{C\times M}$, where $M = H \times W$. Then, the matrix multiplication operation is performed on different channels to model the interdependence of features maps. Specifically, the channel relation vector $\mathbf{s}_l \in \mathbb{R}^{C\times 1}$ of $\mathbf{F}_l$ is obtained by

$$\mathbf{s}_l = \mathbf{F}_{r,l}\mathbf{F}_{r,l}^T\mathbf{1}, \tag{4}$$

where T denotes matrix transpose and $\mathbf{1} \in \mathbb{R}^{C\times 1}$ as an all-one vector. Finally, the aggregation weight of the i-th channel is calculated as the pixel-wise softmax function:

$$\omega_{l,i} = \frac{\exp(\mathbf{s}_l(i,1))}{\sum_{i=1}^{C}\exp\left(\mathbf{s}_l(i,1)\right)}. \tag{5}$$

Thus, the aggregated feature maps $\mathbf{F}_l^c$ can be formulated as

$$\mathbf{F}_l^c = [\omega_{l,1}\mathbf{F}_{l,1}, \omega_{l,2}\mathbf{F}_{l,2}, \cdots, \omega_{l,C}\mathbf{F}_{l,C}]. \tag{6}$$

2) Fusion Strategy: Fig. 5 shows the structure of our fusion module (FM). Let $\mathbf{F}_{IR}$ and $\mathbf{F}_{VIS}$ denoted the upscaled infrared and visible feature maps, respectively. For each source image, its feature maps $\mathbf{F}_l$ $(l \in \{IR, VIS\})$ is processed by the PAM and CAM to obtain the position weighted feature maps $\mathbf{F}_l^p$

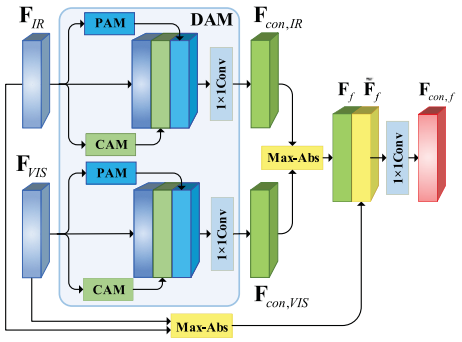


Fig. 5. The structure of the presented feature fusion module.

and the channel weighted feature maps $\mathbf{F}_l^c$. Then, the $\mathbf{F}_l$, $\mathbf{F}_l^c$ and $\mathbf{F}_l^p$ are concatenated and fed to a 1×1 convolutional layer to obtain the DAM-based features

$$\mathbf{F}_{con,l} = \text{Conv}([\mathbf{F}_l, \mathbf{F}_l^c, \mathbf{F}_l^p], k = 1). \tag{7}$$

Next, the obtained features $\mathbf{F}_{con,IR}$ and $\mathbf{F}_{con,VIS}$ are fused by the maximum selection rule

$$\mathbf{F}_f(i,j) = \begin{cases} \mathbf{F}_{con,IR}(i,j), & |\mathbf{F}_{con,IR}(i,j)| \geq |\mathbf{F}_{con,VIS}(i,j)| \\ \mathbf{F}_{con,VIS}(i,j), & |\mathbf{F}_{con,IR}(i,j)| < |\mathbf{F}_{con,VIS}(i,j)|, \end{cases} \tag{8}$$

and the original source features $\mathbf{F}_{IR}$ and $\mathbf{F}_{VIS}$ are also fused by the maximum selection rule

$$\tilde{\mathbf{F}}_f(i,j) = \begin{cases} \mathbf{F}_{IR}(i,j), & |\mathbf{F}_{IR}(i,j)| \geq |\mathbf{F}_{VIS}(i,j)| \\ \mathbf{F}_{VIS}(i,j), & |\mathbf{F}_{IR}(i,j)| < |\mathbf{F}_{VIS}(i,j)|, \end{cases} \tag{9}$$

where (i, j) indicates the feature coordinate. Finally, the above two fused features are further concatenated and pass a 1×1 convolutional layer to generate the output fused features

$$\mathbf{F}_{con,f} = \text{Conv}([\mathbf{F}_f, \tilde{\mathbf{F}}_f], k = 1). \tag{10}$$

C. Residual Compensation Module

In our fusion and super-resolution framework, we need to increase the size of each feature map to the target size via the MUM. However, this process may cause the loss of fine details in the source images. Inspired by back-projection networks [40], we develop a simple yet effective residual compensation mechanism to make up for the lost details. To this end, we create a meta-downscale module (MDM) by imitating the MUM, as shown in Fig. 3(c). Just like the MUM, the MDM also dynamically predicts the weights of the downscale filters by taking the scale factor as input, thus it can arbitrarily reduce the size of feature maps without repeated training. The MUM and MDM are jointly applied to design the residual compensation module (RCM). Specifically, let $\mathbf{F}_l^{dw}$ ($l \in \{IR, VIS\}$) be the original low-resolution feature maps extracted by the feature extraction network, and $\mathbf{F}_l$ be its meta-upscaled version. The residual between $\mathbf{F}_l^{dw}$ and the meta-downscaled result of $\mathbf{F}_l$ is computed as

$$\mathbf{R}_l^{dw} = \mathbf{F}_l^{dw} - \text{MDM}(\mathbf{F}_l). \tag{11}$$

The high-resolution of the residual $\mathbf{R}_l^{dw}$ is obtained with the MUM as

$$\mathbf{R}_l^{up} = \text{MUM}(\mathbf{R}_l^{dw}). \tag{12}$$

The residual map $\mathbf{R}_{IR}^{up}$ ($\mathbf{R}_{VIS}^{up}$) is composed of the fine-grained features and the distorted information produced in MUM and MDM. In SR, the smooth regions and partial salient structures can be easily recovered by the SR algorithms. Therefore, the residuals on the first few layers are mainly high-frequency components that have not been recovered. In this case, the residual coefficient with larger absolute value at each position generally indicates that more information of image details is lost or distorted in the up-sampling. Therefore, to better compensate such information for the fused image, the $\mathbf{R}_{IR}^{up}$ and $\mathbf{R}_{VIS}^{up}$ are merged with the maximum selection rule:

$$\mathbf{R}_f^{up}(i,j) = \begin{cases} \mathbf{R}_{IR}^{up}(i,j), & |\mathbf{R}_{IR}^{up}(i,j)| \geq |\mathbf{R}_{VIS}^{up}(i,j)| \\ \mathbf{R}_{VIS}^{up}(i,j), & |\mathbf{R}_{IR}^{up}(i,j)| < |\mathbf{R}_{VIS}^{up}(i,j)|. \end{cases} \tag{13}$$

The fused feature maps $\mathbf{F}_{con,f}$ are refined by adding $\mathbf{R}_f^{up}$ as

$$\mathbf{F}_{f,r} = \mathbf{F}_{con,f} + \mathbf{R}_f^{up}. \tag{14}$$

Meanwhile, at each super-resolution branch, the feature maps $\mathbf{F}_l$ are also refined by its corresponding residual maps $\mathbf{R}_l^{up}$ as

$$\mathbf{F}_{l,r} = \mathbf{F}_l + \mathbf{R}_l^{up}. \tag{15}$$

It is noted that the RCM can be iteratively used in the proposed framework, as shown in Fig. 1.

D. Loss Function Formulation

In this work, the loss function is designed in a multi-task learning manner via simultaneous fusion and super-resolution to pursue better capacity in feature learning. The Wald's protocol [41] is adopted, namely, the original infrared and visible images are used as the ground truth while their low-resolution versions are used as the input for model training. It is noted that the image fusion task is lacking ground truth data as reference. Therefore, the super-resolution task is likely to be helpful to improve the performance of the fusion task via supervised learning. Specifically, the loss functions consist of a pixel loss L_{pixel} and a contrast loss $L_{contrast}$.

1) Pixel Loss: The pixel loss aims to restrict the intensity difference between the ground truth and the model prediction. Let $\mathbf{I}_{IR}^{up}$ and $\mathbf{I}_{VIS}^{up}$ denote the ground truth infrared and visible images, respectively. Let $\mathbf{I}_{IR}^{sr}$, $\mathbf{I}_{VIS}^{sr}$ and $\mathbf{I}_f^{sr}$ denote the predicted infrared, visible and fused images in high resolution, respectively. The pixel loss L_{pixel} is defined as

$$\begin{aligned} L_{pixel} = &\|\mathbf{I}_{IR}^{up} - \mathbf{I}_{IR}^{sr}\|_1 + \|\mathbf{I}_{VIS}^{up} - I_{VIS}^{sr}\|_1 \\ &+ \|\mathbf{I}_{IR}^{up} - \mathbf{I}_f^{sr}\|_1 + \|\mathbf{I}_{VIS}^{up} - \mathbf{I}_f^{sr}\|_1, \end{aligned} \tag{16}$$

where $\|\cdot\|_1$ denotes the l_1-norm. Similar to [10], the L_{pixel} is defined with the l_1-norm instead of the l_2-norm to achieve better capability in preserving salient information from the source images.

TABLE I

EXPERIMENTAL SETTINGS. THE FIRST COLUMN (ROW) SHOWS THE DOWN-SAMPLING RATIO OF THE INFRARED (VISIBLE) SOURCE IMAGE

Infrared \ Visible	0.5	0.8	1.0
0.5	Setting #1	Setting #2	Setting #3
0.8	Setting #4	Setting #5	Setting #6
1.0	Setting #7	Setting #8	Setting #9

2) Contrast Loss: Inspired by the perceptual contrast enhancement approach presented in [11], [42], we introduce a contrast loss $L_{contrast}$ to further strengthen the salient features and enhance the contrast of the fused image. The $L_{contrast}$ is defined as

$$L_{contrast} = \mathrm{D}\left(\mathbf{I}_f^{sr}\right) - \mathrm{C}\left(\mathbf{I}_f^{sr}\right), \tag{17}$$

where the first term $\mathrm{D}(\mathbf{I}_f^{sr})$ is used to prevent a large deviation between the fused image and the average of two source images, while the second term $\mathrm{C}(\mathbf{I}_f^{sr})$ is used to improve the contrast of the fused image. Specifically, the $\mathrm{D}(\mathbf{I}_f^{sr})$ is defined as

$$\mathrm{D}(\mathbf{I}_f^{sr}) = \|\mathbf{I}_f^{sr} - \bar{\mathbf{I}}^{up}\|_F^2, \tag{18}$$

where $\bar{\mathbf{I}}^{up}$ is the average of $\mathbf{I}_{IR}^{up}$ and $\mathbf{I}_{VIS}^{up}$. $\mathrm{C}(\mathbf{I}_f^{sr})$ accounts for the change in pixel lightness. Theoretically, the effect of this change on contrast should be inversely proportional to the distance between the two pixels. In addition, the contrast of the fused image should have a positive correlation to that of the source images. Based on the above two considerations, the $\mathrm{C}(\mathbf{I}_f^{sr})$ is defined as

$$\mathrm{C}(\mathbf{I}_f^{sr}) = \sum_{(i,j)\in H\times W} \sum_{(i',j')\in H\times W} \frac{|\mathbf{I}_f^{sr}(i,j) - \mathbf{I}_f^{sr}(i',j')|}{d_{pos}^{i,j,i',j'} \times d_{int}^{i,j,i',j'}}, \tag{19}$$

where $i \neq i'$, $j \neq j'$, and $H \times W$ denotes the resolution of $\mathbf{I}_f^{sr}$. $d_{pos}^{i,j,i',j'}$ is as a function related to the distance between (i, j) and (i', j'), while $d_{int}^{i,j,i',j'}$ is a function describing intensity difference between (i, j) and (i', j') in the source images. In this work, we define $d_{pos}^{i,j,i',j'}$ as

$$d_{pos}^{i,j,i',j'} = \sqrt{(i-i')^2 + (j-j')^2}, \tag{20}$$

and $d_{int}^{i,j,i',j'}$ as

$$d_{int}^{i,j,i',j'} = 1 - \tanh\left(\frac{\sum_{l=IR,VIS} |\mathbf{I}_l^{up}(i,j) - \mathbf{I}_l^{up}(i',j')|}{2}\right). \tag{21}$$

In Eq.(21), we employ the tanh function to calculate $d_{int}^{i,j,i',j'}$ as it has a larger derivative than other similar functions like sigmoid when performing back propagation, which can speed up the convergence of the network. By this means, the overall contrast in the fused image can be increased by minimizing the contrast loss.

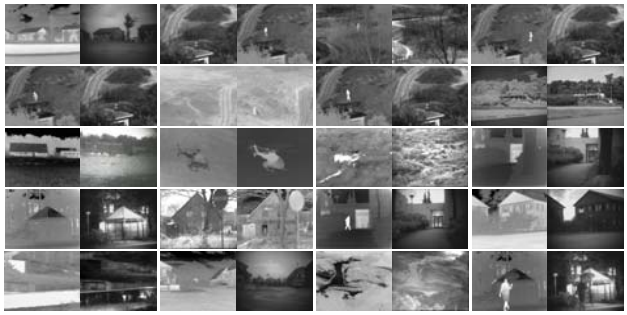


Fig. 6. The testing set which contains 20 pairs of widely-used infrared and visible images from the TNO dataset.

TABLE II

ANALYSIS OF THE EFFECTIVENESS OF DIFFERENT FUNCTIONAL MODULES. THE BOLD AND BLUE FONTS INDICATE THE OPTIMAL AND SUBOPTIMAL VALUES, RESPECTIVELY

Methods	Q_{TE}	Q_Y	Q_{SF}	Q_{MI}	Q_{STD}	Q_{AG}
Baseline	0.3582	0.4855	−0.4245	0.3492	31.3423	3.3916
Baseline+DAM	0.3811	0.5186	−0.4127	0.3569	27.0161	3.5124
Baseline+DAM+RCM	0.3644	0.5560	−0.2252	0.3448	37.4310	4.4822
Baseline+DAM+2RCMs	**0.3911**	**0.6558**	**−0.0931**	**0.3864**	**39.9261**	**4.9691**
Baseline+DAM+3RCMs	0.3712	0.5018	−0.4037	0.3431	31.7554	3.3204

TABLE III

OBJECTIVE ASSESSMENT RESULTS OF OUR MODEL TRAINED WITH DIFFERENT VALUES OF λ. THE BOLD AND BLUE FONTS INDICATE THE OPTIMAL AND SUBOPTIMAL VALUES, RESPECTIVELY

λ	Q_{TE}	Q_Y	Q_{SF}	Q_{MI}	Q_{STD}	Q_{AG}
0	0.3619	0.5173	-0.3895	0.3358	31.4023	3.3983
0.0001	0.3288	0.4801	-0.3302	0.2455	33.2666	4.1282
0.001	0.3644	0.5505	-0.2267	0.3432	38.9912	4.4124
0.005	**0.3911**	**0.6558**	**-0.0931**	**0.3864**	**39.9261**	**4.9691**
0.001	0.3653	0.5233	-0.1468	0.3306	39.8672	4.7532
0.1	0.3661	0.4524	0.2878	0.3184	38.5969	3.7498

E. Training Strategy

A two-phase training strategy is applied to train the proposed model. The model is first trained only with the L_{pixel} to obtain the initial features for super-resolution and fusion. Then, we freeze the parameters of the super-resolution branches, while just fine-tune the parameters of the fusion branch with the following total loss

$$L_{total} = L_{pixel} + \lambda L_{contrast}, \tag{22}$$

where the weighting parameter λ is used to balance the contributions of L_{pixel} and $L_{contrast}$.

IV. EXPERIMENTS

A. Dataset and Training Details

In deep learning-based infrared and visible image fusion, the KAIST[1] and the FLIR[2] are two commonly-used datasets for model training. In the KAIST, there are 95000 infrared and visible image pairs, while the FLIR contains 14452 pairs of infrared and visible images. In our experiments, we randomly select 6200 pairs from each dataset to totally obtain 12400 pairs of infrared and visible images, in which

[1]https://soonminhwang.github.io/rgbt-ped-detection/
[2]https://www.flir.ca/oem/adas/adas-dataset-form/

Fig. 7. Effectiveness validation of DAM and RCM. From left to right: the results of "Baseline", the results of "Baseline + DAM", the results of "Baseline + DAM + RCM", the results of "Baseline + DAM + 2RCMs", and the results of "Baseline + DAM + 3RCMs".

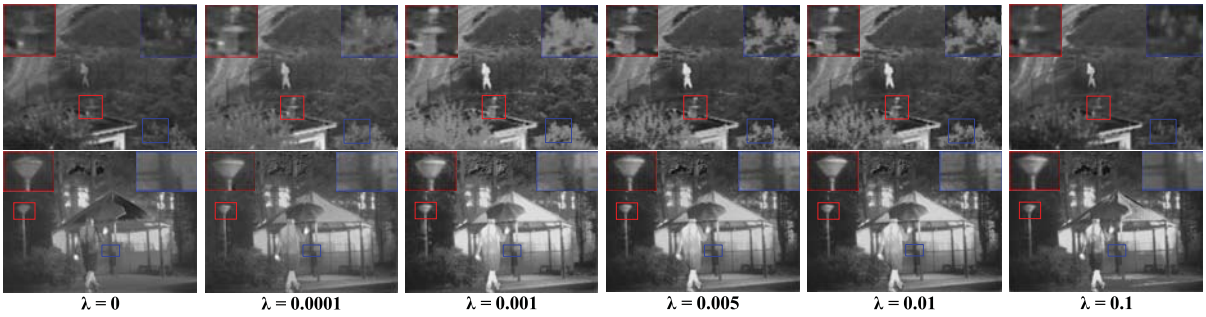


Fig. 8. Fusion results of the proposed method using different values of λ in the contrast loss.

12000 pairs are used as for training and the remaining 400 pairs are used for validation. We convert the RGB visible image into grayscale image by the 'cvtColor' function from the OpenCV library. Moreover, we select 20 pairs of widely-used infrared and visible images from the TNO dataset[3] to construct the testing set, as shown in Fig. 6. In the training stage, we randomly crop each infrared and visible image pair to size 128×128 as the high-resolution source images, and then down-sample them by bilinear interpolation to generate the input low-resolution images. Specifically, for each training batch, two scale factors are randomly and independently selected for down-sampling the infrared and visible images, respectively. Once the model is trained, it has the capacity to handle an arbitrary scale factor, so that the goal of different input resolutions and arbitrary output resolution can be realized.

In our experiments, the training process is completed after 23 epochs, in which the first 20 epochs are trained by the pixel loss L_{pixel} and the last 3 epoches are trained by the total loss L_{total}. We set the learning rate to 1×10^{-4} for all layers during the training. The parameter λ is set to 0.005. The Adam optimizer [43] is used for training and the batchsize is set to 4. The numbers of FEBs and RCMs used in our method are set to 6 and 2, respectively. We implement our model with the PyTorch framework and conduct all the experiments on a computer equipped with a NVIDIA GTX 2080Ti GPU.

B. Fusion Evaluation Metrics

Six popular image fusion metrics are applied to evaluate the fusion performance objectively, which include the Tsallis entropy Q_{TE} [44], Yang's metric Q_Y [45], the spatial frequency (SF)-based metric Q_{SF} [46], the normalized mutual information Q_{MI} [47], the standard deviation Q_{STD} [28] and the average gradient Q_{AG} [48]. Specifically, Q_{TE} uses the Shannon entropy to calculate the mutual information between the fused image and the source images. Q_Y is a widely-used image structural similarity (SSIM)-based fusion metric. Q_{SF} can evaluate the degree of spectrum richness in the spatial domain of the fusion result. Q_{MI} measures the dependence between fused image and the source images. Q_{STD} is used to assess the intensity difference of pixels in the fused image. Q_{AG} can measure the amount of fine details contained in the fusion result. Except for Q_{SF} that a value closer to 0 indicates a better fusion performance, all the other metrics with larger values indicate better results.

[3]https://figshare.com/articles/TNO Image Fusion Dataset/1008029

C. Experiment Settings

To demonstrate the effectiveness of the proposed method, 9 settings which denote the different combinations of source image resolutions are adopted in our experiments, as listed in Table I. We fix the output resolution while change the input resolution. Specifically, the target resolution of the fused image is the same as that of the original high-resolution image, while the low-resolution source images for fusion are obtained by bilinear interpolation using different scale factors. Taking 'Setting #2' as an example, it represents that the resolution of infrared and visible images to be fused is 0.5 and 0.8 times that of the original image, respectively. Considering that the proportional relationship between the input and output resolutions is relative, the above strategy that fixing the output resolution while changing the input resolution is reasonable.

Fig. 9. Fusion results of different methods on the "Camp1828" example with nine settings. From top to bottom: the results on the tasks of Settings 1-9. From left to right: the results generated by DeepFuse, DenseFuse, FEZ, FusionGAN, GTF and our method, respectively.

D. Ablation Study

DAM and RCM are two important components in our fusion framework. In this subsection, we conduct an ablation study to investigate the impacts of these two modules. It is easy to find that either the DAM or the RCM can be flexibly removed from our fusion framework. Therefore, we denote the model without DAM and RCM as "Baseline", the model only with DAM as "Baseline + DAM", the model with both DAM and RCM as "Baseline + DAM + RCM" and the model with t ($t \geq 2$) RCMs as "Baseline + DAM + tRCMs", respectively. For simplicity, the Setting #2 described in Table I is adopted as an example in this experiment. The average objective evaluation results of different models on the whole testing set are given in Table II, and the fusion results on two examples are shown in Fig. 7.

1) Effectiveness of DAM: To highlight the importance of information in different channels and different locations of the feature maps, we introduced the DAM in the model. To demonstrate the effect of DAM, we compare the fusion results obtained by the "Baseline" model with those obtained by the "Baseline + DAM" model. It can be found from Fig. 7 that the "Baseline + DAM" model has obvious advantage on extracting spatial details from the source images as well as obtaining higher contrast in the fused image. Moreover,

Fig. 10. Fusion results of different methods on the "Kaptein1654" example with nine settings. From top to bottom: the results on the tasks of Settings 1-9. From left to right: the results generated by DeepFuse, DenseFuse, FEZ, FusionGAN, GTF and our method, respectively.

as shown in Table II, the utilization of DAM also achieves clear improvement on objective evaluation results. The above results validate the effectiveness of the DAM in our fusion framework.

2) Effectiveness of RCM: In our model, the RCM is used to compensate the information lost in the up-sampling process of feature maps. To demonstrate the effect of the RCM, we first compare the "Baseline + DAM + RCM" model with the "Baseline + DAM" model. It can be seen from Table II that the evaluation scores on most fusion metrics increase when the RCM is adopted, and more spatial details are preserved according to Fig. 7. When the number of RCMs increases to 2, the performances on both detail extraction and contrast preservation are further improved. In addition, the "Baseline + DAM + 2RCMs" model obtains the best objective performance on all the six fusion metrics. However, when three RCMs are used, the perceptual quality of the fused result has degraded, that is because the information in the third residual map that can improve the quality of fused result is reduced, and its positive effect on improving quality is drowned by the negative effect of external information introduced by MDM. Therefore, the number of RCMs is set to 2 by default in our experiments. More experimental analysis is given in the supplementary materials.

TABLE IV

OBJECTIVE EVALUATION RESULTS OF DIFFERENT IMAGE FUSION METHODS. THE BOLD AND BLUE FONTS INDICATE THE OPTIMAL AND SUBOPTIMAL VALUES, RESPECTIVELY. THE TWO INTEGERS WITHIN THE BRACKET AFTER EACH SCORE REPRESENT THE NUMBER OF IMAGE PAIRS ON WHICH THE CORRESPONDING METHOD GETS THE FIRST AND SECOND PLACES AMONG ALL THE METHODS, RESPECTIVELY

Tasks	Methods	Q_{TE}	Q_Y	Q_{SF}	Q_{MI}	Q_{STD}	Q_{AG}
Setting #1	DeepFuse [38]	0.3315(0,0)	0.5928(0,9)	-0.3076(2,4)	0.2810(0,0)	35.0512(1,4)	3.8547(1,5)
	DenseFuse [22]	0.3310(0,1)	0.5933(7,5)	-0.3078(4,11)	0.2815(0,0)	34.7151(0,5)	3.8375(2,12)
	FEZ [24]	0.3790(5,6)	0.5559(0,2)	-0.5268(0,0)	0.2858(0,1)	29.1426(0,2)	2.7173(0,0)
	FusionGAN [28]	0.3759(6,3)	0.4178(0,0)	-0.5040(1,0)	0.3447(5,4)	30.9491(0,3)	2.6537(0,0)
	GTF [49]	0.3454(2,3)	0.5589(2,2)	-0.4312(1,4)	0.4011(6,10)	35.7034(5,3)	3.0395(0,2)
	Ours	**0.3971(7,7)**	**0.6305(11,2)**	**0.0540(12,1)**	**0.4176(9,5)**	**43.4690(14,3)**	**5.2933(17,1)**
Setting #2	DeepFuse [38]	0.3294(0,0)	0.6157(1,11)	-0.3254(2,3)	0.2831(0,0)	34.7229(0,6)	3.8461(1,5)
	DenseFuse [22]	0.3319(0,1)	0.6187(6,2)	-0.3131(2,7)	0.2837(0,0)	34.6086(4,3)	3.9155(1,10)
	FEZ [24]	0.3777(5,6)	0.5724(0,0)	-0.5392(0,0)	0.2862(0,1)	29.1043(0,2)	2.7027(0,0)
	FusionGAN [28]	0.3756(6,3)	0.4469(0,0)	-0.5092(0,2)	0.3461(6,3)	30.8987(1,2)	2.6669(0,2)
	GTF [49]	0.3504(2,3)	0.5960(3,4)	-0.4208(1,7)	**0.3981(8,8)**	35.6850(6,2)	3.2261(0,2)
	Ours	**0.3911(7,7)**	**0.6558(10,3)**	**-0.0931(15,1)**	0.3864(6,8)	**39.9261(9,5)**	**4.9691(18,1)**
Setting #3	DeepFuse [38]	0.3348(0,0)	0.6796(3,6)	-0.2767(2,4)	0.2905(0,1)	34.6387(3,7)	4.2368(4,4)
	DenseFuse [22]	0.3292(0,1)	0.6785(3,5)	-0.2780(2,7)	0.2913(0,2)	34.1575(2,4)	4.1935(4,10)
	FEZ [24]	0.3782(1,2)	0.6379(1,0)	-0.5028(0,0)	0.2934(8,3)	29.1276(0,2)	2.9954(0,0)
	FusionGAN [28]	0.3731(6,5)	0.5107(0,0)	-0.4666(1,1)	0.3427(6,3)	30.3352(1,1)	3.0053(0,0)
	GTF [49]	0.3251(6,3)	0.6708(7,2)	-0.3541(1,7)	0.3430(1,3)	35.4734(8,1)	3.8213(1,5)
	Ours	**0.3878(7,9)**	**0.6840(6,7)**	**-0.1743(14,1)**	**0.3877(5,8)**	**37.4284(6,5)**	**4.4983(11,1)**
Setting #4	DeepFuse [38]	0.3349(0,0)	0.6207(0,9)	-0.3149(2,2)	0.2828(0,0)	34.9880(0,6)	3.9267(1,4)
	DenseFuse [22]	0.3337(0,1)	0.6214(8,6)	-0.3123(1,13)	0.2833(0,0)	34.4903(1,3)	3.9064(2,13)
	FEZ [24]	**0.3912(6,5)**	0.5860(1,0)	-0.5356(0,0)	0.2888(0,1)	29.1160(0,2)	2.7086(0,0)
	FusionGAN [28]	0.3751(6,2)	0.4160(0,0)	-0.5048(0,1)	0.3425(4,6)	31.0144(1,2)	2.6418(0,0)
	GTF [49]	0.3531(3,3)	0.5613(2,4)	-0.4482(0,3)	**0.4150(10,7)**	35.6483(6,2)	2.9980(0,2)
	Ours	0.3910(5,9)	**0.6262(9,1)**	**-0.0946(17,1)**	0.3746(6,6)	**41.0574(12,5)**	**4.9693(17,1)**
Setting #5	DeepFuse [38]	0.3301(0,1)	0.6328(2,4)	-0.3154(1,3)	0.2846(0,0)	34.6336(1,7)	4.0224(1,3)
	DenseFuse [22]	0.3295(0,0)	0.6361(4,10)	-0.3071(2,8)	0.2850(0,0)	34.3472(0,2)	4.0279(3,10)
	FEZ [24]	0.3759(4,5)	0.5909(0,1)	-0.5313(0,0)	0.2873(0,1)	29.1024(0,2)	2.8230(0,0)
	FusionGAN [28]	0.3741(6,2)	0.4474(0,0)	-0.4971(1,1)	0.3424(1,3)	30.8444(0,3)	2.7513(0,1)
	GTF [49]	0.3549(1,3)	0.6140(2,4)	-0.4098(1,7)	0.4076(5,10)	35.6810(5,3)	3.4102(2,4)
	Ours	**0.4179(9,9)**	**0.6918(12,1)**	**0.0046(15,1)**	**0.4515(14,6)**	**41.9220(14,3)**	**5.0051(14,2)**
Setting #6	DeepFuse [38]	0.3351(0,1)	0.6945(0,7)	-0.2726(0,3)	0.3127(0,0)	34.6634(5,4)	4.3382(1,8)
	DenseFuse [22]	0.3305(0,1)	0.6941(6,2)	-0.2715(2,8)	0.2931(0,1)	34.0073(0,7)	4.2920(6,5)
	FEZ [24]	0.3766(8,1)	0.6538(0,1)	-0.4992(0,0)	0.2945(1,2)	29.1219(0,2)	3.0537(0,0)
	FusionGAN [28]	0.3729(4,5)	0.5106(0,0)	-0.4576(2,1)	0.3400(1,8)	30.4308(1,1)	3.0392(0,0)
	GTF [49]	0.3363(2,2)	0.6893(3,8)	-0.3296(3,6)	0.3495(2,6)	35.5288(7,1)	4.0454(5,3)
	Ours	**0.4088(6,10)**	**0.7181(11,2)**	**0.1232(13,2)**	**0.4396(16,3)**	**39.0097(7,5)**	**4.6531(8,4)**
Setting #7	DeepFuse [38]	0.3383(0,1)	0.6555(1,9)	-0.2756(0,5)	0.2877(0,0)	34.7384(3,4)	4.2440(2,9)
	DenseFuse [22]	0.3371(0,1)	**0.6578(7,7)**	-0.2708(2,10)	0.2884(0,0)	34.2293(2,7)	4.2212(8,8)
	FEZ [24]	0.3912(8,4)	0.6251(2,2)	-0.4975(0,0)	0.2889(0,2)	29.1682(0,2)	2.9947(0,0)
	FusionGAN [28]	0.3769(5,3)	0.4220(0,0)	-0.4990(2,1)	0.3456(5,7)	30.9154(1,2)	2.6826(0,0)
	GTF [49]	0.3428(2,2)	0.5648(2,2)	-0.4276(1,4)	**0.3739(9,4)**	35.4754(7,1)	3.1374(0,3)
	Ours	**0.3987(5,9)**	0.6367(8,0)	**-0.1976(15,0)**	0.3632(6,7)	**37.4504(7,4)**	**4.4374(10,0)**
Setting #8	DeepFuse [38]	0.3324(1,1)	0.6664(2,6)	-0.2828(1,2)	0.2905(0,2)	34.4853(0,6)	4.2729(1,9)
	DenseFuse [22]	0.3322(0,1)	0.6682(4,5)	-0.2796(2,7)	0.2896(0,0)	34.1389(1,4)	4.2347(7,5)
	FEZ [24]	0.3803(6,2)	0.5563(0,2)	-0.5131(0,0)	0.2586(0,2)	28.7631(0,2)	2.9507(0,0)
	FusionGAN [28]	0.3761(4,4)	0.4536(0,0)	-0.4982(2,1)	0.3465(1,4)	30.8818(2,1)	2.7324(0,0)
	GTF [49]	0.3738(4,3)	0.6860(8,0)	-0.3722(3,5)	**0.4449(11,2)**	37.2937(6,2)	3.6775(4,2)
	Ours	**0.4086(5,9)**	**0.7016(6,7)**	**-0.1030(12,5)**	0.4202(8,10)	**39.8319(11,5)**	**4.7524(8,4)**
Setting #9	DeepFuse [38]	0.3392(0,0)	0.7144(1,4)	-0.2275(1,4)	0.3104(0,1)	34.5201(2,7)	**4.6983(1,14)**
	DenseFuse [22]	0.3357(0,1)	0.7152(3,6)	-0.2253(2,10)	0.3130(1,0)	33.8374(1,3)	4.6424(10,1)
	FEZ [24]	0.3878(2,9)	0.6844(0,2)	-0.4654(0,0)	0.2967(0,3)	29.1823(0,2)	3.3221(0,0)
	FusionGAN [28]	0.3713(3,3)	0.5300(0,0)	-0.4403(2,1)	0.3436(0,7)	30.3984(1,1)	3.1609(0,0)
	GTF [49]	0.3361(1,3)	0.7021(4,5)	-0.2913(4,4)	0.3569(0,9)	35.6049(7,1)	4.3379(5,3)
	Ours	**0.4334(14,4)**	**0.7574(12,3)**	**-0.0817(11,1)**	**0.4933(19,0)**	**39.2519(9,6)**	4.6566(4,2)

E. Impact of the Contrast Loss

In our method, the contribution of the contrast loss $L_{contrast}$ is controlled by the weighting parameter λ. By changing λ (i.e., 0, 0.0001, 0.001, 0.005, 0.01, 0.1), the impact of the contrast loss can be observed. To characterize this impact, Fig. 8 shows the fusion results of the proposed method on two testing examples using different values of λ (under Setting #2 as well). It can be seen that the some fused regions suffer from low contrast when λ is set to a very small value (e.g., 0.0001). The contrast can be improved by increasing λ, but the value can not be too large (e.g., 0.1) as it may lead to the loss of important details in the fused image. We experimentally find that a setting of $\lambda \in [0.001, 0.01]$ can generally obtain satisfactory visual effects on both detail extraction and contrast preservation. Table III lists the average objective evaluation

TABLE V

OBJECTIVE ASSESSMENT RESULTS OF SIMULTANEOUS IMAGE FUSION AND SUPER-RESOLUTION METHODS. THE BOLD AND BLUE FONTS INDICATE THE OPTIMAL AND SUBOPTIMAL VALUES, RESPECTIVELY. THE TWO INTEGERS WITHIN THE BRACKET AFTER EACH SCORE REPRESENT THE NUMBER OF IMAGE PAIRS ON WHICH THE CORRESPONDING METHOD GETS THE FIRST AND SECOND PLACES AMONG ALL THE METHODS, RESPECTIVELY

Methods	Q_{TE}	Q_Y	Q_{SF}	Q_{MI}	Q_{STD}	Q_{AG}
Li's [8]	0.3636(3,6)	0.5351(0,2)	-0.4001(0,7)	0.2998(0,1)	29.8778(0,1)	3.2103(0,0)
Yin's [7]	0.3950(7,9)	0.6102(8,12)	-0.2959(8,11)	**0.4249(11,9)**	35.3479(3,16)	3.8693(2,18)
Ours	**0.3971(10,5)**	**0.6305(12,6)**	**0.0540(12,2)**	0.4176(9,10)	**43.4690(17,3)**	**5.2933(18,2)**

results of the proposed method using different values of λ on the whole testing set. It can be seen that the setting of $\lambda = 0.005$ obtains the best performance on all the six metrics. Based on the above results, we adopt $\lambda = 0.005$ as the default setting in our experiments.

F. Comparison With State-of-the-Art Methods

To demonstrate the effectiveness of the proposed method, 9 settings described in Table I are used as the task settings. We compare our method with five state-of-the-art fusion approaches including DeepFuse [38], DenseFuse [22], FEZ [24], FusionGAN [28] and GTF [49]. The first four are all recently proposed deep learning-based methods. However, these methods cannot be directly used to fuse low-resolution source images of arbitrary resolutions. To address this problem, we first adopt the Meta-SR [10] approach (for fair comparison as we use the meta-upscale module in our fusion framework) to improve the resolution of each source image to the target resolution and then use the compared fusion methods to merge them. In the experiments, the results of all the compared methods are generated by the corresponding source codes published by their respective authors. In addition, for the sake of fair comparison, all the deep learning-based methods adopt their pretrained models released by their corresponding authors to obtain the fusion results.

Fig. 9 and Fig. 10 provide two examples (denoted as Camp1828 and Kaptein1654, respectively) of the fusion results obtained by different methods. For better comparison, two close-ups are given in each example. It can be seen that our method obtains the most competitive performance among all the methods in terms of visual quality when considering the factors like detail extraction and contrast preservation. In contrast, other methods suffer from some undesirable visual effects such as loss of spatial details or low contrast. It is clear that the proposed method has advantage over other methods in most cases (i.e., settings and metrics).

Table IV lists the objective evaluation results of different fusion methods on 9 settings. The average scores over all the 20 source image pairs are reported. For each setting and each metric, the best result and the second best result among all the methods are indicated in bold and blue, respectively. The two integers in the bracket after each score represent the number of image pairs on which the corresponding method gets the first and second places among all the methods, respectively. It can be seen from Table IV that the proposed method owns clear advantages over other methods in all the nine settings and on all the six metrics. More experimental analysis can be seen in supplementary materials.

Fig. 11. The fusion results of different simultaneous image fusion and super-resolution methods. From left to right: the results of Li's method, the results of Yin's method, and the results of our method.

G. Comparison With Simultaneous Fusion and Super-Resolution Methods

As mentioned in Section II, there exist a few methods that can simultaneously conduct image fusion and super-resolution with integer scale factors. Our method is also competent for this task. In this subsection, we compare our method with two representative methods of this category: the variational and fractional differential model-based method [8] (Li's method) and the sparse representation-based method [7] (Yin's method). Considering that these two methods can only upscale the source images with an integer scale factor, the factor is set to 2 for all the methods in this experiment. Fig. 11 shows the fusion results of all the three methods. It can be clearly seen that our method can obtain better visual quality in terms of both detail extraction and contrast preservation. The objective assessment results of these three methods are listed in Table V. The proposed method obtains the best performance on five metrics.

V. CONCLUSION

In this paper, a novel meta learning-based deep framework for infrared and visible image fusion is proposed. By adopting a meta-upscale module, our fusion framework can tackle the source images of different resolutions and generate the fused image of arbitrary resolution just with a single learned model, which is distinctively different from existing methods. For the design of network architecture, the novelty mainly includes two aspects: 1) a dual attention mechanism-based feature fusion module is presented to address the position attention

and channel attention simultaneously, and 2) a residual compensation module that can be iteratively used in the fusion framework is developed to improve the detail extraction ability. In addition, the loss function is formulated in a multi-task learning manner via simultaneous fusion and super-resolution, which can help to learn more effective features. We also design a new contrast loss to further improve the fusion quality. A series of qualitative and quantitative experiments are conducted to verify the effectiveness of our method. The results indicate that our method can provide the state-of-the-art performance for arbitrary-resolution infrared and visible image fusion, leading to high potential in real-world applications.

REFERENCES

[1] S. Gao, Y. Cheng, and Y. Zhao, "Method of visual and infrared fusion for moving object detection," *Opt. Lett.*, vol. 38, pp. 1981–1983, Jun. 2013.

[2] R. Singh, M. Vatsa, and A. Noore, "Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition," *Pattern Recognit.*, vol. 41, no. 3, pp. 880–893, Mar. 2008.

[3] Q. Zhang, Y. Wang, M. D. Levine, X. Yuan, and L. Wang, "Multisensor video fusion based on higher order singular value decomposition," *Inf. Fusion*, vol. 24, pp. 54–71, Jul. 2015.

[4] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[5] X. Jin *et al.*, "A survey of infrared and visual image fusion methods," *Infr. Phys. Technol.*, vol. 85, pp. 478–501, Sep. 2017.

[6] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.

[7] 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, Jul. 2013.

[8] 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.

[9] 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.

[10] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-SR: A magnification-arbitrary network for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1575–1584.

[11] M. Bertalmío, V. Caselles, E. Provenzi, and A. Rizzi, "Perceptual color correction through variational techniques," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1058–1072, Apr. 2007.

[12] H. Li, H. Qiu, Z. Yu, and Y. Zhang, "Infrared and visible image fusion scheme based on NSCT and low-level visual features," *Infr. Phys. Technol.*, vol. 76, pp. 174–184, May 2016.

[13] 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.

[14] 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.

[15] B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," *Inf. Fusion*, vol. 13, no. 1, pp. 10–19, Jan. 2012.

[16] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[17] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[18] 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.

[19] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

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

[21] J. Li *et al.*, "DRPL: Deep regression pair learning for multi-focus image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4816–4831, 2020.

[22] 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.

[23] 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, May 2018, Art. no. 1850018.

[24] F. Lahoud and S. Süsstrunk, "Fast and efficient zero-learning image fusion," 2019, *arXiv:1905.03590*. [Online]. Available: http://arxiv.org/abs/1905.03590

[25] 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.

[26] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "A symmetric encoder-decoder with residual block for infrared and visible image fusion," 2019, *arXiv:1905.11447*. [Online]. Available: http://arxiv.org/abs/1905.11447

[27] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Trans. Image Process.*, vol. 29, pp. 3845–3858, 2020.

[28] 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.

[29] 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.

[30] 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.

[31] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3224–3232.

[32] U. Upadhyay and S. P. Awate, "Robust super-resolution GAN, with manifold-based and perception loss," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1372–1376.

[33] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–16.

[34] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1–10.

[35] S. Liu *et al.*, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.

[36] 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. 4700–4708.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[38] 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. 4714–4722.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[40] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1664–1673.

[41] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[42] G. Piella, "Image fusion for enhanced visualization: A variational approach," *Int. J. Comput. Vis.*, vol. 83, no. 1, pp. 1–11, Jun. 2009.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

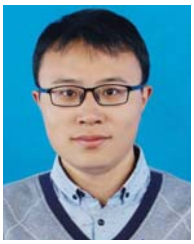
[44] N. Cvejic, C. N. Canagarajah, and D. R. Bull, "Image fusion metric based on mutual information and Tsallis entropy," *Electron. Lett.*, vol. 42, no. 11, pp. 626–627, May 2006.

[45] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Inf. Fusion*, vol. 9, no. 2, pp. 156–160, Apr. 2008.

[46] 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, Apr. 2007.
[47] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'information measure for performance of image fusion,"' *Electron. Lett.*, vol. 44, no. 18, pp. 1066–1067, 2008.
[48] 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.
[49] 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.

Huafeng Li received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering from Chongqing University in 2009 and 2012, respectively. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His research interests include image processing, computer vision, machine learning, and information fusion.

Yueliang Cen received the B.S. degree in network engineering from Yunnan University, China, in 2014. She is currently pursuing the master's degree in software engineering with the School of Information Engineering and Automation, Kunming University of Science and Technology. Her research interests include image processing and computer vision.



Yu Liu (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China in 2011 and 2016, respectively. He is currently an Associate Professor with the Department of Biomedical Engineering, Hefei University of Technology. His research interests include image processing, computer vision, information fusion, and machine learning. In particular, he is interested in image fusion, image restoration, visual recognition, and deep learning. He is serving as an Editorial Board Member for *Information Fusion*.

Xun Chen (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Science and Technology of China in 2009 and the Ph.D. degree in biomedical engineering from The University of British Columbia (UBC), Canada. He has been a Research Scientist with the Department of Electrical and Computer Engineering, UBC. He is currently a Full Professor and the Head of the Department of Electrical Engineering and Information Science, University of Science and Technology of China. He has published over 100 scientific articles in prestigious IEEE/Elsevier journals and conferences. His research interests include statistical signal processing and machine learning in biomedical applications. He is serving as an Area Editor for *Signal Processing: Image Communication* and an Associate Editor for IEEE SIGNAL PROCESSING LETTERS and *Frontiers in Neuroscience*.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language process, image processing, and machine learning.