STDFusionNet: An Infrared and Visible Image Fusion Network Based on Salient Target Detection

Article *in* IEEE Transactions on Instrumentation and Measurement · January 2021

DOI: 10.1109/TIM.2021.3075747

5 authors, including:

Jiayi Ma
Wuhan University
347 PUBLICATIONS **31,069** CITATIONS
SEE PROFILE

Tang Linfeng
Wuhan University
18 PUBLICATIONS **2,987** CITATIONS
SEE PROFILE

Meilong Xu
Wuhan University
2 PUBLICATIONS **398** CITATIONS
SEE PROFILE

Hao Zhang
Wuhan University
39 PUBLICATIONS **4,364** CITATIONS
SEE PROFILE

STDFusionNet: An Infrared and Visible Image Fusion Network Based on Salient Target Detection

Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao

***Abstract*—In this paper, we propose an infrared and visible image fusion network based on the salient target detection, termed as STDFusionNet, which can preserve the thermal targets in infrared images and the texture structures in visible images. Firstly, a salient target mask is dedicated to annotating regions of the infrared image that people or machines pay more attention to, so as to provide spatial guidance for the integration of different information. Secondly, we combine this salient target mask to design a specific loss function to guide the extraction and reconstruction of features. Specifically, the feature extraction network can selectively extract salient target features from infrared images and background texture features from visible images, while the feature reconstruction network can effectively fuse these features and reconstruct the desired results. It is worth noting that the salient target mask is only required in the training phase, which enables the proposed STDFusionNet to be an end-to-end model. In other words, our STDFusionNet can fulfill salient target detection and key information fusion in an implicit manner. Extensive qualitative and quantitative experiments demonstrate the superiority of our fusion algorithm over the state-of-the-art, where our algorithm is much faster and the fusion results look like high-quality visible images with clear highlighted infrared targets. Moreover, the experimental results on the public datasets verify that our algorithm can improve the EN, MI, VIF and SF metrics with 1.25%, 22.65%, 4.3% and 0.89% gains, respectively. Our code is publicly available at: https://github.com/Linfeng-Tang/STDFusionNet.**

***Index Terms*—Image fusion, salient target detection, deep learning, infrared image, mask.**

I. Introduction

The image captured by a single sensor or under a single shooting setting can only describe the imaging scene from a limited perspective, hence fusing complementary images from various sensors or different shooting settings contributes to enhanced understanding of the scene. In all image fusion scenarios, the infrared and visible image fusion is probably the most widely used [1]. Infrared images capture thermal radiation emitted from objects, which can effectively highlight salient targets but lack texture details. In contrast, visible images usually contain rich structure information, but are easily affected by the environment to lose targets. This complementarity gives us the opportunity to fuse them to obtain desired results that contain both significant thermal targets and rich textural details. Due to the excellent characteristics of fused results, infrared and visible image fusion has been widely used in military actions, target detection, recognition, surveillance, and many other domains.

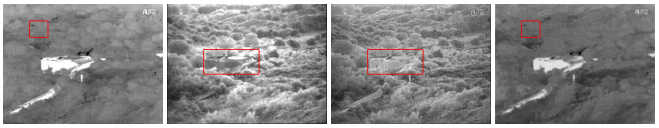


Fig. 1. The weakening of useful information in existing methods. From left to right: infrared image, visible image, results of U2Fusion [9] and FusionGAN [10].

In the past few decades, numerous traditional infrared and visible image fusion algorithms have been proposed, which can be classified into five classes, including multi-scale decomposition-based methods [2], [3], saliency-based methods [4], sparse representation-based methods [5], optimization-based methods [6], and hybrid-based methods [7]. Although the above algorithms have achieved relatively satisfactory fusion performance in most cases, there are still some drawbacks. i) Existing traditional methods generally use the same transformation or representation to extract features from source images without considering the inherent properties of infrared and visible images. ii) Measurement of activity levels and design of fusion rules in most methods are manual and tend to become more complex [8].

In recent years, as the deep learning techniques are maturing, researchers have proposed several deep fusion algorithms. Generally, the predominant deep fusion methods can be divided into two categories, *i.e.* convolutional neural network (CNN)-based methods [12], [11] and generative adversarial network (GAN)-based methods [13]. The CNN-based fusion methods [14], [9], [15] usually rely on the powerful fitting ability of neural networks to realize the extraction and reconstruction of effective information under the guidance of well-designed loss functions. The GAN-based fusion methods establish an adversarial game between the fused image and source images, so as to make the fused image approximate the desired probability distribution without supervision [10], [16]. Although the existing deep learning-based methods have achieved relatively good fusion performance compared to traditional methods, there are still a challenge that cannot be ignored. Specifically, previous deep fusion methods treat different regions of various source images without distinction when constructing loss functions, which means that these methods introduce a lot of redundant or even

This work was supported by the National Natural Science Foundation of China under Grant 61773295, the Key Research and Development Program of Hubei Province under Grant 2020BAB113, and the Natural Science Foundation of Hubei Province under Grant 2019CFA037.

J. Ma, L. Tang, M. Xu and H. Zhang are with the Electronic Information School, Wuhan University, Wuhan, 430072, China (e-mail: jyma2010@gmail.com, linfeng0419@gmail.com, melonxu@whu.edu.cn, zhpersonalbox@gmail.com).

G. Xiao is with the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, College of Computer and Control Engineering, Minjiang University, Fuzhou, 350108, China (e-mail: gbx@mju.edu.cn).

Fig. 2. Schematic illustration of STDFusionNet. From left to right: the infrared image, the visible image, the fusion result of a traditional method GTF [6], the fusion result of a deep learning-based method DenseFuse [11], and the result of our proposed STDFusionNet. The red box and green box show that GTF and DenseFuse suffer detail loss, blurred edges, and artifacts, and our result better highlights prominent targets and has abundant textures.

invalid information in the fusion process. As a result, the useful information is weakened in the fused image. We provide a typical example in Fig. 1 to illustrate this more intuitively, in which U2Fusion [9] is a representative CNN-based method and FusionGAN [10] is a typical GAN-based method. It can be seen that U2Fusion weakens the salient target, while FusionGAN weakens the background textures.

To address the above challenges, we propose a new network based on the salient target detection for infrared and visible image fusion, namely STDFusionNet. First, for infrared images, humans and machines primarily pay attention to the regions where salient targets such as pedestrians, vehicles, hunkers, *etc*., are located. For visible images, the rich background textures help to make the imaging scene more vivid. Therefore, we can defined that the most meaningful information for fusion process is the significant thermal targets in the infrared image and the background texture structures in the visible image. Based on this definition, we annotate the salient targets in the infrared image to obtain the salient target masks. Then, the obtained salient masks are introduced into the design of the specific loss function, which selectively drive the network to extract and reconstruct the above-defined effective features. In addition, due to the significant differences in multi-modal source images, we adopt the pseudo-siamese network to extract different types of information from source images with distinction, such as the salient target intensity and background texture structures. It is worth emphasizing that the salient target mask is only utilized to guide the training of networks, and is not required to be fed into the network during the testing phase, and hence our network is an end-to-end model. Under these specific designs, our STDFusionNet effectively addresses the problems of the feature extraction effectiveness and desired information definition, *etc.*

To intuitively show the performance of our method, we provide a typical example in Fig. 2, with a traditional GTF [6] and a deep learning-based DenseFuse [11] for comparison. GTF views the fused image as an infrared image with additional visible gradients. Even though the targets in its result can be highlighted, the textures are not sharp, and artifacts are presented, leading to unnatural fused image. On the contrary, in the result of DenseFuse, the texture details are better preserved, while the thermal information of the target is weakened. Our STDFusionNet has the advantages of both GTF and DenseFuse. Specifically, our method implicitly achieves the salient target detection and the extraction and reconstruction of effective information, and the fused image could highlight significant targets while retaining abundant textures (*e.g.*, the bush, road, and wall).

The main contributions of this work include the following three aspects:

- We introduce the salient target mask to the specific loss function, which can guide the network to detect the thermal radiation targets in the infrared image, fusing them with the background texture details in the visible image.
- We explicitly define the desired information in the fusion process as the salient target in the infrared image and the background textures in the visible image. To the best of our knowledge, this is the first precise definition for the target of infrared and visible image fusion.
- Extensive experiments demonstrate the superiority of our method over state-of-the-art alternatives. Compared with the competitors, our approach could generate fusion results looking like high-quality visible images with highlighted targets, which contribute to improving target recognition and scene understanding.

The rest of this paper is organized as follows. In section II, we briefly introduce the related works on image fusion. In section III, the proposed method is introduced in detail. Section IV illustrates the fusion performance of our method on public datasets with comparisons to other approaches, followed by some concluding remarks in section V.

II. Related Work

In this section, we review the existing infrared and visible image fusion approaches, including traditional fusion methods and deep learning-based fusion methods.

A. Traditional Fusion Methods

Traditional fusion methods generally manually design activity level measurement or fusion rules to practice image fusion in the spatial or transform domain, which can be divided into five categories according to their principles, including multi-scale transform-based methods [17], [2], saliency-based methods [4], [18], sparse representation-based methods [5], [19], optimization-based methods [6], [20] and hybrid methods [21], [7]. The main ideas of these methods are discussed below.

The multi-scale transform-based methods believe that objects in the physical world are typically composed of components of various scales, and the multi-scale transform is consistent with the human visual system. Therefore, the fused images obtained by multi-scale transform can have pleasing visual

effect [1]. In general, infrared and visible image fusion schemes based on multi-scale decomposition typically involve three steps. Firstly, all source images are decomposed into a series of multi-scale representations. Subsequently, the multi-scale representations of original images are fused according to specific fusion rules. Eventually, the fused image is obtained by performing corresponding inverse transforms on the fused multi-scale representations [17].

The saliency-based methods are usually built on the basis that salient targets are more easily perceived by human vision than their adjacent objects or pixels. Saliency is applied to infrared and visible image fusion in two main ways, *i.e.*, weight calculation and salient target extraction. The former is usually combined with multi-scale transforms, where the source images are decomposed into base and detail layers through the multi-scale transform. Then saliency detection is used to obtain a saliency map of the base or detail layer, and then a weight map of base or detail layer is obtained from the saliency map [7]. The latter uses saliency detection to extract information about the significant regions from the infrared and visible images and then integrates the crucial information into the final fused image [18].Generally, the saliency-based methods can maintain the integrity and pixel intensity of the significant object regions and improve the visual quality of the fused image [4].

The premise of the sparse representation-based methods is to learn an over-complete dictionary from a great number of high-quality images, which are usually achieved by the joint sparse representation [22] and the convolutional sparse representation [5]. Then, the sparse representation coefficients of source images can be obtained by the learned over-complete dictionary, and be fused according to the given fusion rule. Finally, the fused image is reconstructed from the fused sparse representation coefficients with the learned over-complete dictionary.

The optimization-based methods generate the desired fusion result via minimizing an objective function [6], [20]. Therefore, the key to such methods lies in the design of objective functions. The construction of the objective functions should consider two aspects, saying the overall intensity fidelity and texture structure preservation. The former constrains the fused result to have the desired brightness distribution, while the latter drives the fused result to contain rich texture details. The above-mentioned infrared and visible image fusion methods all have their strengths and weaknesses, and the hybrid models combine their strengths to improve the fusion performance [21], [7].

B. Deep Learning-based Fusion Methods

Relying on the excellent feature extraction capabilities of neural networks, deep learning has promoted tremendous progress in image fusion. Early deep learning-based methods only adopt the neural network to construct a weight map or extract features [9]. Liu *et al.* adopted the pre-trained convolutional neural network to implement activity level measurement of source images and generate a weight map, in which the whole fusion process is based on pyramids [12]. However, the neural network is not specifically trained for image fusion, which limits the fusion performance.

With further research, some deep methods based on auto-encoder are proposed. These methods usually pre-train an auto-encoder to implement feature extraction and image recovery, while the feature fusion is fulfilled by traditional rules. Li *et al.* introduced the dense block into the encoder and decoder to design a new image fusion network, termed as DenseFuse [11]. In the fusion layer, DenseFuse is achieved using the conventional addition and l1-norm strategy. Considering that a network without down-sampling cannot extract multi-scale features from source images, the nest connection-based networks are introduced in the NestFuse to extract information from source images in a multi-scale perspective [23]. The spatial and channel attention models are used to fuse the extracted information, but it is worth mentioning that the attention mechanism used for fusion is still unlearnable.

Since the unsupervised distribution estimation ability of the generative adversarial network is very suitable for the image fusion task, more and more GAN-based fusion methods are proposed. Ma *et al.* established an adversarial game between the fused result and the visible image to further enhance the preservation of texture structures. However, this single adversarial mechanism can easily lead to unbalanced fusion. To ameliorate this problem, they later proposed the dual-discriminator conditional generative adversarial network (DDcGAN) [13] to realize image fusion, in which both the infrared image and the visible image participates in the adversarial games. It is worth noting that the generative adversarial network with dual discriminators is not easy to train. In this context, a generative adversarial network with multi-classification constrained is proposed [24] to achieve information balance in the fusion process, in which the multi-distribution simultaneous estimation is done by a single adversarial game.

Due to the strong ability of feature representation in neural networks, the varied information can be represented in a unified way [9]. A growing number of researchers are dedicated to exploring the general image fusion framework. Zhang *et al*. first utilized two convolutional layers to extract the salient features from source images. Then, they selected appropriate fusion rules according to the type of input images to fusion the source images features, and recovered the fused images from the convolutional features by two convolutional layers [14]. Their proposed network framework only needs to be trained on one type of image fusion dataset and subsequently adjusts the fusion rules according to the type of source images, thus implementing a unified network to solve various fusion tasks. In contrast, Zhang *et al*. proposed a network structure based on proportional maintenance of gradient and intensity which adapt to different fusion tasks via adjusting the weights of the loss terms when constructing the loss function [25]. Considering the cross-fertilization between different fusion tasks, U2Fusion was trained sequentially on a unified model for different fusion missions, and a unified model for multiple fusion tasks was obtained [9].

Compared with the above-mentioned methods, the proposed STDFusionNet has two main technical contributions. First, the desired information in the image fusion process is defined as the salient target in the infrared image and the texture information

in the visible image. The defined desired information can provide a more explicit optimization direction for parameter learning. Second, we design a special loss in conjunction with the salient target mask to guide the network to achieve salient target detection and information fusion. This enables the fused images generated by STDFusionNet to retain as much important information as possible in the source images and reduce the effect of redundant information.

III. Method

In this section, we describe the proposed infrared and visible image fusion network based on the salient target detection, STDFusionNet. First, we provide the problem formulation of the STDFusionNet, in which the core ideas are discussed. Then, we describe the designed loss function in detail. Finally, we give the network architecture of the proposed STDFusionNet.

A. Problem Formulation

The target of image fusion is extracting significant information from multiple source images and fusing the complementary information to generate a synthesized image. The key to this problem is how to define the most meaningful information and how to fuse the complementary information. In infrared and visible image fusion, the most critical information is the salient targets and the texture structures, which are contained in infrared images and visible images, respectively. Therefore, we explicitly define the desired information as the salient target information in infrared images and the background texture structure information in visible images. Consequently, there are two keys to image fusion based on this definition.

The first key is to determine the salient target in the infrared image. As we observed, the significant information of infrared images is mainly presented in the regions containing objects (*e.g.*, pedestrians, vehicles and bunkers) that can emit more heat. Hence, the proposed network should learn to automatically detect these regions from infrared images. The second key is to accurately extract the desired information from the detected regions and perform effective fusion and reconstruction. In other words, the fused result should accurately contain the salient target in the infrared image and the background texture in the visible image. The specific loss function and effective network structure are designed to address the above two key problems.

First, we propose a specific loss function to constrain the fusion process, in which the salient target mask is introduced to guide the network to detect salient areas, while the preservation of thermal targets and background texture is achieved by ensuring the intensity and gradient consistency in specific regions. Second, we design an effective network structure to realize feature extraction, fusion and reconstruction. Concretely, the feature extraction network adopts a pseudo-siamese network architecture to treat source images differently, so as to selectively extract salient target features from the infrared image I_{ir} and background texture features from the visible image I_{vi}. Eventually, the feature reconstruction network fuses the extracted features and reconstructs the fused image I_f, highlighting the salient targets in the infrared image while preserving the texture details of the visible image. Under the above design, our model can implicitly realize salient target detection and desired information fusion.

B. Loss Function

The loss function determines the type of information retained in the fused image and the proportional relationship between various information. The loss function of our STDFusionNet consists of two kinds of losses, *i.e.*, the pixel loss and the gradient loss. The pixel loss constrains the pixel intensity of the fused image to be consistent with source images, while the gradient loss forces the fused image to contain more detailed information. We construct the pixel loss and gradient loss for the salient regions and background areas. Combined with the salient target mask I_m, the desired result I_d can be defined as follows:

$$I_d = I_m \circ I_{ir} + (1 - I_m) \circ I_{vi} \ , \tag{1}$$

where the operator $\circ$ denotes element-wise multiplication.

Similarly, the fused image generated by STDFusionNet can be segmented into a prominent region $I_m \circ I_f$ containing the thermal infrared target and a background region $(1 - I_m) \circ I_f$ with texture details.

Therefore, we construct the corresponding losses in the salient and background regions respectively for guiding the optimization of the STDFusionNet. On the one hand, we constrain the fused image to have the same pixel intensity distribution as the desired image. The salient pixel loss $\mathcal{L}_{\text{pixel}}^{\text{salient}}$ and the background pixel loss $\mathcal{L}_{\text{pixel}}^{\text{back}}$ are formulated as:

$$\mathcal{L}_{\text{pixel}}^{\text{salient}} = \frac{1}{HW} \left\| I_m \circ (I_f - I_{ir}) \right\|_1 , \tag{2}$$

$$\mathcal{L}_{\text{pixel}}^{\text{back}} = \frac{1}{HW} \left\| (1 - I_m) \circ (I_f - I_{vi}) \right\|_1 , \tag{3}$$

where H and W are the height and width of the image, respectively, $\|\cdot\|_1$ stands for the l_1-norm. On the other hand, the gradient loss is introduced to enhance the constrains on the network in order to force the fused images with sharper textures and salient targets with sharpened edges. Similar to the definition of the pixel loss, the gradient loss also contains the salient gradient loss $\mathcal{L}_{\text{grad}}^{\text{salient}}$ and the background gradient loss $\mathcal{L}_{\text{grad}}^{\text{back}}$, which are more precisely formulated as follows:

$$\mathcal{L}_{\text{grad}}^{\text{salient}} = \frac{1}{HW} \left\| I_m \circ (\nabla I_f - \nabla I_{ir}) \right\|_1 , \tag{4}$$

$$\mathcal{L}_{\text{grad}}^{\text{back}} = \frac{1}{HW} \left\| (1 - I_m) \circ (\nabla I_f - \nabla I_{vi}) \right\|_1 , \tag{5}$$

where ∇ denotes the gradient operator, in this paper, we employ the Sobel operator to compute the gradient of an image.

Different from the previous method, we treat pixel loss and gradient loss in the same region equally, so the final loss function is defined as:

$$\mathcal{L} = (\mathcal{L}_{\text{pixel}}^{\text{back}} + \mathcal{L}_{\text{grad}}^{\text{back}}) + \alpha(\mathcal{L}_{\text{pixel}}^{\text{salient}} + \mathcal{L}_{\text{grad}}^{\text{salient}}), \tag{6}$$

where α is the hyper-parameter that controls the loss balance in different regions. Due to the introduction of salient region loss *i.e.*, $\mathcal{L}_{\text{pixel}}^{\text{salient}}$ and $\mathcal{L}_{\text{grad}}^{\text{salient}}$, the STDFusionNet has the ability to detect and extract salient targets in infrared images in an implicit manner.

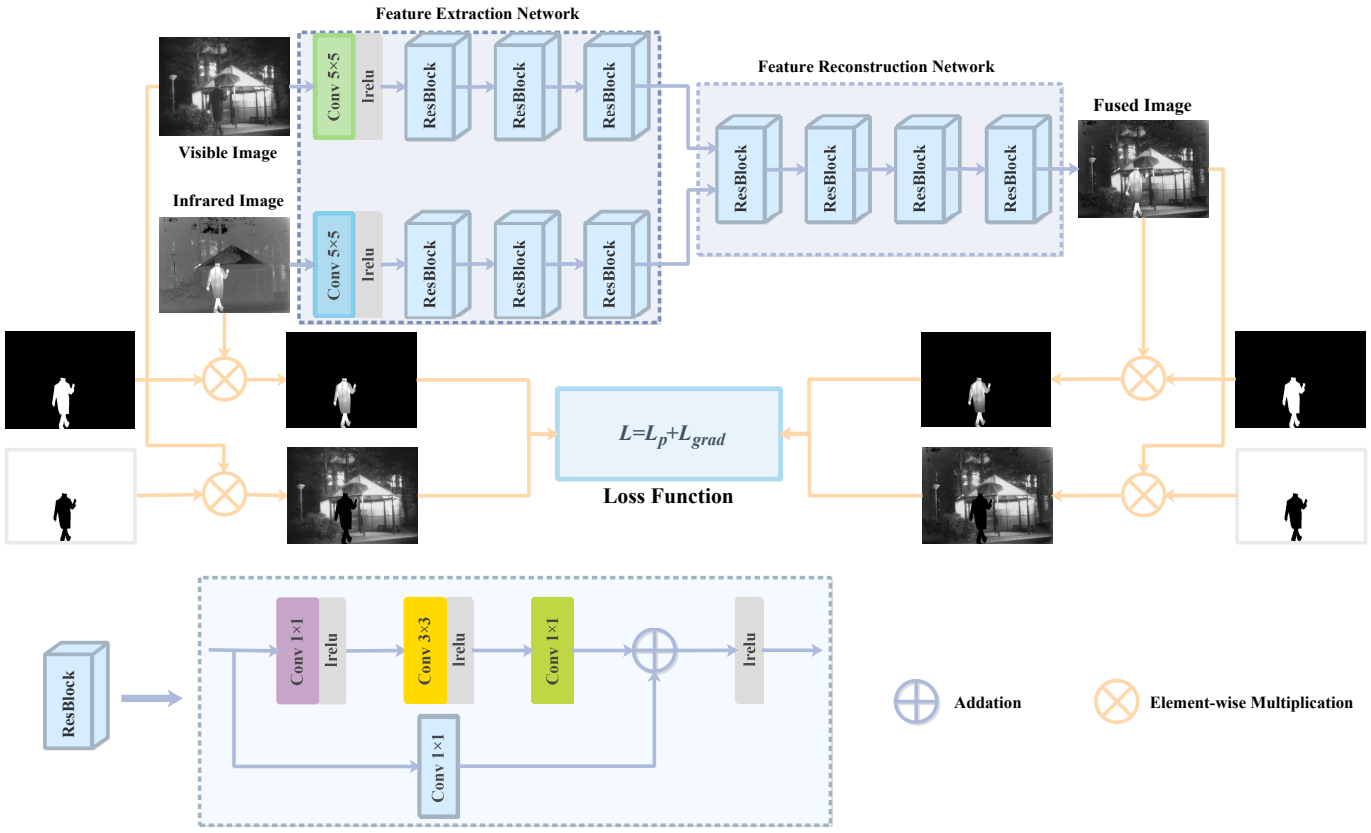


Fig. 3. The architecture of the proposed infrared and visible image fusion network based on the salient target detection. The mask is only needed to construct loss function in the training of the model, and is not needed in the testing phase.

C. Network Architecture

As illustrated in Fig. 3, our network architecture consists of two parts, *i.e.*, the feature extraction network and the feature reconstruction network.

Feature Extraction Network. The feature extraction network is constructed on the basis of the convolutional neural network, and the ResBlock is introduced to enhance the network extraction and alleviate the problem of vanishing/exploding gradients [26]. As shown in Fig. 3, the feature extraction network consists of a common layer and three ResBlocks that can reinforce the extracted information. The common layer consists of a convolutional layer with a kernel size of 5×5 and a Leaky Rectified Linear Unit activation layer. Each Resblock consists of three convolutional layers, named Conv1, Conv2, Conv3, and a skip-connected identity mapping convolutional layer, termed as identity conv. The kernel size of all convolutional layers is 1×1 except for Conv2, which has a kernel size of 3×3. Both Conv1 and Conv2 use the Leaky Rectified Linear Unit as the activation function, while the output of Conv3 and identity conv are summed and followed by the Leaky Rectified Linear Unit activate function. The identity conv is designed to overcome the inconsistent dimensionality of the ResBlock input and output. It is worth noting that considering the different properties of infrared and visible images, both feature extraction networks use the same network architecture, but the respective parameters are trained independently. In combination with the proposed loss function, the feature extraction network can extract the salient feature and texture detail features from source images.

Feature Reconstruction Network. The feature reconstruction network consists of four ResBlocks, which play the role of feature fusion and image reconstruction. It is worth noting that the activation function of the last layer uses Tanh to ensure that the range of variation of the fused image is consistent with that of the input images. The input of the feature extraction network is the concatenation of infrared convolutional features and visible convolutional features in the channel dimension, and its output is the fused image. It is well known that information loss is a catastrophic problem in image fusion missions. Therefore, in all convolutional layers of STDFusionNet, the padding is set to $SAME$, and stride is set to 1. As a result, our network does not introduce any downsampling, and the size of the fused image is consistent with source images.

The purpose of the salient target mask is to highlight the objects (*e.g.*, the pedestrians, vehicles, and bunkers) that radiate numerous heats in infrared images. Therefore, we use the labelme toolbox [27] to annotate salient targets in infrared images and convert them to binary salient target masks. Then, the salient target masks are inverted to obtain the background masks. After that, we multiply the salient target masks and texture background masks with the infrared images and visible images at the pixel level to obtain the source salient target regions and source background texture regions, respectively. Moreover, the fused images are also multiplied with the salient target masks and the texture background masks at the pixel

Fig. 4. Four source and mask image pairs. The top row contains visible images, the second row contains infrared images and the corresponding mask images are in the bottom row.

level to receive the fused salient target regions and the fused background regions. Subsequently, the original salient regions, original salient regions, original background regions, fused salient regions, and fused background regions are applied to construct the specific loss function, which guides the network to realize salient targets detection and information fusion implicitly.

IV. Experiments

In this section, we first describe the experimental settings, including datasets, evaluation metrics and training details. Then, we demonstrate the efficiency of the proposed STDFusionNet on public datasets, and compare it with nine state-of-the-art fusion methods, including two traditional methods, *i.e.*, GTF [6] and MDLatLRR [2], and seven deep learning-based methods, *i.e.*, DenseFuse [11], NestFuse [23], FusionGAN [10], GANMcC [24], IFCNN [14], PMGI [25] and U2Fusion [9]. The implementation of all these nine methods are publicly available, and we set the parameters as reported in the original papers. In addition, we provide the generalization experiment, efficiency comparison, visualization of salient target detection and ablation experiments to verify the effectiveness of specific designs.

A. Experimental Settings

1) Datasets: Our experiments are executed on two datasets, namely the TNO dataset [28] and the RoadScene dataset [9].

The TNO dataset is a common dataset for infrared and visible image fusion, containing various types of military-related scenes. The dataset contains 60 pairs of infrared and visible images, with 3 sequences containing 19, 23, and 32 image pairs, respectively. A typical set of source images and their mask images are shown in Fig. 4. In order to remedy the shortage in quantity of existing datasets, Xu *et al.* released the RoadScene dataset based on the FLIR video [9]. The RoadScene dataset contains 221 pairs of aligned infrared and visible image containing rich scenes of roads, vehicles, and pedestrians. The release of this dataset effectively alleviates the challenges of few image pairs and low spatial resolution in the benchmark dataset.

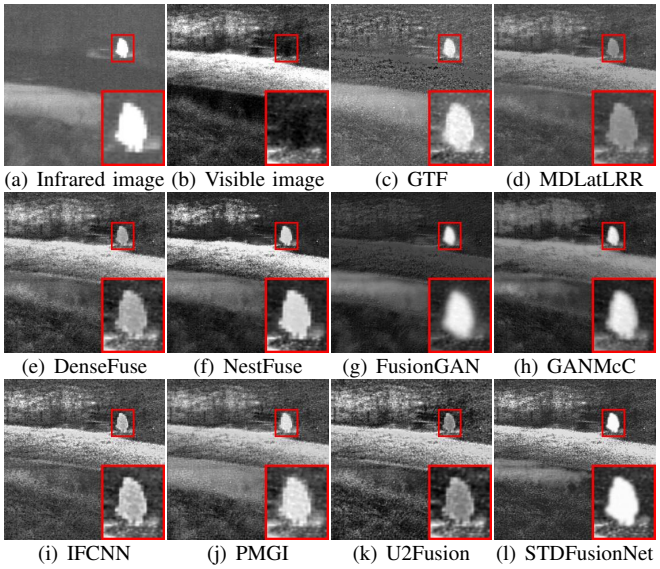


Fig. 5. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the bench image pair. For a clear comparison, we select a salient region (*i.e.*, the red box) in each image and then zoom in it on the bottom right corner.

2) Evaluation Metrics: The assessment of fusion performance can be divided into subjective and objective evaluations. Subjective evaluation relies on human visual perception. For the infrared and visible image fusion, the desired result is better to contain significant thermal targets and rich texture structures. The objective evaluation is a supplement to the subjective evaluation, which usually uses some quantitative metrics to measure the fusion performance. In this paper, four popular metrics are selected, including entropy (EN) [29], mutual information (MI) [30], visual information fidelity (VIF) [31], and spatial frequency (SF) [32]. The definitions of them are as follows.

The EN measures the amount of information contained in a fused image, which is defined based on information theory. The mathematical definition of EN as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l, \tag{7}$$

where L denotes the number of gray levels and p_l is the normalized histogram of the corresponding gray level in the fused image. A larger entropy indicates that the fusion image contains more information and that the method achieves better fusion performance.

The MI metric measures the amount of information transferred from the source images to the fused image. In information theory, MI measures the dependence of two random variables, and in image fusion evaluation, the MI fusion metric is defined as follows:

$$MI = MI_{A,F} + MI_{B,F}, \tag{8}$$

where $MI_{A,F}$ and $MI_{B,F}$ denote the amount of information transferred from source images A and B to fused image F, respectively. The MI between two random variables can be

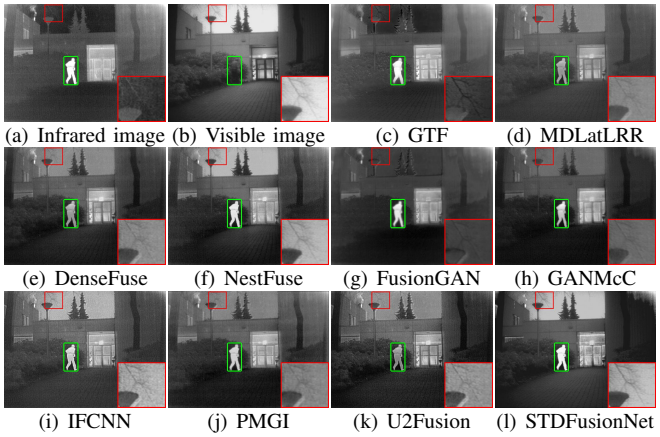


Fig. 6. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the Kaptein_1123 image pair. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

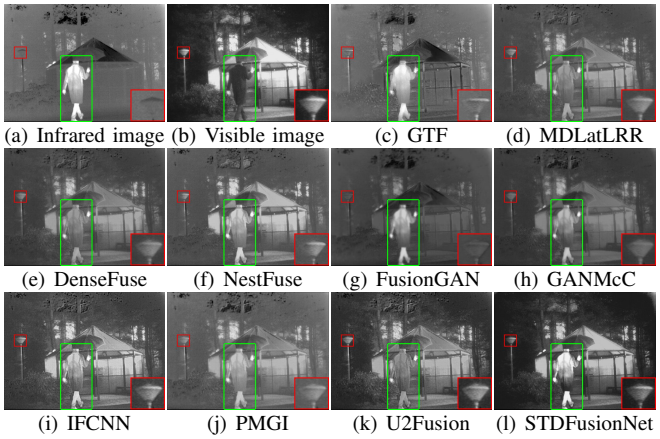


Fig. 7. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the Kaptein_1654 image pair. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

calculated by the Kullback-Leibler measure, which is defined as follows:

$$MI_{X,F} = \sum_{x,f} p_{X,F}(x,f) \log \frac{p_{X,F}(x,f)}{p_X(x)p_F(f)}, \tag{9}$$

where $p_X(x)$ and $p_F(f)$ denote the marginal histograms of the source image X and the fused image F, respectively. $p_{X,F}(x,f)$ means the joint histogram of the source image X and the fused image F. The larger the MI, the more information is transferred from source images to the fused image and the better fusion performance.

The VIF metric measures the information fidelity of the fused image, which is consistent with the human visual system. Computing the VIF metric usually involves the following four steps. First, the source images and the fused image are divided into different blocks; second, evaluate each block for distortion of the visual information; third, the VIF for each sub-band is calculated; finally, calculating the overall metric based on VIF.

The SF metric is a reference-free metric that measures the spatial frequency information contained in the fused image through the row frequency and column frequency. The mathematical definition of SF is as follows:

$$SF = \sqrt{RF^2 + CF^2}, \tag{10}$$

where $RF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - F(i,j-1))^2}$ and $CF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - F(i-1,j))^2}$. A large SF metric indicates that the fused image contains abundant textures and detail information, so the fusion method has excellent performance.

3) Training Details: We train our model on the TNO dataset, and the number of image pairs for training is 20. In order to obtain more training data, we crop each image by setting the stride to 24, and each patch is of the same size 128×128. As a result, the number of produced image patch pairs for training is $6,921$. In the testing phase, we select 20 image pairs from TNO dataset for the comparative experiment and 20 image pairs from RoadScene dataset for the generalization experiment. It is worth noting that each source image is normalized to $[-1, 1]$. We adopt Adam as the optimizer solver for training the model. The proposed algorithm is implemented on the TensorFlow [33] platform. The training parameters are set as follows: the batchsize equal to 32, the number of iteration is set to 30, and the learning rate is set to 10^{-3}. As we have observed, the salient regions take up only a slight proportion of the infrared image. In order to balance the loss of the salient and background regions, in this work, α is set to 7. It is important to note that the source images are fed directly into the fusion network without any cropping during testing. What's more, all the experiments are conducted on a NVIDIA TITAN V GPU and 2.00 GHz Intel(R) Xeon(R) Gold 5117 CPU.

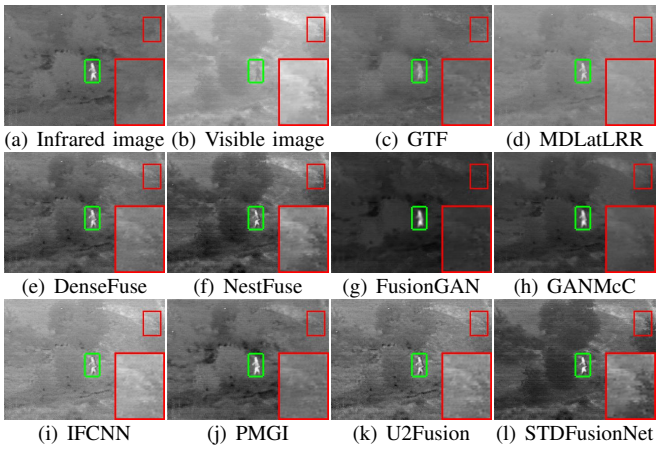


Fig. 8. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the Tree_4915 image pair. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

B. Comparative Experiment

In order to comprehensively evaluate the performance of our method, we compare the proposed STDFusionNet with other nine methods on the TNO dataset.

1) Qualitative Results: To observe the differences in the fusion performance of various algorithms intuitively, we first select four typical pairs of source images (bench, Kaptein_1123, Kaptein_1654, Tree_4915) from the TNO dataset for subjective evaluation. The fused results of different algorithms are shown in Figs. 5 - 8. In Fig. 5, we select a salient region (*i.e.*, red box) in each fused image, then zoom in and place it in the bottom right corner for clear comparison. As shown in Fig. 5, MDLatLRR loses the thermal emission target information, resulting in failure to capture infrared targets in the salient region. While DenseFuse, IFCNN, and U2Fusion retain the thermal emission target information but suffer from serious noise contamination, which comes from visible images. Moreover, FusionGAN preserves thermal radiation information to some extent, but suffers from the shortcoming of blurred infrared target edges. GTF, NestFuse, GANMcC, PMGI, and STDFusionNet are able to highlight the salient target. Especially, STDFusionNet generates fused result that maintaines the contrast of the salient targets well.

In the other three scenes, we select a background region with abundant detail in each fused image, and then zoom in it and put it in the bottom right corner. Also, we label the salient target in a green box. From the fusion results of the remaining three image pairs, we can find that our STDFusionNet not only highlights the salient targets in the scene effectively, but also has a distinct advantage in maintaining the detailed texture of the background region. Specifically, in the Kaptein_1123 scene, the texture of the tree branches in the fused image generated by our method is the clearest, and STDFusionNet is the only method in which the sky is not contaminated by thermal radiation information. While in the Kaptein_1654 scene, the streetlights in the background region are almost consistent with the visible image. Moreover, in the Tree_4915 image pair, it is almost impossible to distinguish the shrubs from their surroundings by other methods except for our method and NestFuse. However, NestFuse weakens the thermal radiation targets in the significant regions. It is worth noting that STDFusionNet can highlight the infrared targets in the salient regions and effectively distinguish the shrubs from their surroundings.

By comparison, it can be found that STDFusionNet is able to selectively preserve salient targets of infrared images and texture details of visible images during the fusion process. This mainly benefits from the manually extracted salient target mask and the constructed loss function during network training.

2) Quantitative Results: The results of four popular quantitative metrics on 20 image pairs from the TNO dataset are shown in Fig. 9 and Table I. Among the four metrics, our method has a significant superiority on three metrics, *i.e.*, EN, MI, and VIF. As for the SF metric, our STDFusionNet only lags behind IFCNN by a narrow margin.

It is important to note that STDFusionNet has the highest value on almost all image pairs on the VIF metric, which is consistent with the conclusions of the subjective evaluation and indicates that STDFusionNet generates fused images with better visual effects. The largest EN demonstrates that the fused images generated by our proposed method have more abundant information than the other nine comparison approaches. The largest MI indicates that our method transfers more information from source images to fused images. Although the SF metric of our algorithm are not the best, the comparable results still indicate that our fused results have sufficient gradient information.

C. Generalization Experiment

The generalization ability of the network is an important basis for evaluating the performance of a deep model. In order to evaluate the generalization ability of our STDFusionNet, we use the image pairs of the RoadScene dataset to test the model trained on the TNO dataset. Since the visible images contained in the RoadScene dataset are in color, we use a specific fusion strategy [34] to achieve the image fusion that preserves the color. Specifically, the RGB visible images are first converted to YCbCr color space. Then the Y channel and the grayscale infrared image are used for fusion as the structural details are mainly in the Y channel. Finally, through the inverse conversion, the fused image can be converted into RGB color space with the Cb and Cr (chrominance) channels of the visible image.

1) Qualitative Results: The fused results of the different methods are shown in Figs. 10 - 13. From the results, we can observed that our STDFusionNet selectively preserves useful information in both infrared and visible images. Compared to fused images generated by other methods, our fused images are very close to the infrared images in salient regions, and the texture structure of the visible images is almost completely preserved in background regions.

Although other methods can highlight the distinctive targets, the background of the fusion images is extremely unpleasant. In particular, the sky in the fused image is heavily contaminated with thermal information, and it is not even possible to accurately estimate the current time and weather from the fused image, which is fatal for road scenes. Moreover, other methods are undesirable for retaining texture details in background areas, such as the writing on walls, bicycles and tree stumps, fences, street lights, *etc*. In contrast, STDFusionNet effectively preserves the background region detail information while maintaining and even enhancing the intensity and contrast of thermal infrared targets in salient regions.

2) Quantitative Results: We also select 20 infrared and visible image pairs from the RoadScene dataset for objective evaluation, and the performance of different methods on the four metrics is shown in Table I and Fig. 14. Similar to the results in the TNO dataset, our STDFusionNet has the best average values for the three metrics, *i.e.*, MI, VIF and SF, but the advantage is not as pronounced as in the TNO dataset. On the EN metric, our method only follows NestFuse by a narrow margin.

In general, both qualitative and quantitative results demonstrate that our STDFusionNet has good generalization performance, which is less affected by the characteristics of imaging sensors.

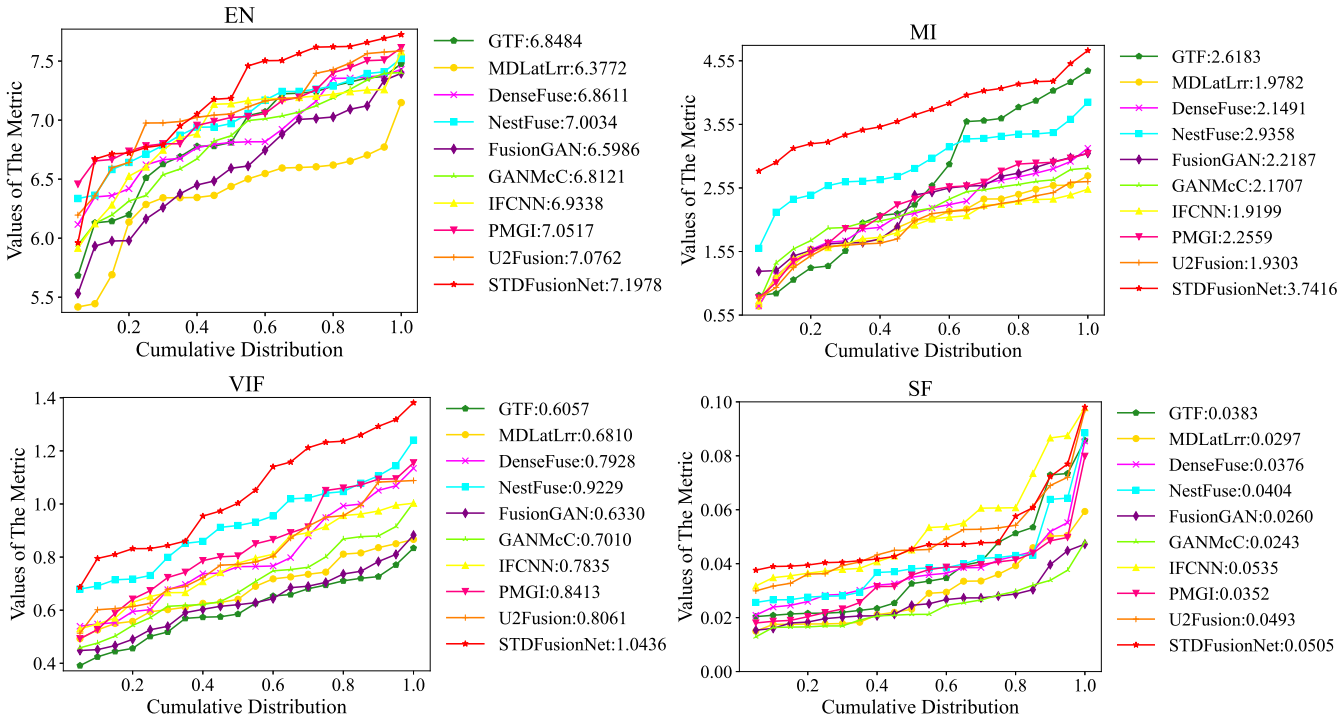


Fig. 9. Quantitative comparisons of the four metrics, *i.e.*, EN, MI, VIF and SF, on twenty image pairs from the TNO dataset. The nine state-of-the-art methods such as GTF, [6] MDLatLRR [2], DenseFuse [11], NestFuse [23], FusionGAN [10], GANMcC [24], IFCNN [14], PMGI[25] and U2Fusion [9] are used for comparison. A point on the curve (x, y) denotes that there are (100*x)% percents of image pairs which have metric values no more than y.

TABLE I

QUANTITATIVE COMPARISONS OF THE FOUR METRICS, *i.e.*, EN, MI, VIF, SF, ON TWENTY IMAGE PAIRS FROM THE TNO DATASET AND THE ROADSENCE DATASET, RESPECTIVELY. **RED** INDICATES THE BEST RESULT AND *BLUE* REPRESENTS THE SECOND BEST RESULT.

	TNO				RoadScene			
	EN	**MI**	**VIF**	**SF**	**EN**	**MI**	**VIF**	**SF**
GTF [6]	6.8484 ± 0.5058	2.6183 ± 1.2131	0.6057 ± 0.1228	0.0383 ± 0.0200	7.3974 ± 0.2669	3.5454 ± 0.6440	0.6455 ± 0.1225	0.0335 ± 0.0073
MDLatLRR [2]	6.3772 ± 0.4305	1.9782 ± 0.5423	0.6810 ± 0.1147	0.0297 ± 0.0134	6.8413 ± 0.2784	3.0232 ± 0.5338	0.7282 ± 0.1270	0.0305 ± 0.0074
DenseFuse [11]	6.8618 ± 0.3880	2.1487 ± 0.6490	0.7930 ± 0.1864	0.0377 ± 0.0145	7.1794 ± 0.2615	3.1297 ± 0.5293	0.7705 ± 0.1390	0.0373 ± 0.0082
NestFuse [23]	7.0034 ± 0.3489	*2.9358 ± 0.5606*	*0.9229 ± 0.1650*	0.0404 ± 0.0157	**7.4875 ± 0.1753**	*3.9642 ± 0.5538*	*0.9262 ± 0.1271*	0.0454 ± 0.0112
FusionGAN [10]	6.5984 ± 0.5161	2.2194 ± 0.6300	0.6330 ± 0.1235	0.0260 ± 0.0089	7.0985 ± 0.2051	3.0262 ± 0.4277	0.6036 ± 0.0663	0.0313 ± 0.0040
GANMcC [24]	6.8099 ± 0.4491	2.1722 ± 0.5346	0.7010 ± 0.1565	0.0243 ± 0.0087	7.2510 ± 0.1892	3.0797 ± 0.5311	0.7180 ± 0.1127	0.0319 ± 0.0049
IFCNN [14]	6.9338 ± 0.4377	1.9199 ± 0.4643	0.7835 ± 0.1576	**0.0535 ± 0.0196**	7.2027 ± 0.1683	3.1281 ± 0.4737	0.7830 ± 0.1173	*0.0516 ± 0.0130*
PMGI [25]	7.0527 ± 0.3281	2.2563 ± 0.6806	0.8413 ± 0.2002	0.0352 ± 0.0146	7.3089 ± 0.1400	3.5906 ± 0.5444	0.8314 ± 0.1246	0.0382 ± 0.0062
U2Fusion [9]	*7.0762 ± 0.3915*	1.9303 ± 0.5256	0.8061 ± 0.1786	0.0493 ± 0.0161	7.1955 ± 0.2966	2.7669 ± 0.5204	0.7371 ± 0.1404	0.0499 ± 0.0102
STDFusionNet	**7.1978 ± 0.4793**	**3.7416 ± 0.5181**	**1.0436 ± 0.2107**	*0.0505 ± 0.0156*	*7.4213 ± 0.1926*	**4.6754 ± 0.7310**	**0.9528 ± 0.1588**	**0.0553 ± 0.0114**

D. Efficiency Comparison

Running efficiency is also an important factor in evaluating model performance. We provide the average running time of different methods on the TNO and RoadScene datasets, which is shown in Table II. From the results, we can see that deep learning-based algorithms have a considerable advantage in runtime due to the utilization of GPUs for acceleration, especially our STDFusionNet. In contrast, traditional methods consume more time to generate the fused image. In particular, MDLatLRR needs to decompose the source image into low-lank parts and saliency parts by latent low-rank representation, so it is particularly time-consuming. In general, our STDFusionNet has the smallest average running time and the lowest standard deviation of running time on both datasets. This illustrates the robustness of our network for different resolution source images and further proves the efficiency of the designed network.

TABLE II

MEAN AND STANDARD DEVIATION OF THE RUNNING TIMES OF ALL METHODS ON THE TNO AND ROADSCENE DATASETS (UNIT: SECOND, **RED** INDICATES THE BEST RESULT AND *BLUE* REPRESENTS THE SECOND BEST RESULT).

Method	**TNO**	**RoadScene**
GTF [6]	2.6122 ± 1.9535	1.8188 ± 0.7396
MDLatlRR [2]	135.0391 ± 72.0068	86.8480 ± 19.8430
DenseFuse [11]	0.7732 ± 0.8658	0.7892 ± 0.763
NestFuse [23]	0.2982 ± 0.4067	0.2187 ± 0.3496
FusionGAN [10]	0.4810 ± 0.6025	0.5118 ± 0.4155
GANMcC [24]	0.7258 ± 0.7856	0.7050 ± 0.4239
IFCNN [14]	*0.0885 ± 0.3358*	*0.0796 ± 0.3172*
PMGI [25]	0.2597 ± 0.4320	0.2721 ± 0.3574
U2Fusion [9]	0.7155 ± 0.7284	0.7820 ± 0.3512
STDFusionNet	**0.0461 ± 0.0497**	**0.0292 ± 0.0333**

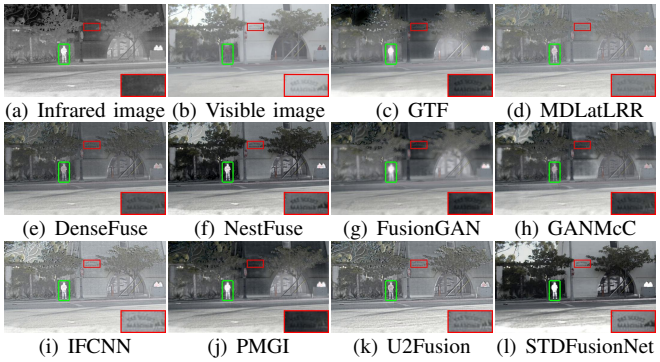


Fig. 10. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the RoadScene dataset. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

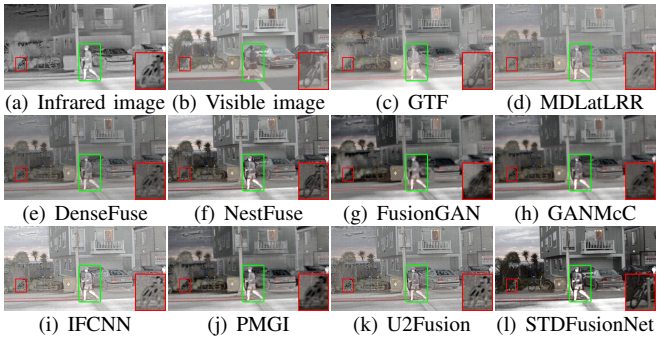


Fig. 11. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the RoadScene dataset. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

E. Visualization of Salient Target Detection

As mentioned earlier, the proposed STDFusionNet can fulfill salient target detection in an implicit manner. Several visual examples are provided to confirm this. The salient region of infrared images and the results of subtracting the visible background regions from fused results are shown in Fig. 15. From these results, we can see that the results of subtracting visible background region from fused images are almost consistent with infrared image salient regions. And there is a slight difference between the salient regions detected by our method and the annotated, which is the additional salient thermal targets detected by our method. These phenomena demonstrate that our STDFusionNet can implicitly performs salient target detection well.

F. Ablation Experiment

1) Desired Information Analysis: In our model, the desired information is explicitly defined as the salient target in the infrared image and the background texture structure in the visible image. To verify the rationality of the desired information definition, we train two models on the TNO dataset based on whether or not to use desired information definition to guide the optimization of the network. More specifically, the

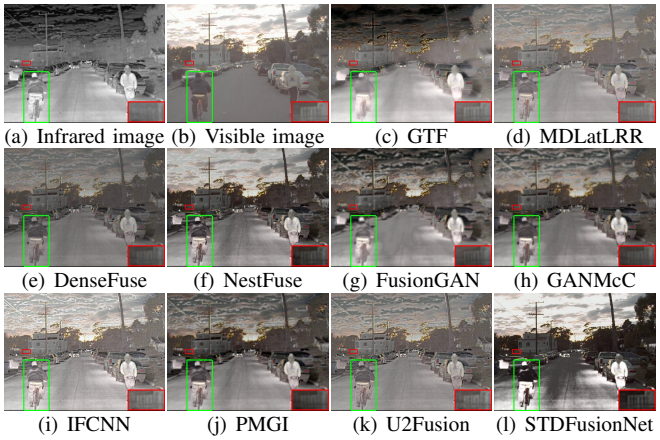


Fig. 12. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the RoadScene dataset. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

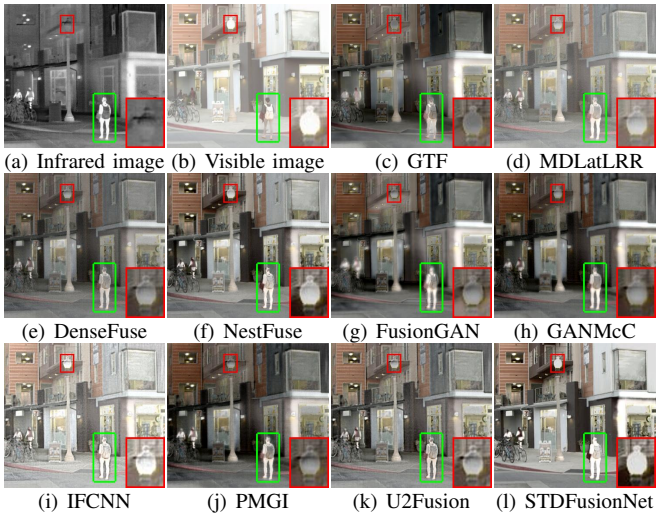


Fig. 13. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on the RoadScene dataset. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and then zoom in it on the bottom right corner and highlight a salient region (*i.e.*, the green box).

main difference between these two models is whether or not salient target masks are introduced into the loss function. Since there is no need to treat the salient and background regions distinctively, the control trade-off parameter α is set to 1 when the salient target masks are removed.

It can be seen from Fig. 16 that with the desired information definition, the fused results of STDFusionNet can not only highlight distinctive targets in salient regions, but also maintain the texture details in the background regions. In contrast, when not using the desired information definition, the network only fuses the infrared and visible images in a coarse manner, resulting in the thermal emission information of infrared images and the texture information of visible images not being well preserved. What's more, as presented in Table III, there is a significant degradation in the performance of the model without

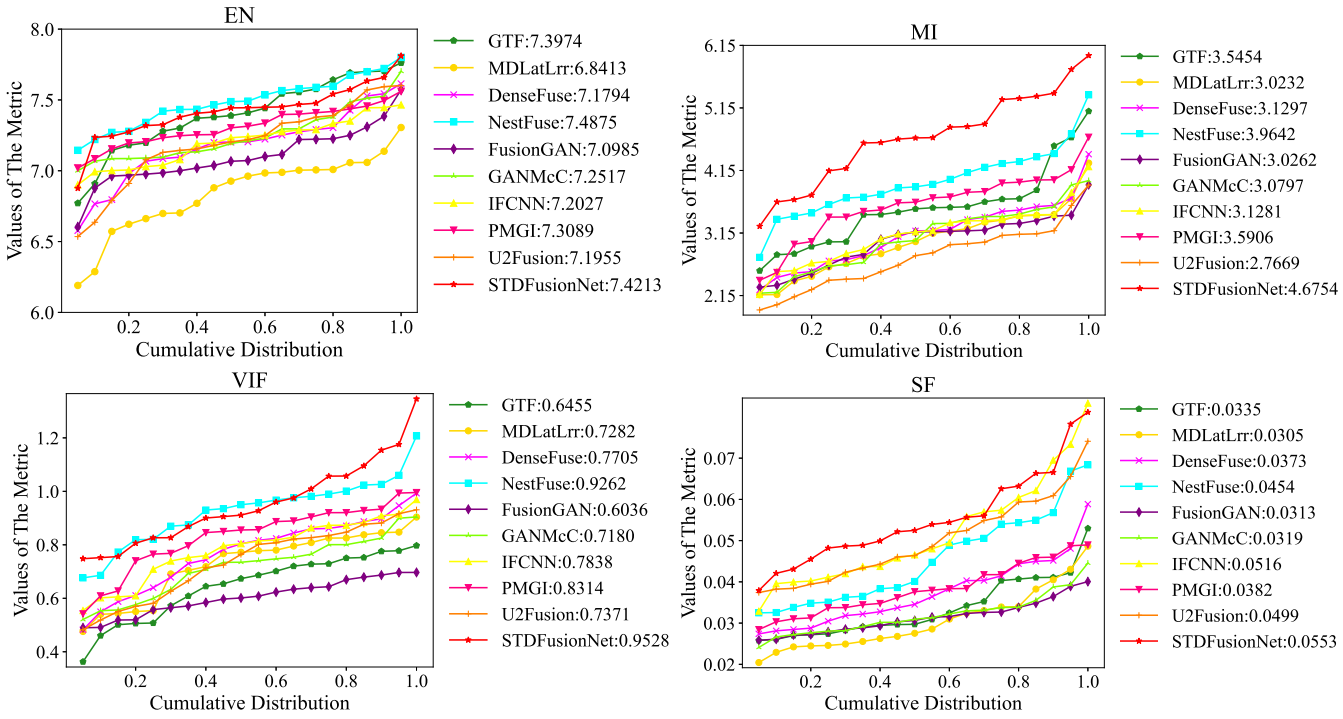


Fig. 14. Quantitative comparisons of the four metrics, *i.e.*, EN, MI, VIF, SF, on twenty image pairs from the RoadSence dataset. The nine state-of-the-art methods such as GTF, [6] MDLatLRR [2], DenseFuse [11], NestFuse [23], FusionGAN [10], are used for comparison. A point on the curve (x, y) denotes that there are (100*x)% percents of image pairs which have metric values no more than y.

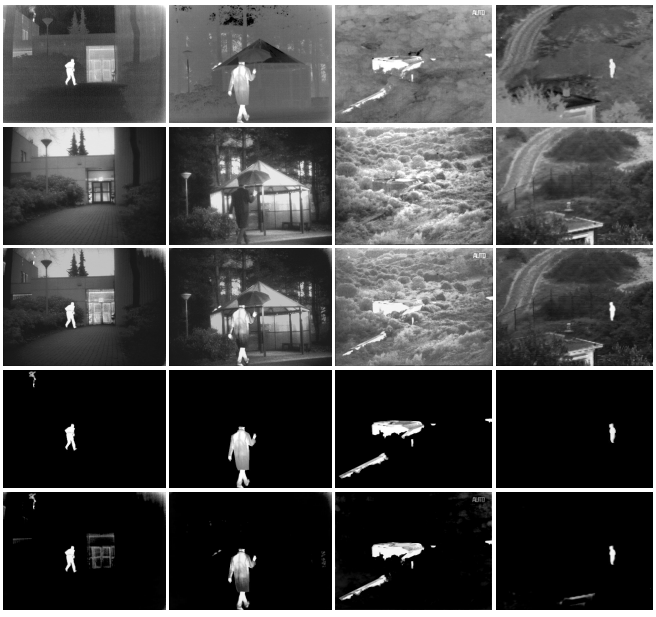


Fig. 15. Visualization of salient target detection on four typical infrared and visible image pairs. From left to right: Kaptein_1123, Kaptein_1654, Bunker and Nato_camp_1816. From top to bottom: infrared images, visible images, results of STDFusionNet, salient regions of infrared images, and difference between fused images and the background regions of visible images.

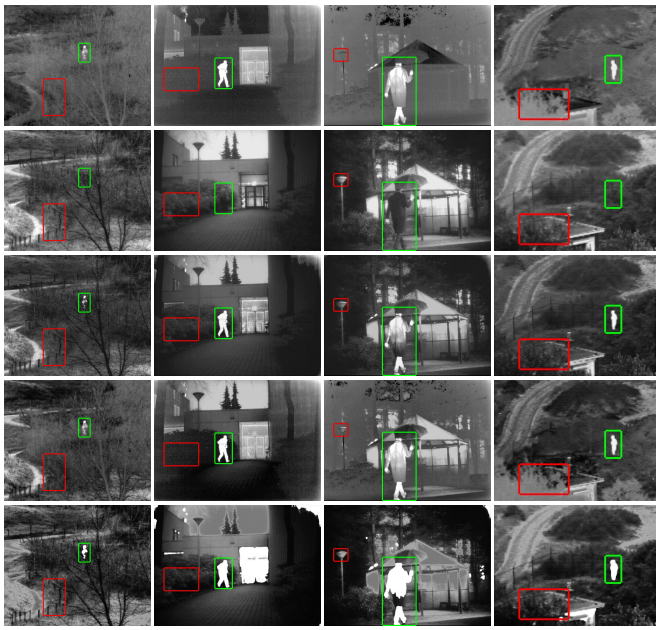


Fig. 16. Visualized results of ablation on four typical infrared and visible image pairs. From left to right: Sandpath, Kaptein_1123, Kaptein_1654 and Nato_camp_1816. From top to bottom: infrared images, visible images, results of STDFusionNet, STDFusionNet without desired information definition, and STDFusionNet without gradient loss.

explicitly defining desired information. Specifically, compared to STDFusionNet, without introducing salient target masks, the EN, MI, VIF, and SF metrics decrease by 15.8%, 52.6%, 42.3%, and 21.5%, respectively. These results prove that our definition of the desired information is reasonable, and it is of great significance to improve the fusion performance.

TABLE III
QUANTITATIVE EVALUATION RESULTS OF ABLATION. THE W/O DESIRED INFORMATION INDICATES THAT STDFUSIONNET WITHOUT DESIRED INFORMATION DEFINITION AND W/O GRADIENT LOSS DENOTES STDFUSIONNET WITHOUT GRADIENT LOSS. **RED** INDICATES THE BEST RESULT AND *BLUE* REPRESENTS THE SECOND BEST RESULT).

	w/o Desired information	w/o Gradient loss	STDFusionNet
EN	*6.5010 ± 0.5142*	6.0294 ± 0.9807	**7.2462 ± 0.4308**
MI	1.9518 ± 0.5197	*3.5511 ± 0.8877*	**4.1213 ± 0.5382**
VIF	0.6142 ± 0.1692	*0.6869 ± 0.2640*	**1.0652 ± 0.1954**
SF	0.0348 ± 0.0060	**0.0691 ± 0.0223**	*0.0489 ± 0.0159*

2) Gradient Loss Analysis: When constructing the loss function, in addition to pixel constraints, the gradient loss is also introduced to force the salient targets in the fused image to have sharper textures and contours. We implement the ablation experiment to demonstrate the effectiveness of the gradient loss. Specifically, we train a model without additional gradient loss, and the results are shown in Fig. 16. It can be seen that when removing the gradient loss, the salient regions hardly have any texture information, and there is also a severe distortion in the salient target shape. In addition, several artifacts occur in the background region. What's more, the results of the quantitative comparison are exhibited in Table III, where all metrics present decreasing tendencies except for the SF metric. These experimental results demonstrate the importance of the gradient loss, which can ensure the textures sharpness of salient targets in the fused image.

V. CONCLUSION

In this paper, we propose a novel infrared and visible image fusion network based on salient target detection, named STDFusionNet. We explicitly define the desired information for infrared and visible image fusion as the salient region of the infrared image and the background region of the visible image. Based on this definition, we introduce the salient target mask into the loss function to precisely guide the optimization of the network. As a result, our model can fulfill salient target detection and information fusion in an implicit manner, and the result not only contains salient thermal targets, but also has rich background textures. Extensive qualitative and quantitative experiments demonstrate the superiority of our STDFusionNet over state-of-the-art methods in terms of both subjective visual effect and quantitative metrics. Moreover, our method is much faster than other comparative methods.

REFERENCES

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

[2] H. Li, X.-J. Wu, and J. Kitler, "Mdlatlrr: A novel decomposition method for infrared and visible image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4733–4746, 2020.

[3] J. Ma and Y. Zhou, "Infrared and visible image fusion via gradientlet filter," *Computer Vision and Image Understanding*, vol. 197-198, p. 103016, 2020.

[4] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

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

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

[7] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

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

[9] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

[11] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[12] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 03, p. 1850018, 2018.

[13] 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 Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

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

[15] H. Xu, X. Wang, and J. Ma, "Drf: Disentangled representation for visible and infrared image fusion," *IEEE Transactions on Instrumentation and Measurement*, 2021.

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

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

[18] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, pp. 94–102, 2017.

[19] M. Wu, Y. Ma, F. Fan, X. Mei, and J. Huang, "Infrared and visible image fusion via joint convolutional sparse representation," *JOSA A*, vol. 37, no. 7, pp. 1105–1115, 2020.

[20] J. Zhao, G. Cui, X. Gong, Y. Zang, S. Tao, and D. Wang, "Fusion of visible and infrared images using global entropy and gradient constrained regularization," *Infrared Physics & Technology*, vol. 81, pp. 201–209, 2017.

[21] F. Fakhari, M. R. Mosavi, and M. M. Lajvardi, "Image fusion based on multi-scale transform and sparse representation: an image energy approach," *IET Image Processing*, vol. 11, no. 11, pp. 1041–1049, 2017.

[22] N. Yu, T. Qiu, F. Bi, and A. Wang, "Image features extraction and fusion based on joint sparse representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1074–1082, 2011.

[23] H. Li, X.-J. Wu, and T. Durrani, "Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9645–9656, 2020.

[24] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.

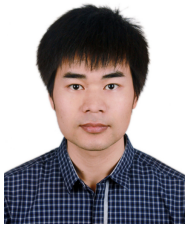
[25] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity." in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 797–12 804.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

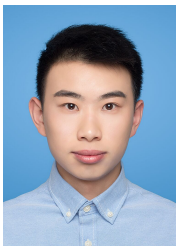
[27] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[28] A. Toet, "TNO Image Fusion Dataset," 4 2014. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[29] J. W. Roberts, J. A. van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, p. 023522, 2008.
[30] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics Letters*, vol. 38, no. 7, pp. 313–315, 2002.
[31] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information Fusion*, vol. 14, no. 2, pp. 127–135, 2013.
[32] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on communications*, vol. 43, no. 12, pp. 2959–2965, 1995.
[33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.
[34] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs." in *ICCV*, vol. 1, no. 2, 2017, p. 3.



Jiayi Ma received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or co-authored more than 150 refereed journal and conference papers, including IEEE TPAMI/TIP, IJCV, CVPR, ICCV, ECCV, *etc*. His research interests include computer vision, machine learning, and remote sensing. Dr. Ma has been identified in the 2020 and 2019 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion*, an Associate Editor of *Neurocomputing* and *Entropy*, and a Guest Editor of *Remote Sensing*.



Linfeng Tang received the B.E. degree from the School of Computer Science and Engineering, Central South University, Changsha, China, in 2020. He is currently a Master student with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Meilong Xu is currently pursuing the bachelor's degree majoring in electronic engineering with the Electronic Information School, Wuhan University. His research interests are in the areas of computer vision and pattern recognition.

Hao Zhang received the B.E. degree from the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China, in 2019. He is currently a Master student with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Guobao Xiao is currently a professor at Minjiang University, China. He was a Postdoctoral Fellow (2016-2018) in the School of Aerospace Engineering at Xiamen University, China. He received the Ph.D. degree in Computer Science and Technology from Xiamen University, China, in 2016. He has published over 30 papers in the international journals and conferences including IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, Pattern Recognition, Pattern Recognition Letters, Computer Vision and Image Understanding, ICCV, ECCV, ACCV, AAAI, ICIP, ICARCV, *etc*. His research interests include machine learning, computer vision, pattern recognition and bioinformatics. He has been awarded the best PhD thesis in Fujian Province and the best PhD thesis award in China Society of Image and Graphics. He serves on the reviewer panel for some international journals and top conferences.