CLF-Net: Contrastive Learning for Infrared and Visible Image Fusion Network

Zhengjie Zhu (朱正杰), Xiaogang Yang (杨小冈), Ruitao Lu (卢瑞涛), Tong Shen (申通), Xueli Xie (谢学立), and Tao Zhang (张涛)

***Abstract*— In this article, we propose an effective infrared and visible image fusion network based on contrastive learning, which is called CLF-Net. A novel noise contrastive estimation framework is introduced into the image fusion to maximize the mutual information between the fused image and source images. First, an unsupervised contrastive learning framework is constructed to promote fused images selectively retaining the most similar features in local areas of different source images. Second, we design a robust contrastive loss based on the deep representations of images, combining with the structural similarity loss to effectively guide the network in extracting and reconstructing features. Specifically, based on the deep representation similarities and structural similarities between the fused image and source images, the loss functions can guide the feature extraction network in adaptively obtaining the salient targets of infrared images and background textures of visible images. Then, the features are reconstructed in the most appropriate manner. In addition, our method is an unsupervised end-to-end model. All of our methods have been tested on public datasets. Based on extensive qualitative and quantitative analysis results, it has been demonstrated that our proposed method performs better than the existing state-of-the-art fusion methods. Our code is publicly available at https://github.com/zzj-dyj/CLF-Net**

***Index Terms*— Contrastive learning, image fusion, infrared image, noise contrastive estimation (NCE), unsupervised learning.**

I. Introduction

AS AN important technology in image processing, image fusion can be utilized to effectively integrate complementary image information from different visual sensors to obtain an information-rich fusion image. Visible and infrared sensors are the two most commonly used visual sensors [1]. The effective fusion of these two types of image information has been widely applied in object recognition [2], detection [3], [4], image enhancement [5], surveillance [6], remote sensing [7], and other fields. Based on the theory of optical imaging, visible images have abundant texture details and high spatial resolution. However, they are also affected by dark environments, fog, and other types of environmental interference. An infrared image is based on the thermal radiation of an object, which can highlight salient targets in the interference environment. However, the infrared image has a low signal-to-noise ratio and lacks texture details. Thus, the fused result has the advantages of the two kinds of source images, which have rich details and salient targets.

In recent years, significant research has been performed regarding deep learning fusion methods due to their powerful representation abilities [8]. These methods can be divided into two categories: the non-end-to-end methods and the end-to-end methods. In the non-end-to-end methods, the design of the feature fusion strategy is the main focus. Currently, the strategies of feature fusion, which are hand-calculated, mainly include addition, l_1-norm [9], attention weighting [10], and so on. However, it is difficult to obtain appropriate hand-calculated features when dealing with different fusion tasks. To eliminate the difficulty of designing a hand-calculated fusion strategy, some end-to-end methods have been proposed [11], [12], [13], [14], [15], [16], [17], [18], [19]. In these methods, the lack of ground truth in the fusion task is a problem that cannot be ignored. To solve this unsupervised problem, the methods [11], [12], and [13] adopt the generative adversarial network (GAN) framework. In addition, methods [15] and [16] guide the trend of image fusion by designing specific loss functions and weighting them. The loss functions of the above methods usually include intensity, gradient, and structure. However, these loss functions do not treat different regions of the source images differently, which results in considerable information redundancy. In VIF-Net [18], the modified structural similarity (M-SSIM) loss, which adaptively calculates the SSIM score by comparing pixel intensity information in sliding windows of different source images, was proposed. By introducing salient target masks, STDFusionNet [19] has a specific loss function to guide the network in effectively merging salient targets in infrared images with background textures in visible images. Although good fusion performance has been achieved using these methods, their loss functions are only based on the shallow features of images but do not make full use of the deep features. In our opinion, it is also effective to reasonably combine deep features to guide training.

To solve the above problems, we propose a new idea inspired by contrastive learning methods [34] in current self-supervised learning tasks. Specifically, Ma *et al.* [19] defined the desired information in the fusion process as the combination of salient targets in infrared images and background textures in visible images. From our perspective, this approach can be more simply stated as follows: we expect that the salient target in the fusion image looks more like that in the infrared

Manuscript received 9 May 2022; revised 7 August 2022; accepted 12 August 2022. Date of publication 7 September 2022; date of current version 19 September 2022. The Associate Editor coordinating the review process was Damodar Reddy Edla. *(Corresponding author: Xiaogang Yang.)*

The authors are with the College of Missile Engineering, Rocket Force University of Engineering, Xi'an 710038, China (e-mail: zzj19980327@163.com; doctoryxg@163.com; lrt19880220@163.com; shentong521@live.com; 18509242741@163.com; sunshinetaoz@163.com).

Digital Object Identifier 10.1109/TIM.2022.3203000

image and the background area looks more like that in the visible image. How do researchers define the term "like"? The answer is contrast. By comparing the similarities and differences between the fusion images and the source images, people can easily choose the fusion image that meets their expectations. To achieve this goal, we propose an effective infrared and visible image fusion network based on contrastive learning (CLF-Net). First, we construct an adaptive contrastive learning framework. In this framework, we focus on the deep representation instead of the image itself, and related local features are maximally reserved by comparing the differences in the dot products (i.e., the cosine similarity) between the feature vectors of the fusion image and the source images. Second, under the above framework, we design a robust contrastive loss, combining with the structural similarity loss to guide the network in extracting and reconstructing features. Specifically, based on the representation similarities and structural similarities between the fused image and source images in the same spatial location, the loss function can be used to guide the feature extraction network adaptively to obtain the salient targets of the infrared images and background textures of the visible images. In addition, because contrastive loss and structural similarity loss are both adaptive, our method is an unsupervised learning process. It is also noted that the contrastive learning framework only participates in the training process of the network. Thus, our CLF-Net is an end-to-end model.

The main contributions of our method can be summarized as follows.

1) We introduce a novel noise contrastive estimation (NCE) framework into image fusion tasks to maximize mutual information (MI) between fused images and source images.
2) We construct an unsupervised contrastive learning framework to promote fused images selectively retaining the most similar feature from different source images. A robust contrastive loss is designed to guide the network to adaptively extract and reconstruct features based on the deep representation.
3) Extensive experiments demonstrate that better performance in terms of qualitative and quantitative analysis is achieved using our method compared with the existing state-of-the-art methods.

The remainder of this article is structured as follows. In Section II, we briefly review the related works on deep-learning-based fusion methods and contrastive learning for computer vision. In Section III, we elaborate on our proposed method. Extensive comparative validation experiments are described in Section IV, followed by the conclusion of our work.

II. Related Work

In this section, we review the existing work and the approaches that are most relevant to our method, including deep-learning-based fusion methods and contrastive learning for computer vision.

A. Deep-Learning-Based Fusion Methods

In recent years, deep learning methods have been widely applied in image fusion tasks and have achieved remarkable results. These methods can be divided into two categories: the non-end-to-end methods and the end-to-end methods.

Initially, some non-end-to-end methods based on autoencoders were proposed. In these methods, the design of the feature fusion strategy is the focus. Li and Wu [9] proposed DenseFuse, which introduces a dense block into the feature extraction layer and exploits traditional addition and L1-norm strategies in the fusion layer. Inspired by the architecture in [20], NestFuse [10] was proposed. In this method, a downsampling network is used to extract multiscale features from source images, and an attention weighting strategy is adopted to fuse the features. Although good performance has been achieved using these methods, the hand-calculated fusion strategy is approximate, which limits further improvement of the fusion performance [8].

To solve the above limitations in the non-end-to-end methods, some end-to-end fusion frameworks have been studied. A GAN-based fusion framework, which was first proposed by Ma *et al.* [11], was established as an adversarial game to constrain the fusion image and obtain more details from the visible images. Based on the game of multiclassification discrimination, GANMcC [13] is used to obtain fused images that more closely resemble the distribution of the source images. In addition to GAN-based methods, several CNN-based end-to-end methods have also been proposed. Hou *et al.* [18] designed a simple end-to-end network that uses M-SSIM and the total variation function to guide the network. Zhang *et al.* [14] utilized two convolutional layers to extract deep features from the source images. Then, they selected elementwise fusion rules to fuse the source image features and reconstructed the fused images by two convolutional layers. Xu *et al.* [16] proposed an adaptive network based on proportional gradient and intensity maintenance, which preserves the adaptive similarity between the fusion result and source images. Ma *et al.* [19] designed a salient target mask to label some salient infrared targets and then designed a specific loss function to guide the extraction and reconstruction of the features. The loss functions in the above methods are all designed based on the shallow features of the image. However, in our opinion, an effective method makes full use of the deep features of images to guide network training.

B. Contrastive Learning for Computer Vision

Contrastive learning has attracted increasing attention in the field of computer vision due to its excellent performance [23]. The concept of contrastive learning was proposed a long time ago, but in recent years, remarkable achievements in the field of computer vision have been achieved using this approach [24]. The core problem of contrastive learning is how to construct the set of positive and negative samples. Hjelm [25] proposed Deep InfoMax, which constructs comparative learning tasks based on local features in images. He *et al.* [26] proposed an efficient comparative learning structure momentum contrast (MoCo), which uses a momentum encoder to encode a single positive sample and multiple negative samples and updates the encoder parameters with momentum. Chen *et al.* [27] proposed a general framework that maximizes the similarity of the two data augmentation projections of the

same image and minimizes the similarity with other images by conducting two random data augmentations on the input image, to achieve a constant visual representation of the same object under different perspectives or interference. Then, the two teams of He and Hinton learned from each other and successively proposed MoCo v2 [28] and SimCLR v2 [29], which are mainly improvements of data augmentation methods and backbone networks. Subsequently, Caron *et al.* [30] took a different approach; instead of aiming to increase the number of negative cases in the optimization direction, all kinds of samples are clustered, and then, all kinds of class clusters are compared. Grill *et al.* [31] proposed a new self-supervised image representation learning method that did not use negative examples and made one encoder stop gradient, which only carried out momentum updates on the parameters of another encoder. Chen and He [32] took the concepts behind BYOL and combined them with the study of Siamese networks, found that the stop gradient is the key to avoiding network collapse, and proposed the SimSiam network.

Given the continuous progress of the theory of contrastive learning, this method has been widely used in many image tasks. For the task of conditional image generation, Kang and Park [33] proposed ContraGAN, which is based on a novel conditional contrastive loss that can learn both data-to-class and data-to-data relations. For the task of image-to-image translation, Park *et al.* [34] proposed contrastive learning, in which the MI between the corresponding image patches in the source domain and target domain is maximized through the framework of contrastive learning to complete the image-to-image translation for unpaired image-to-image translation.

To the best of our knowledge, there are few studies on the application of contrastive learning in the task of infrared and visible image fusion. Inspired by contrastive learning, Luo *et al.* [47] adopted a contrastive difference loss to avoid the trivial solution and promote the disentanglement ability of the autoencoder. The contrastive difference loss can maximize the distinction between the common and private features of source images. However, IFSepR does not construct the positive sample pairs and the NCE framework, which is the main difference from our methods. Therefore, inspired by the NCE framework, we have proposed a novel image fusion algorithm named CLF-Net. The results also show that the image fusion performance can be effectively improved using this network.

III. Method

In this section, we describe the proposed contrastive learning technique for infrared and visible image fusion networks in detail. First, we present the general network architecture of the proposed CLF-Net. Next, we introduce the NCE framework, which is the basis for contrastive learning. Then, we construct a novel adaptive patchwise contrastive learning framework, which reveals the design details of the contrastive loss function. Finally, we describe the designed loss function in detail.

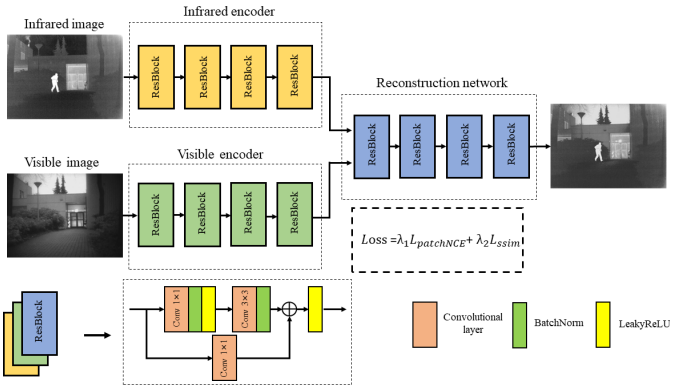


Fig. 1. Architecture of the proposed CLF-Net.

A. Network Architecture

The architecture of CLF-Net is shown in Fig. 1 and consists of two parts: the feature extraction network and the feature reconstruction network.

1) Feature Extraction Network: It consists of two specific encoders. Both encoders are constructed based on a ResBlock to alleviate the well-known problems of vanishing or exploding gradients [22]. As shown in Fig. 1, the feature extraction network consists of four ResBlocks that can reinforce the extracted information. The residual mapping of each ResBlock is composed of two convolutional layers, which are used to extract features. These two layers have kernel sizes of 1×1 and 3×3. The identity mapping, which consists of a convolutional layer with a kernel size of 1×1 is used to adjust the input and output dimensions and maintain their consistency. For infrared images and visible images, the structure of the feature extraction networks (i.e., the infrared encoder and visible encoder) is consistent, but the parameters of these networks are independent of each other.

2) Feature Reconstruction Network: It is directly composed of four ResBlocks. The deep features from the two different encoders are directly concatenated and reconstructed into the fused image. At the end of the feature reconstruction network, we have replaced the activation function leaky rectified linear unit (LeakyReLU) with tanh to ensure that the range of change between the fused image and the source images is consistent.

In all convolutional layers of the ResBlock for the whole process of feature extraction, fusion, and reconstruction, the stride is set to 1, the padding is set to 0 when the kernel size is 3×3, and the padding is set to 1 when the kernel size is 1×1. As a result, there is no downsampling process in CLF-Net, which also means that no information is lost.

B. NCE Framework

NCE is presented as a new estimation principle for parameterized statistical models [35]. The core idea is to determine some characteristics of the original data by learning the difference between the original data distribution sample and the selected noise distribution. This process effectively simplifies the model estimation problem to a dichotomous problem and greatly reduces the computational complexity [37].

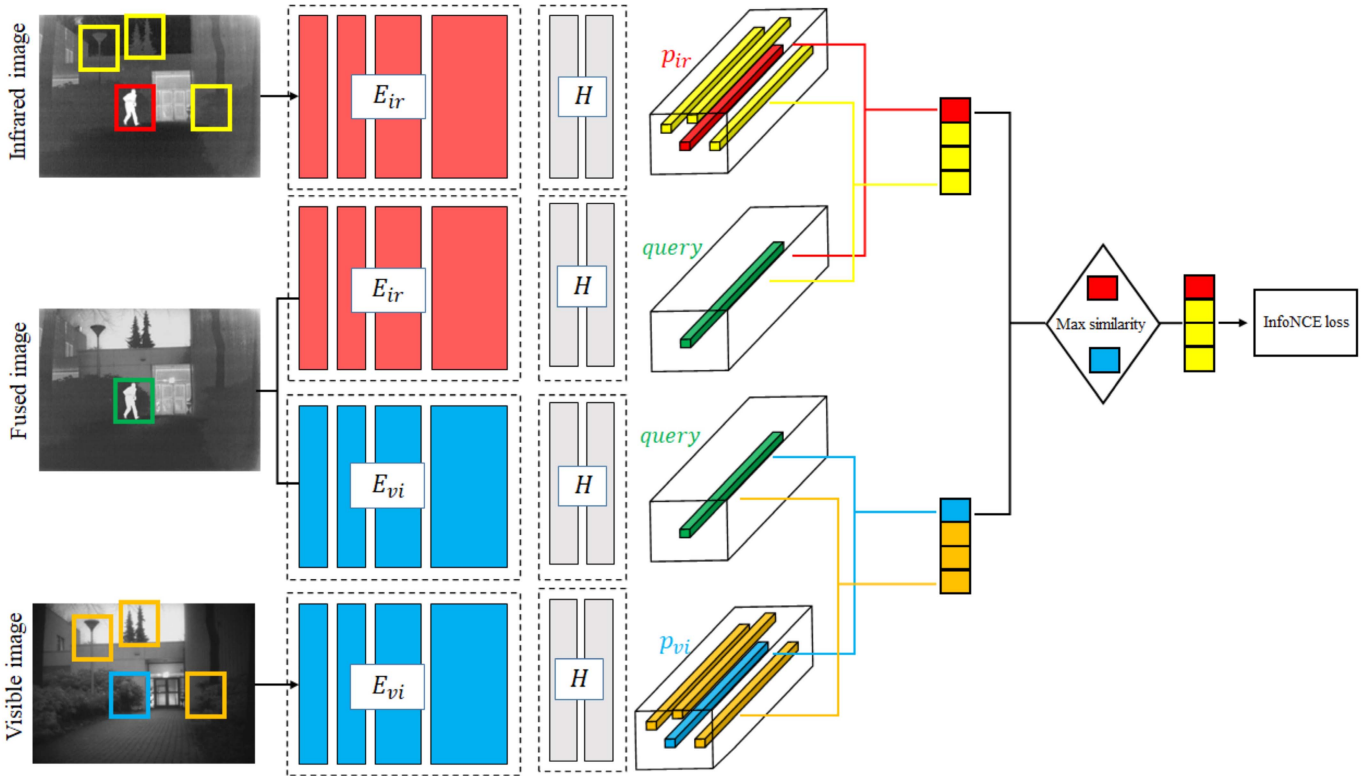


Fig. 2. Adaptive patchwise contrastive learning: we randomly sample a query patch from the fused image and select the positive patch from the infrared image and visible image at the same position (i.e., the green, red, and blue boxes). Next, N random negative patches are selected from other positions of the infrared image and visible image (i.e., the yellow and orange boxes). Then, we reuse the infrared encoder and visible encoder and add the two-layer MLP network, in which the positive patches in the fusion image and source images will be encoded into feature vectors query, p_{ir}, and p_{vi}. Finally, the similarities between query and p_{ir} or query and p_{vi} are calculated, and the most similar one will be retained to calculate the InfoNCE loss.

Based on the idea of NCE and introducing the concept of MI, a new form of contrastive loss function called InfoNCE [36] is proposed. Specifically, we assume that there is an encoded query and a set of encoded samples $\{k_1^-, k_2^-, \ldots, k_N^-\}$, including a positive example and N negative examples. The query, positive example, and N negative examples are mapped into the K-dimensional vectors q, $k^+ \in \mathrm{R}^K$, and $k^- \in \mathrm{R}^{N\times K}$, respectively, where $k_n^- \in \mathrm{R}^K$ denotes the nth negative example. When q is similar to the positive example k^+ and dissimilar to all other negative examples k^-, the value of the InfoNCE loss will be small. The similarity is measured by the dot product between the l_2 normalized query and other examples. This result is then scaled by a temperature τ and passed as logits. The InfoNCE loss is defined as follows:

$$l\left(q, k^+, k^-\right) = -\log\left[\frac{\exp\left(q \cdot k^+/\tau\right)}{\exp(q \cdot k^+/\tau) + \sum_{n=1}^{N} \exp\left(q \cdot k_n^-/\tau\right)}\right]. \tag{1}$$

Based on the above InfoNCE framework, several important design ideas, including how to design the contrastive learning structure and how to build the specific loss function, are presented in Sections III-C and III-D.

C. Adaptive Patchwise Contrastive Learning

In the general contrastive learning methods, data augmentation is often used to establish a positive pair for positive samples, and $n-1$ negative pairs are established by using all $n-1$ other images in the same training batch with the augmented images of positive samples. Then, the similarity between positive pairs is maximized, and the similarity between negative pairs is minimized to fully extract the general features of the unlabeled datasets. However, some studies [37] have shown that the more negative pairs there are, the better the contrastive learning effect. This requires the support of abundant training datasets. Obviously, for image fusion tasks, the lack of sufficient training datasets has always been an urgent problem to be solved. Combined with the characteristics of the image fusion tasks, we construct an unsupervised patchwise contrastive learning framework based on the work by Park *et al.* [34].

Since the image fusion task focuses more on the salient target of the infrared image and the background texture information of the visible image, we start from the local features of the image to construct a contrastive learning task based on the image patches.

Specifically, as shown in Fig. 2, we randomly sample a patch of the fused image and a positive patch of the infrared image and visible image at the same position (i.e., the green, red, and blue boxes). Next, N random negative patches are selected from other positions of the infrared image and visible image (i.e., the yellow and orange boxes). Then, we reuse the infrared and visible encoders and add the two-layer multilayer perceptron (MLP) network, which is used to encode patches at

any spatial location in the source images and fusion image as feature vectors. For example, the positive patches in the fused image and source images are encoded as feature vectors query, p_{ir}, and p_{vi}. Finally, the similarities between the query and p_{ir} or the query and p_{vi} are calculated and the most similar one will be retained to calculate the InfoNCE loss.

It is worth noting that both positive and negative samples used to calculate the InfoNCE loss are sampled from the source image and fusion image encoded by the same encoder. For the selection strategy of negative samples, we will elaborate on the extending experiment.

D. Loss Function

In this section, we discuss the calculation of the loss function combined with SSIM and patchNCE, which is used to guide the CNN network in finding the most appropriate parameters through unsupervised learning. The SSIM loss mainly focuses on the structural characteristics of the image itself, while the patchNCE loss mainly focuses on the deep features of the image.

The SSIM combines image brightness, contrast, and structure to measure image quality [30]. For any two images, the SSIM is described as follows:

$$\text{SSIM}(A, B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)}. \quad (2)$$

We set $C_1 = 1 \times 10^{-4}$ and $C_2 = 9 \times 10^{-4}$, which are the same values as in [39]. According to the above parameter settings, as suggested in [18], we set the SSIM loss as follows:

$$E(I|W) = \frac{1}{m \times n} \sum_{i=1}^{m \times n} P_i \quad (3)$$

$$\text{Score}(I_f, I_{\text{ir}}, I_{\text{vi}}|W) = \begin{cases} \text{SSIM}(I_f, I_{\text{ir}}|W) \\ \text{if } E(I_{\text{ir}}|W) > E(I_{\text{vi}}|W) \\ \text{SSIM}(I_f, I_{\text{vi}}|W) \\ \text{if } E(I_{\text{ir}}|W) \le E(I_{\text{vi}}|W) \end{cases} \quad (4)$$

$$L_{\text{SSIM}} = 1 - \frac{1}{N} \sum_{W=1}^{N} \text{Score}(I_f, I_{\text{ir}}, I_{\text{vi}}|W) \quad (5)$$

where W represents the sliding window from the top left to the bottom right with a stride of 1, P_i represents the value of pixel i, m and n represent the size of the sliding window, and N represents the number of sliding windows in a single image. The size of the window is 16×16 in our work.

Above, we discussed the SSIM loss function. On the one hand, based on the average intensity of pixels in the local window, the SSIM loss can not only retain salient targets in the infrared image but also retain bright areas and some conspicuous textures in the visible image. On the other hand, the SSIM loss can use the shallow feature of image structure to ensure the structural consistency of input and output. For the feature extraction network, we expect that the infrared encoder can retain more salient target features and the visible encoder can retain more detailed texture features, which is the most obvious complementary features between infrared and visible images. Thus, we introduce a novel contrastive loss function, which directly uses the encoded deep representations to promote the encoder to retain sufficient complementary information.

Specifically, based on the adaptive patchwise contrastive learning framework mentioned above, we can construct the contrastive loss as follows. First, since the two encoders E_{ir} and E_{vi} used in our image fusion task can extract an effective feature stack, we can make use of them. At the same time, we pass the feature maps through a small neural network projection head H, which is a two-layer MLP. The infrared image and visible image are encoded by corresponding encoders, and the fusion image is encoded by two encoders; thus, there are four feature sequences that can be obtained

$$\begin{cases} z_{\text{ir}} = H(E_{\text{ir}}(I_{\text{ir}})) \\ z_{\text{vi}} = H(E_{\text{vi}}(I_{\text{vi}})) \\ z_{f\_\text{ir}} = H\big(E_{\text{ir}}(I_f)\big) \\ z_{f\_\text{vi}} = H\big(E_{\text{vi}}(I_f)\big). \end{cases} \quad (6)$$

We denote $s \in \{1, \ldots, S\}$, where S is the number of the spatial locations sampled from the last image feature layer. For any specific spatial location in the image feature level, we refer to the patch feature as $z^s \in \text{R}^C$ and the remaining features in the same feature level as $z^{S/s} \in \text{R}^{(S-1)\times C}$, where C is the number of channels. As shown in Fig. 2, the patchNCE loss of any specific spatial location can be obtained as shown in (7), and then, the contrastive loss can be obtained as shown in (8)

$$l^s_{\text{patchNCE}}(Z^s) = \begin{cases} l\Big(z^s_{f\_\text{ir}}, z^s_{\text{ir}}, z^{S/s}_{\text{ir}}\Big) \\ if\big(z^s_{f\_\text{ir}} \cdot z^s_{\text{ir}}\big) > \big(z^s_{f\_\text{vi}} \cdot z^s_{\text{vi}}\big) \\ l\Big(z^s_{f\_\text{vi}}, z^s_{\text{vi}}, z^{S/s}_{\text{vi}}\Big) \\ if\big(z^s_{f\_\text{ir}} \cdot z^s_{\text{ir}}\big) \le \big(z^s_{f\_\text{vi}} \cdot z^s_{\text{vi}}\big) \end{cases} \quad (7)$$

$$L_{\text{patchNCE}} = \text{E}_{x\sim X} \sum_{s=1}^{S} l^s_{\text{patchNCE}}(Z^s) \quad (8)$$

where Z^s is a general term for the set $\{z^s_{f\_\text{ir}}, z^s_{f\_\text{vi}}, z^s_{\text{ir}}, z^s_{\text{vi}}\}$.

Above, we discussed the calculation of the contrastive loss function. This loss function focuses more on the deep representations, which are extracted by encoders. As the training process progresses, the patchNCE loss can effectively adjust the encoder and projection head to retain the most similar parts of the source images to the fused image.

Based on the above two loss functions, the total loss function can be defined as

$$L = \lambda_1 L_{\text{patchNCE}} + \lambda_2 L_{\text{SSIM}} \quad (9)$$

where λ_1 and λ_2 are the hyperparameters that control the loss balance between the two loss functions.

In general, the SSIM loss maintains the structural consistency between input and output, while the patchNCE loss maintains the consistency of the deep features of input and output. The two loss functions complement each other and guide the network to achieve satisfactory results.

IV. Experiments

In this section, we first elaborate on the experimental settings that are used in our work, including the datasets, evaluation metrics, training details, and discussion of training hyperparameters. Then, we compare the proposed method with nine other popular methods, including the DenseFuse [9], RFN-Nest [17], FusionGAN [11], GANMcC [13], IFCNN [14], PMGI [15], U2Fusion [16], STDFusionNet [19], and IFSepR [47], on the TNO dataset, RoadScene dataset, medical images, and multifocus images. Next, we provide an additional ablation experiment and efficiency evaluation experiment to further verify the performance of the proposed methods. Finally, we discuss the negative sample selection in the adaptive patchwise contrastive learning framework.

A. Experimental Settings

1) Datasets: All training and testing datasets that are used come from the TNO dataset [40] and the RoadScene dataset [16].

The TNO dataset mainly describes various military-related scenes and is the most commonly used dataset in infrared and visible image fusion tasks. In addition, the RoadScene dataset was published based on FLIR videos, in which a large number of road scenes, including roads, vehicles, and pedestrians, are described. The TNO dataset contains 60 infrared and visible image pairs and three video sequence screenshots, while the RoadScene dataset contains 221 infrared and visible image pairs. These image pairs play an important role in the training and verification of the model.

2) Evaluation Metrics: The evaluation of the fusion performance includes a subjective evaluation and an objective evaluation. The subjective evaluation is based on peoples' visual perception; usually, the fusion image containing salient infrared targets and rich texture information has the best effect. The objective evaluation is a measure of the fusion performance using quantitative metrics. In this article, six popular metrics are selected, including entropy (EN) [41], MI [42], VIF [43], standard deviation (SD) [44], average gradient (AG) [45], and spatial frequency (SF) [46]. EN measures the amount of information contained in a fused image based on the information theory. The MI measures the dependence of the source images and fused images. VIF measures the information fidelity of the fused result by calculating the distortion of the images, which is consistent with the human visual system. The SD can reflect the distribution and contrast of the fused image, which is based on the statistical concept. The AG quantifies the gradient information of the fused image and the SF measures the gradient distribution of the fused image. Both AG and SF reveal the detail and texture information.

3) Training Details: We use the TNO dataset to train our model. Twenty image pairs are selected and the training data are expanded through cropping. We use a sliding window of 128×128 to crop the image into small image patches, and the sliding step is set to 32. Finally, a total of 4404 image patch pairs are obtained. We select 20 image pairs from the TNO dataset for the comparative experiment. To adequately explore the generalization power of our method, we select 20 image pairs from the RoadScene dataset for the generalization experiment. Each source image is initially normalized to $[-1, 1]$. The MLP consists of a linear layer with an output size of 2048, followed by batch normalization, a LeakyReLU, and a final linear layer with an output dimension of 256, as in BYOL [31]. In the training process, the training parameters are set as follows: the batch size, max epoch, learning rate, and temperature τ are initialized as 4, 20, 2×10^{-3}, and 0.07, respectively. In addition, the proposed algorithm is implemented on the PyTorch platform and all the experiments are conducted on an NVIDIA GeForce RTX 2070 super GPU and Intel i7-10875H CPU.

TABLE I

Discussion of Training Hyperparameters on Six Metrics and Training Time. Red Represents the Best Result, Blue Represents the Second Best Result, and Bold Represents the Third Best Result

S	λ_1/λ_2	EN [41]	MI [42]	VIF [43]	SD [44]	AG [45]	SF [46]	Time
200	100	7.0205	4.2142	**1.0439**	9.7641	3.9440	0.0372	17 38'
	10	7.0718	4.4776	1.1035	9.4760	3.7830	0.0359	17 41'
	1	7.2262	3.6359	1.0484	9.6135	4.9897	0.0495	17 38'
	0.1	**7.1150**	3.8393	1.0305	9.3776	**4.4037**	0.0452	17 43'
	0.01	7.0448	3.8064	0.9989	9.3136	4.1845	0.0431	17 42'
100	1	7.1681	**3.8174**	1.0437	9.4785	4.5395	**0.0453**	8 59'
300	1	7.0178	3.7795	0.9873	**9.5160**	4.2921	0.0455	26 27'

4) Discussion of Training Hyperparameters: There are some hyperparameters that directly affect the final model performance, including the number of samples S in (7) and λ_1 and λ_2 in (9). It is worth noting that the number of samples S is related to the number of negative samples N. Specifically, the positive sample comes from a patch in a certain source image, while the negative samples not only come from other patches in the same source image but also from other source images in the same training batch. Therefore, the number of negative samples $N = S - 1 + (B - 1) \times S$, where B is the number of batch size.

We analyze the influences of the selection of different hyperparameters from the quantitative metrics and training time of an epoch, which is shown in Table I.

Hou *et al.* [18] argued that: when the weight of L_{SSIM} in the loss function is relatively low, this leads to low contrast and low quality in the fused image. In contrast, when the weight of L_{SSIM} in the loss function is relatively high, visible details are lost to a certain degree.

As shown in Table I, the selection of different proportions of λ_1 and λ_2 has a direct impact on quantitative metrics. Specifically, when λ_1/λ_2 is greater than 1, MI, VIF, and SD are better, which indicates that L_{patchNCE} is more helpful in improving the fusion image contrast and making it more consistent with human visual effects. When λ_1/λ_2 is less than 1, EN, AG, and SF are better, which indicates that the fused image contains more texture details. It seems that this is contrary to the conclusion in VIF-Net. Our perspective is that even if the weight of L_{patchNCE} is low, it can make up for the deficiency of L_{SSIM} and retain visible details, which can also be proved in the ablation experiment. Therefore,

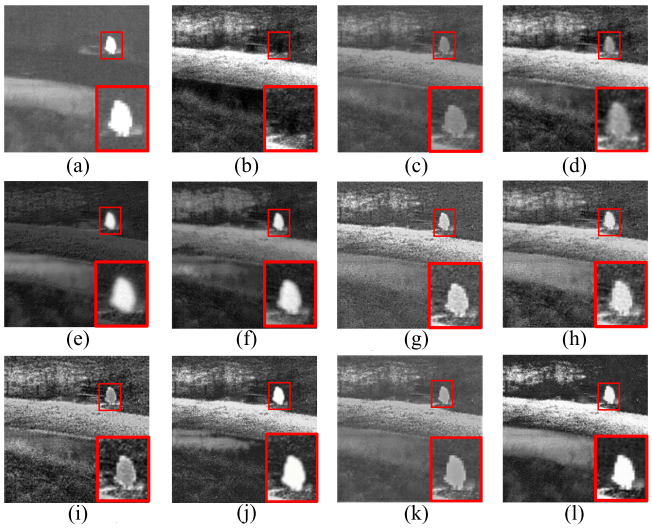


Fig. 3. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on bench. We have selected the salient object (i.e., the red box) and zoomed in on it in the bottom-right corner for ease of comparison. The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

to combine the advantages of the two types of loss functions, the hyperparameters λ_1 and λ_2 are set as 1 in this article.

In addition, according to Table I, it can be found that with the increase of sampling number of S, the training time increases significantly. Thus, a larger number of samples have a greater burden on training efficiency and equipment. In addition, from the perspective of quantitative metrics, too low or too high sample number will make the model poor. Too little samples will make the model unable to better distinguish positive samples and negative samples, while too much samples will make it more likely that the negative samples contain more samples close to positive samples. For example, if multiple samples are collected on the thermal infrared target, these samples can actually be regarded as the same class and taking most of them as negative samples will affect the model performance. Therefore, the hyperparameter S selected in this article is 200.

B. Comparative Experiment

For a comprehensive analytical evaluation of our approach, we compare our proposed CLF-Net with eight other approaches on the TNO dataset.

1) Qualitative Results: To intuitively compare the performance of different algorithms, we select four typical image pairs from the TNO dataset (bench, Nato_camp_1811, Kaptein_1123, and 2_men_in_front_of_house). The qualitative comparison results are shown in Figs. 3–6. In Figs. 3 and 4, we select the salient object (i.e., the red box) and zoom in it in the bottom corner for ease of comparison. As shown in Fig. 3, the infrared target information is lost when using DenseFuse, and obvious thermal radiation targets are not captured, while all relatively obvious thermal radiation targets can be captured by NestFuse, IFCNN, PMGI, and U2Fusion. However, noise interference from the visible images affects the results to varying degrees,

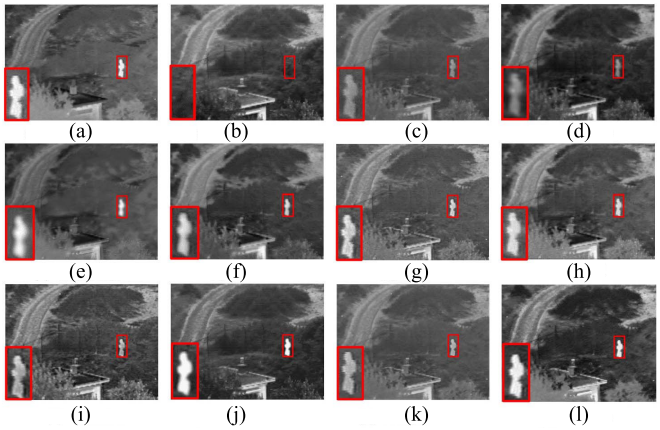


Fig. 4. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on Nato_camp_1811. We have selected the salient object (i.e., the red box) and zoomed in on it in the bottom-right corner for ease of comparison. The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

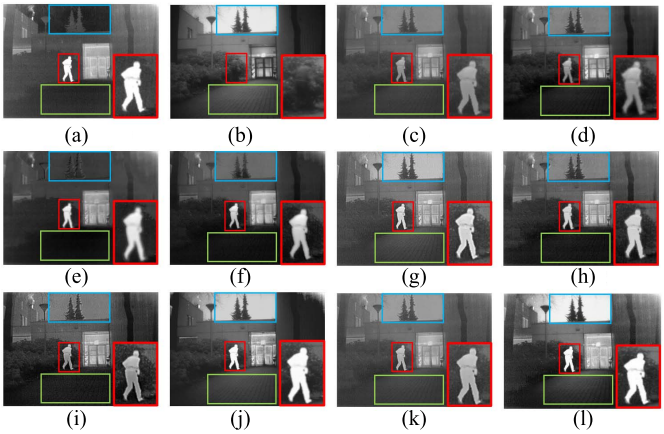


Fig. 5. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on Kaptein_1123. We have selected the salient object (i.e., the red box) and zoomed in on it in the bottom-right corner for ease of comparison. In addition, we have selected the two background areas of sky and ground (i.e., the blue box and green box, respectively). The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

among which the most seriously impacts results are obtained using U2Fusion. In addition, conspicuous targets with high contrast can be captured using FusionGAN and GANMcC. However, the edge of the target is fuzzy, which affects target recognition. In contrast, when using STDFusionNet and CLF-Net, the most prominent infrared targets of the highest quality can be captured. However, compared with STDFusionNet, the edge of the salient target in CLF-Net is clearer and closer to the infrared image. In Fig. 4, the best performance for the fusion of salient objects is still achieved using our method.

In Figs. 5 and 6, we not only select the salient object (i.e., the red box) and zoom in on it in the bottom-right corner, but we also select the background areas. As shown in Fig. 5, first, for salient targets, enough information is retained using CLF-Net to obtain infrared targets with clear edges and high

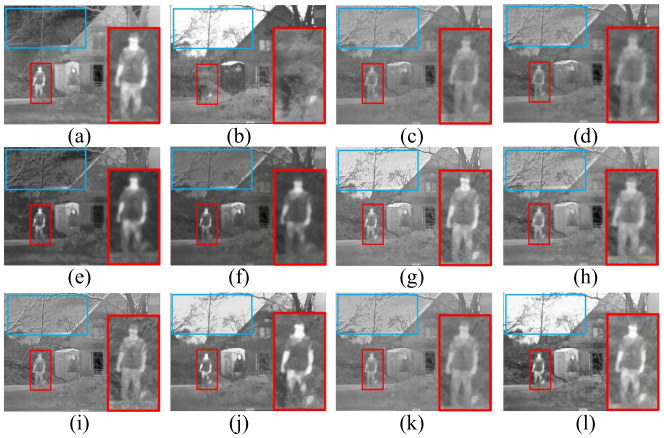


Fig. 6. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on 2_men_in_front_of_house. We have selected the salient object (i.e., the red box) and zoomed in on it in the bottom-right corner for ease of comparison. In addition, we have selected the background areas of the sky (i.e., the blue box). The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

contrast. In addition, the texture information in the background areas of the ground in the visible images can be retained (i.e., the green box) when using DenseFuse, NestFuse, IFCNN, STDFusionNet, and CLF-Net. However, for other backgrounds in the image, such as the sky (i.e., the blue box), the five algorithms perform differently. In CLF-Net, there is minimal disturbance from the thermal infrared information, and the brightest sky, which is the same as the sky in the visible image, is obtained.

Through the above comparison experiments, it is found that a sufficient amount of texture information of the visible images can be adaptively retained in CLF-Net, while the clearest infrared salient target is extracted. This indicates that the fused images generated by our method have excellent subjective visual effects.

2) Quantitative Results: To quantitatively analyze our method and eight other algorithms, 20 image pairs are selected from the TNO dataset for testing. The results of six general quantitative metrics are shown in Fig. 7 and Table II. Among the six metrics, the best performance in terms of EN, VIF, and SD is achieved using our method, and significant advantages in terms of VIF are observed. In addition, comparable performance on the MI, AG, and SF metrics is achieved using our method. For the MI metric, except for STDFusionNet, the best performance is achieved using our method compared with other algorithms. For the AG metric, our method follows behind U2Fusion and IFCNN, and for the SF metric, our method follows behind IFCNN by only a narrow margin.

As shown in Fig. 7, the highest value on almost all image pairs on the VIF metric is obtained using our method. The VIF metric is consistent with the human visual system. By obtaining the highest value on the VIF metric, it is demonstrated that our algorithm has a better human visual effect, which is consistent with the results obtained in the qualitative analysis. The larger EN is, the more information that is contained in the fused image. The largest amount of information is contained in

Fig. 7. Quantitative comparison results of CLF-Net and nine state-of-the-art methods on 20 images from the TNO dataset. Six metrics are used for comparison: EN, MI, VIF, SD, AG, and SF. The compared methods are DenseFuse, RFN-Nest, FusionGAN, GANMcC, IFCNN, PMGI, U2Fusion, STDFusionNet, IFSepR, and CLF-Net.

TABLE II

QUANTITATIVE COMPARISON RESULTS OF CLF-NET AND NINE STATE-OF-THE-ART METHODS ON 20 IMAGES FROM THE TNO DATASETS. SIX METRICS ARE USED FOR COMPARISON: EN, MI, VIF, SD, AG, AND SF. RED REPRESENTS THE BEST RESULT, BLUE REPRESENTS THE SECOND BEST RESULT, AND BOLD REPRESENTS THE THIRD BEST RESULT

	TNO					
	EN [41]	MI [42]	VIF[43]	SD [44]	AG [45]	SF [46]
DenseFuse [9]	6.3989	2.1046	0.6342	8.5050	2.7006	0.0266
RFN-Nest [17]	6.9657	2.0058	0.7614	9.1554	2.8707	0.0246
FusionGAN [11]	6.5718	2.2516	0.6119	8.6336	2.5011	0.0259
GANMcC [13]	6.7397	2.1975	0.6482	8.8980	2.6367	0.0247
IFCNN [14]	6.8325	**2.3595**	0.7720	8.9772	5.0494	0.0501
PMGI [15]	7.0370	2.2342	**0.7993**	9.5324	3.7842	0.0359
U2Fusion [16]	6.9903	1.7694	0.7401	**9.4385**	5.3138	0.0479
STDFusionNet[19]	**7.0225**	4.1589	0.9510	9.2720	4.6846	**0.0492**
IFSepR [47]	6.5597	2.0902	0.6836	8.7226	3.5794	0.0370
CLF-Net	7.2262	3.6359	1.0484	9.6135	**4.9897**	0.0495

the fused image of our CLF-Net. The best performance on SD, which reflects a result with high contrast and is also consistent with the results of qualitative analysis, is achieved using our proposed method. The larger MI is, the more the information transfers from the source images to the fused image, which indicates that our method retains a large amount of information

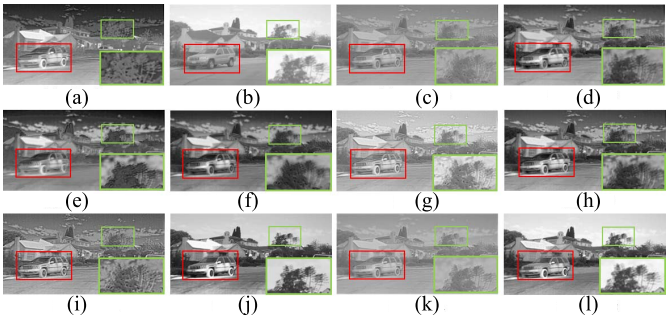


Fig. 8. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on FLIR_00993. We select a background area of the tree (i.e., the green box) and zoom in it in the bottom right corner for ease of comparison and mark the salient road scene object (i.e., the red box). The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

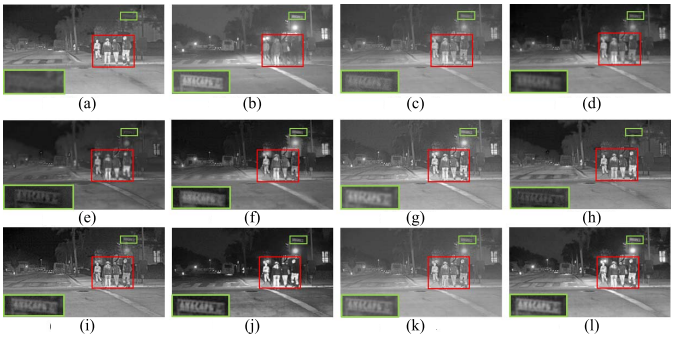


Fig. 9. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on FLIR_03952. We select a background area of the banner (i.e., the green box) and zoom in it in the bottom-left corner for ease of comparison and select the salient road scene object (i.e., the red box). The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

from the infrared and visible images. In addition, both SF and AG can reveal the details and textures, respectively, of the fused image. Although the SF and AG metrics using our methods are not the best, the comparable results still mean that the fused images obtained using our method contain adequate gradient information.

C. Generalization Experiment on RoadScene Dataset

To explore the generalization capability of our method, we compare our proposed CLF-Net with eight other methods on the RoadScene dataset.

1) Qualitative Results: We select four typical image pairs from the RoadScene dataset for analysis. As shown in Figs. 8–11, we select the salient targets of typical road scenes with red boxes (i.e., the cars and people). In the comparison with other methods, the salient object of the fused image generated by CLF-Net has a clear edge contour and the highest contrast, which maximizes the retention of thermal targets in the infrared images. Due to the unique imaging mode of the infrared image, it is difficult to distinguish the time or weather in the infrared image because the sky of the infrared image is always dark regardless of the time of day. In Figs. 8, 10, and 11, the scenes were all taken during the day. However, it is difficult to estimate whether the scene

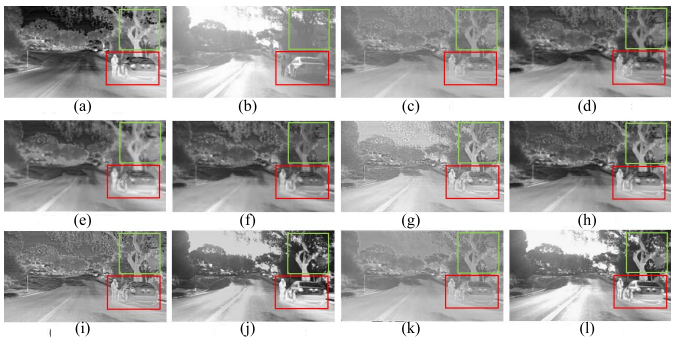


Fig. 10. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on FLIR_04302. We select a background area of the bole (i.e., the green box) and the salient road scene object (i.e., the red box). The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

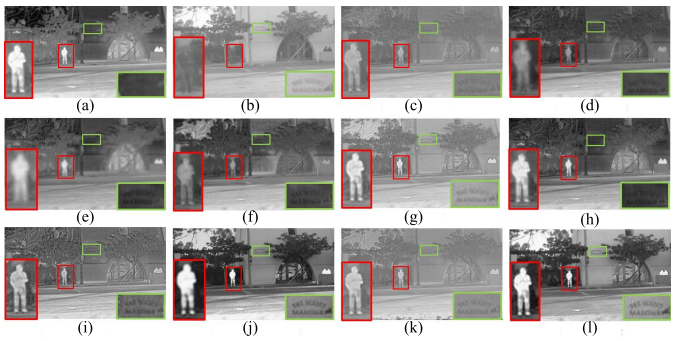


Fig. 11. Qualitative comparison results of CLF-Net and eight state-of-the-art methods on FLIR_04598. We select a background area of the writing on wall (i.e., the green box) and mark the salient road scene object (i.e., the red box) and zoom in them in the bottom corner for ease of comparison. The first two images in the first row are (a) infrared image and (b) visible image. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

represents day or night from the resulting fusion image, except when using IFCNN, STDFusionNet, and CLF-Net. Among these methods, the best performance is achieved using STDFusionNet and CLF-Net. Meanwhile, compared with STDFusionNet, CLF-Net has the brightest sky background and is closest to the visible images. In addition, we select some background areas, such as the tree, banner, and writing on the wall, with green boxes. Through comparison, the most detailed texture information of visible images is retained when using our method. In Fig. 10, it is worth noting that the sunlight makes the visible images taken by the camera appear slightly overexposed, resulting in blurred details. However, the influence of overexposure is effectively reduced in the fusion image generated by our method.

2) Quantitative Results: Twenty image pairs from the RoadScene dataset are selected for quantitative evaluation, and the quantitative comparison results are shown in Fig. 12 and Table III. Specifically, the largest average values in terms of EN, MI, VIF, AG, and SF are obtained using our method. For the SD metric, our proposed algorithm has a comparable performance with the NestFuse and STDFusionNet method by a narrow margin. In general, good results in both qualitative

Fig. 12. Quantitative comparison results of CLF-Net and nine state-of-the-art methods on 20 images from the RoadScene dataset. Six metrics are used for comparison: EN, MI, VIF, SD, AG, and SF. The compared methods are DenseFuse, RFN-Nest, FusionGAN, GANMcC, IFCNN, PMGI, U2Fusion, STDFusionNet, IFSepR, and CLF-Net.

TABLE III

QUANTITATIVE COMPARISON RESULTS OF CLF-NET AND NINE STATE-OF-THE-ART METHODS ON 20 IMAGES FROM THE ROADSCENE DATASETS. SIX METRICS ARE USED FOR COMPARISON: EN, MI, VIF, SD, AG, AND SF. RED REPRESENTS THE BEST RESULT, BLUE REPRESENTS THE SECOND BEST RESULT, AND BOLD REPRESENTS THE THIRD BEST RESULT

	RoadScene					
	EN [41]	MI [42]	VIF[43]	SD [44]	AG [45]	SF [46]
DenseFuse [9]	6.6546	2.8861	0.6291	9.4781	3.1148	0.0319
RFN-Nest [17]	7.2604	2.7260	0.7049	10.0316	3.3159	0.0306
FusionGAN [11]	7.0384	2.8801	0.5674	**10.2825**	3.0991	0.0318
GANMcC [13]	7.2160	2.7471	0.6739	10.2207	3.5105	0.0335
IFCNN [14]	6.9926	2.9782	0.7290	10.2938	5.4211	0.0570
PMGI [15]	**7.3049**	**3.3469**	**0.7676**	10.0579	4.2302	0.0409
U2Fusion [16]	7.1546	2.7608	0.6806	9.9800	5.9251	**0.0582**
STDFusionNet[19]	7.4365	4.7425	0.9989	10.5826	**5.8935**	0.0652
IFSepR [47]	6.7080	2.6837	0.6107	9.7581	4.0815	0.0498
CLF-Net	7.4540	5.1649	1.0838	10.4617	6.0765	0.0683

analysis and quantitative analysis are achieved using our method, indicating that our model has good generalization ability.

D. Generalization Experiment on Medical Image

In this section, we have compared the image fusion methods on medical images. The medical images for the experiment are collected from [48] and include 24 pairs of images.

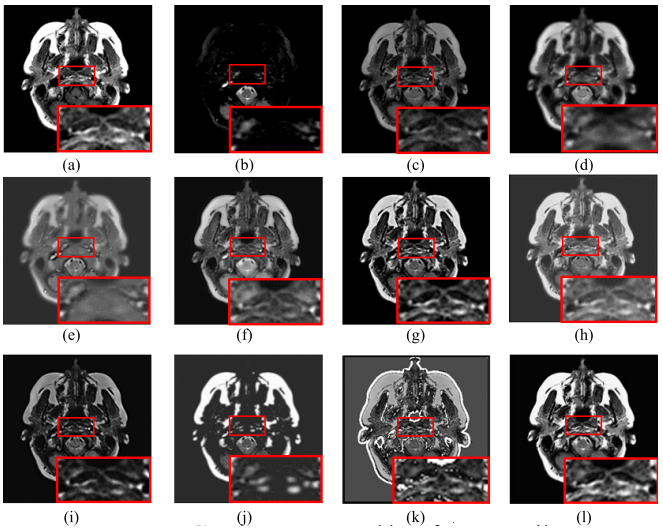


Fig. 13. Qualitative comparison results of CLF-Net and nine state-of-the-art methods on medical image. We have selected the significant region of MR-T1 (i.e., the red boxes) and zoomed in on it in the bottom-right corner for ease of comparison. The first two images in the first row are (a) MR-T1 and (b) MR-T2. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

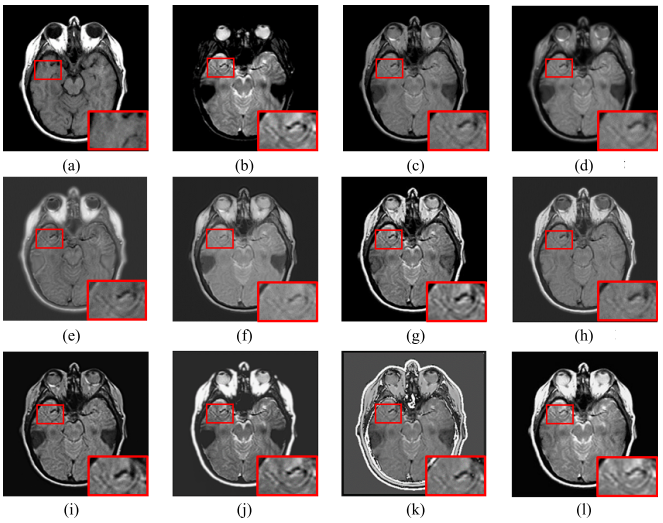


Fig. 14. Qualitative comparison results of CLF-Net and nine state-of-the-art methods on medical image. We have selected the texture details of MR-T2 (i.e., the red box) and zoomed in it in the bottom-right corner for ease of comparison. The first two images in the first row are (a) MR-T1 and (b) MR-T2. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

1) Qualitative Results: We select MR-T1 and MR-T2 as source images. MR-T1 contains bright skull features and MR-T2 contains rich texture information. Two typical image pairs from [48] are shown in Figs. 13 and 14. We select the significant region of MR-T1 in Fig. 13 and the texture detail of MR-T2 in Fig. 14 with red boxes. In Fig. 13, different fusion algorithms have inconsistent effects on bright skull features in MR-T1. Among them, STDFusionNet and our CLF-Net have the best retention effect for bright skull area. The images generated by IFSepR have some distortion but contain rich texture details. For the marked significant region, RFN-Nest

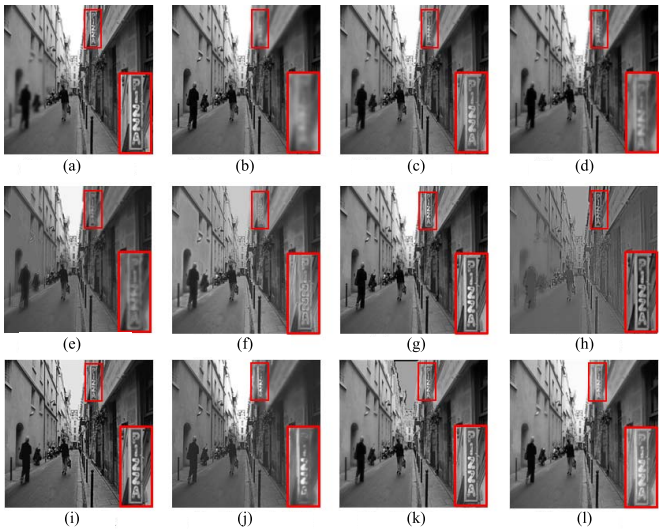


Fig. 15. Qualitative comparison results of CLF-Net and nine state-of-the-art methods on multifocus images. We have selected the blurred area (i.e., the red box) and zoomed in it in the bottom-right corner for ease of comparison. The first two images (a) and (b) in the first row are two different multifocus images in the same scene. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR, and (l) CLF-Net.

TABLE IV

QUANTITATIVE COMPARISON RESULTS OF CLF-NET AND NINE STATE-OF-THE-ART METHODS ON 24 MEDICAL IMAGES. SIX METRICS ARE USED FOR COMPARISON: EN, MI, VIF, SD, AG, AND SF. RED REPRESENTS THE BEST RESULT, BLUE REPRESENTS THE SECOND BEST RESULT, AND BOLD REPRESENTS THE THIRD BEST RESULT

	Medical Image					
	EN [41]	MI [42]	VIF[43]	SD [44]	AG [45]	SF [46]
DenseFuse [9]	3.8466	2.9867	0.6885	8.9087	4.0286	0.0557
RFN-Nest [17]	4.5150	2.3400	0.6031	9.0241	4.0271	0.0463
FusionGAN [11]	4.4461	2.1360	0.4726	9.3438	3.3050	0.0387
GANMcC [13]	4.5014	2.4103	0.5568	**9.1036**	4.4184	0.0575
IFCNN [14]	4.2599	2.5167	0.7121	8.9105	7.2057	0.1041
PMGI [15]	**4.6867**	2.5797	**0.7066**	9.1562	4.8224	0.0601
U2Fusion [16]	4.3619	2.3839	0.5392	8.3771	5.4786	0.0758
STDFusionNet[19]	4.0163	**2.6296**	0.6788	8.2683	5.6081	0.0726
IFSepR [47]	4.9990	2.4846	0.1763	8.6086	11.7412	0.2052
CLF-Net	4.8480	2.8499	0.8314	9.0204	**7.1763**	**0.1040**

and FusionGAN have almost lost the features of this region. In contrast, IFCNN, PMGI, and CLF-Net retain the features of the significant region well, which are closer to the source image.

In Fig. 14, the analysis of the effect of the bright skull area is consistent with that in Fig. 13. In addition, for the orbital region, MR-T2 is brighter than MR-T1, only GANMcC, IFCNN, STDFusionNet, and CLF-Net retain the brightness of MR-T2 better. Finally, compared with Fig. 13, the MR-T2 image of Fig. 14 contains richer and more useful texture information. DenseFuse, FusionGAN, GANMcC, and PMGI retain less texture information, and IFCNN, U2Fusion, STDFusionNet, IFSepR, and CLF-Net retain more texture information.

2) Quantitative Results: The quantitative comparison results are shown in Table IV and Fig. 17. AG and SF reflect

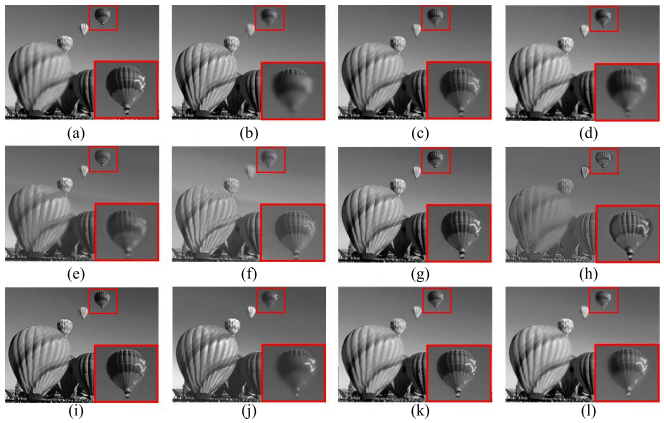


Fig. 16. Qualitative comparison results of CLF-Net and nine state-of-the-art methods on multifocus images. We have selected the blurred area (i.e., the red box) and zoomed in it in the bottom-right corner for ease of comparison. The first two images (a) and (b) in the first row are two different multifocus images in the same scene. These are followed by (c) DenseFuse, (d) RFN-Nest, (e) FusionGAN, (f) GANMcC, (g) IFCNN, (h) PMGI, (i) U2Fusion, (j) STDFusionNet, (k) IFSepR and (l) CLF-Net.

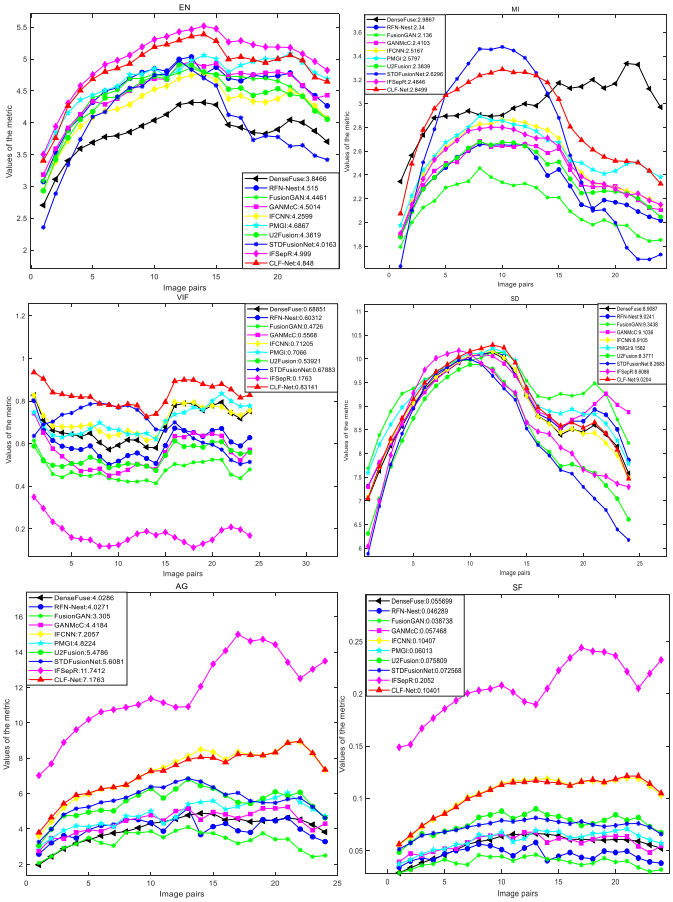


Fig. 17. Quantitative comparison results of CLF-Net and nine state-of-the-art methods on 24 pairs of medical images. Six metrics are used for comparison: EN, MI, VIF, SD, AG, and SF. The compared methods are DenseFuse, RFN-Nest, FusionGAN, GANMcC, IFCNN, PMGI, U2Fusion, STDFusionNet, IFSepR, and CLF-Net.

the texture information of images, and IFSepR, CLF-Net, and IFCNN are the top three in these two metrics, followed by U2Fusion and STDFusionNet, indicating that these algorithms better retain rich texture information in the image, which

TABLE V

QUANTITATIVE COMPARISON RESULTS OF CLF-NET AND NINE STATE-OF-THE-ART METHODS ON 36 MULTIFOCUS IMAGES. SIX METRICS ARE USED FOR COMPARISON: EN, MI, VIF, SD, AG, AND SF. RED REPRESENTS THE BEST RESULT, BLUE REPRESENTS THE SECOND BEST RESULT, AND BOLD REPRESENTS THE THIRD BEST RESULT

	Multi-focus Image					
	EN [41]	MI [42]	VIF[43]	SD [44]	AG [45]	SF [46]
DenseFuse [9]	7.2860	10.1347	1.7746	10.2895	6.0100	0.0748
RFN-Nest [17]	7.4388	5.7572	1.6765	10.3452	4.3296	0.0430
FusionGAN [11]	7.0896	5.8655	1.2374	9.7979	5.3408	0.0643
GANMcC [13]	7.0103	5.5935	1.1978	10.7624	6.1092	0.0731
IFCNN [14]	7.3324	6.3636	1.7750	10.3278	**7.2626**	0.0919
PMGI [15]	6.6468	3.9707	0.7788	9.3841	6.5229	0.0800
U2Fusion [16]	**7.4039**	5.8298	1.6015	10.5110	8.3527	0.0940
STDFusionNet[19]	7.1489	**6.9463**	1.4562	9.4457	6.5482	0.0828
IFSepR [47]	7.3363	8.3457	1.6208	**10.4764**	7.1306	**0.0935**
CLF-Net	7.4048	6.2178	**1.6919**	10.2464	7.3844	0.0936

is consistent with qualitative analysis. In the VIF metric, CLF-Net, IFCNN, and PMGI are the top three algorithms, indicating that the images obtained by these algorithms have a better human visual effect. In addition, IFSepR has the lowest VIF, which is related to image distortion. Finally, our algorithm also has good performance in EN and MI, which demonstrates its great ability in transferring information.

Medical image fusion task is similar to infrared and visible image fusion task to some extent. Based on the above qualitative and quantitative analysis, our proposed method has good application potential in the medical image fusion task.

E. Generalization Experiment on Multifocus Image

In this section, we evaluate our method in the field of multifocus image fusion. The multifocus images for the test are collected from [49] and include 36 pairs of images in different scenarios.

1) Qualitative Results: The source images are derived from the same image, which are blurred in different nonoverlapping regions. Two typical image pairs from [49] are shown in Figs. 15 and 16. As for the overall image fusion effect, the distortion of PMGI fusion results is serious, which contains a large amount of noise. IFSepR can fuse the blurred region, but the distortion occurs in the sky. The fusion performance of RFN-Nest, FusionGAN, GANMcC, and STDFusionNet for the blurred region is poor, mainly reflected in the difficulty in clearly identifying the blurred text. In contrast, IFCNN, U2Fusion, and DenseFuse retain the details of the blurred area, and the text is clearly visible. CLF-Net preserves the details of the blurred area to a certain extent and the text is legible.

In Fig. 16, IFCNN, U2Fusion, and IFSepR have the best reservation for the detail features of the fuzzy region, which is closest to the source image. DenseFuse and CLF-Net can also better retain enough texture details, while RFN-Nest, FusionGAN, and STDFusionNet are barely able to discern the blurred details.

2) Quantitative Results: The quantitative comparison results of multifocus image fusion are shown in Table V and Fig. 18. U2Fusion ranks first in AG and SF, which

Fig. 18. Quantitative comparison results of CLF-Net and nine state-of-the-art methods on 36 pairs of multifocus images. Six metrics are used for comparison: EN, MI, VIF, SD, AG, and SF. The compared methods are DenseFuse, RFN-Nest, FusionGAN, GANMcC, IFCNN, PMGI, U2Fusion, STDFusionNet, IFSepR, and CLF-Net.

demonstrates that it retains a wealth of detailed texture information. IFCNN has the best effect in the VIF metric, and the images generated are closer to human vision. Our algorithm ranks second in EN, AG, and SF, and third in VIF, indicating that our method is competitive in multifocus image fusion.

Compared with infrared and visible image fusion, the strategy of multifocus image fusion is a little different. Specifically, for the blurred region in the source image, the multifocus fusion image is more inclined to retain the corresponding unblurred part in the single source image. Therefore, some algorithms with good performance in infrared and visible image fusion have poor performance in multifocus image fusion. However, our proposed algorithm based on contrastive learning is also biased to retain information in local areas reserving the region with the highest similarity in the source image. Therefore, our method still has strong application potential in multifocus image fusion.

Compared with infrared and visible image fusion, the results of our method on medical image and multifocus image fusion are slightly inferior. In the future work, we can carry out in-depth research from two aspects: on the one hand, we can use specific training sets to make the model targeted.;n the other hand, based on the characteristics of medical and multifocus images, some new modules can be introduced to optimize the model.

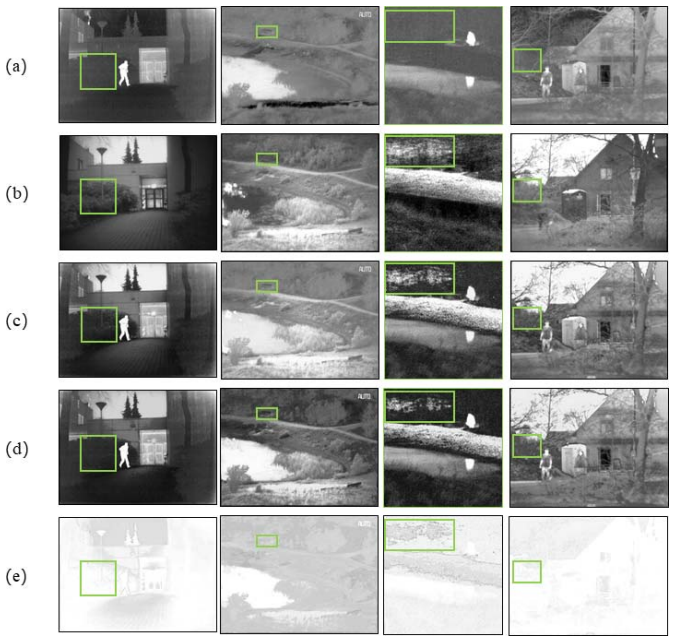


Fig. 19. Visualization of the results of ablation experiment on four typical TNO image pairs. From left to right: Kaptein_1123, lake, bench, and 2_men_in_front_of_house. From top to bottom: (a) infrared images, (b) visible images, (c) fused images of CLF-Net, (d) fused images of CLF-Net without patchNCE loss, and (e) difference between (c) and (d).

TABLE VI

QUANTITATIVE COMPARISON RESULTS OF ABLATION EXPERIMENT ON 20 IMAGES FROM THE TNO DATASETS. SIX METRICS ARE USED FOR COMPARISON: EN, MI, VIF, SD, AG, AND SF. RED REPRESENTS THE BEST RESULT

	EN [41]	MI [42]	VIF [43]	SD [44]	AG [45]	SF [46]
W/o patchNCE	7.1474	3.4332	0.9703	9.4721	4.0682	0.0389
CLF-Net	7.2262	3.6359	1.0484	9.6135	4.9897	0.0495

F. Ablation Experiment

In our model, in addition to introducing the SSIM loss function, we mainly design the loss function based on a comparative learning framework: the patchNCE loss. The two loss functions work together to guide the CNN network to explore the most appropriate parameters. The PatchNCE loss contrasts the source images with the fused image at the deep feature level to guide the network to fully retain the significant target of the infrared image and the detailed texture of the visible image. To verify the effectiveness of the patchNCE loss, we conduct relevant ablation experiments, in which we have removed the patchNCE loss by setting the parameter λ_1 to 0 and only use the SSIM loss to train our network.

The results of ablation experiments are shown in Fig. 19 and Table VI. As shown in Fig. 19(d), the model can only realize the fusion of infrared and visible images with the use of the SSIM loss. However, there are still some flaws in the fusion of the background texture information. Meanwhile, to intuitively compare the difference between the two fusion images with or without patchNCE loss, we directly perform the subtraction operation for the two images and invert the results, which are shown in Fig. 19(e). The darker the image is, the greater the difference between the two fusion images. Fig. 19(e) shows that there are many dark parts in the overall background area and the edge part of the salient target, which indicates that the patchNCE loss has made up for a large amount of background texture information and edge information of the salient target. In addition, the results of the quantitative comparison are demonstrated in Table III. Compared with the fused images of CLF-Net without patchNCE loss, our method has significantly improved in all metrics, especially in the AG and SF metrics, which shows that the fused images generated by our method have richer gradient and texture information, which is consistent with the previous analyses.

TABLE VII

MEAN AND STANDARD DEVIATION OF THE RUNNING TIME OF DIFFERENT METHODS ON THE TNO AND ROADSCENE (UNIT: SECOND). RED REPRESENTS THE BEST RESULT, BLUE REPRESENTS THE SECOND BEST RESULT, AND BOLD REPRESENTS THE THIRD BEST RESULT

Methods	TNO	RoadScene	Parameters(M)
DenseFuse [9]	**0.0912 ± 0.0540**	**0.0484 ± 0.0223**	0.296772
RFN-Nest [17]	0.2692 ± 0.1246	0.1142 ± 0.0357	30.096996
FusionGAN [11]	1.2279 ± 0.5322	0.6627 ± 0.1029	1.323586
GANMcC [13]	2.4147 ± 1.0291	1.3869 ± 0.2020	2.271299
IFCNN [14]	0.0246 ± 0.0273	0.0148 ± 0.0186	0.334348
PMGI [15]	0.2702 ± 1.1260	0.1436 ± 0.0317	0.042017
U2Fusion [16]	1.5748 ± 0.6614	0.8602 ± 0.1272	0.659217
STDFusionNet[19]	0.6971 ± 0.3126	0.3671 ± 0.0762	**0.282513**
IFSepR [47]	2.7371 ± 0.5484	1.4836 ± 0.1634	0.513827
CLF-Net	0.0857 ± 0.0530	0.0459 ± 0.0239	0.190632

G. Efficiency and Complexity Evaluation Experiment

For image fusion tasks, model generation efficiency is also an important indicator to evaluate the model performance, which can be intuitively evaluated by running time. By testing on the TNO and RoadScene datasets, we list the mean and standard deviation of the running time of nine different algorithms, including our algorithm in Table VII. The mean of the running time can reflect the comprehensive running efficiency of the model, and the standard deviation of the running time can reflect the robustness of the model for the fusion of source images with different resolutions. According to the results, our methods are superior to most algorithms except for IFCNN, which indicates that our CLF-Net is competitive in the evaluation of running efficiency.

In addition, we also test the number of parameters of the model. The result shows that the number of parameters of CLF-Net is relatively low compared with other comparison algorithms, only higher than PMGI slightly, which indicates that our network is lightweight.

H. Expending Experiment

In this section, we discuss negative sample selection in the adaptive patchwise contrastive learning framework. There are three possible implementations, as described in the following example. If we select a query from the fused image obtained by the infrared encoder and a positive sample from the infrared image, then we can acquire the negative examples from the following three strategies: 1) the visible image encoded by the visible encoder; 2) the visible image encoded by the infrared encoder; and 3) the rest of the infrared image, which is the

TABLE VIII

QUANTITATIVE COMPARISON RESULTS OF EXPENDING EXPERIMENT ON 20 IMAGES FROM THE TNO DATASETS. SIX METRICS ARE USED FOR COMPARISON: EN, MI, VIF, SD, AG, AND SF. RED REPRESENTS THE BEST RESULT

	strategy (a)	strategy (b)	strategy (c)
EN [41]	6.9340	6.5267	7.2262
MI [42]	3.3141	2.7148	3.6359
VIF [43]	0.8139	0.4436	1.0484
SD [44]	9.3358	9.5201	9.6135
AG [45]	4.2819	4.2593	4.9897
SF [46]	0.0479	0.0465	0.0495
Time	16'52"	17'29"	16'50"

implementation adopted in this study. The comparative test results are shown in Table VIII.

From Table VIII, it can be seen that the best performance in all metrics is achieved using strategy 3), the method proposed in this article. The reasons are analyzed as follows. First, for strategy 1), the features extracted from different encoders are inconsistent, which results in a significant difference between the negative examples obtained from the visible encoder and the query obtained by the infrared encoder. Based on this issue, the contrastive loss is maintained at a low value from the beginning of training. As a result, this strategy cannot be implemented as a good guiding role in the optimization of network parameters. Second, for strategy 2), positive and negative patches are encoded by the same encoder, which can ensure that positive and negative patches are in the same space. However, since the function of the infrared encoder is specific, it has no practical significance in obtaining visible image features from the infrared encoder. Thus, there is still a significant difference between the positive sample and the negative sample, which also exists at the beginning of the training. In addition, strategy 2) has considerable computational redundancy, which can also be seen from the training time per epoch.

V. CONCLUSION

In this article, we propose a novel end-to-end infrared and visible image fusion network, named CLF-Net. A novel unsupervised NCE framework is introduced into the image fusion. Based on the framework, we design an adaptive patchwise contrastive loss, which focuses on the deep representation similarity. The structural similarity loss is also adopted, which focuses on the structural similarity. Both loss functions guide the network in adaptively extracting and fusing features. As a result, not only is the significant thermal target in the infrared image retained, but the rich texture features in the visible image are also retained. Based on extensive qualitative and quantitative experiments, our CLF-Net is superior to most advanced methods in both visual perception and quantitative metrics. In the future, on the one hand, we will optimize the model to further improve the performance, and on the other hand, we will apply infrared and visible image fusion to a wider range of tasks, such as target detection and RGB-thermal (RGBT) tracking.

REFERENCES

[1] 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.

[2] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognit.*, vol. 40, no. 6, pp. 1771–1784, 2007.

[3] R. Lu, X. Yang, W. Li, J. Fan, D. Li, and X. Jing, "Robust infrared small target detection via multidirectional derivative-based weighted contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[4] R. Lu *et al.*, "Infrared small target detection based on local hypergraph dissimilarity measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[5] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 34–41, Sep./Oct. 2001.

[6] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*. Berlin, Germany: Springer, 2006, pp. 528–539.

[7] S. G. Simone, A. Farina, F. C. Morabito, S. B. Serpico, and L. Bruzzone, "Image fusion techniques for remote sensing applications," *Inf. Fusion*, vol. 3, no. 1, pp. 3–15, 2002.

[8] H. Zhang, H. Xu, and X. Tian, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[9] 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.

[10] 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 Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[11] 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.

[12] 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.

[13] 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 Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[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," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[15] 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 *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12797–12804.

[16] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2020.

[17] H. Li, X.-J. Wu, and J. Kittler, "RFN-nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.

[18] R. Hou *et al.*, "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 640–651, 2020.

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

[20] Z. Zhou *et al.*, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 3–11.

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

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[24] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[25] R. Devon Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," 2018, *arXiv:1808.06670*.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[27] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[28] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[29] T. Chen *et al.*, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[30] M. Caron *et al.*, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[31] J. B. Grill *et al.*, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[32] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[33] M. Kang and J. Park, "ContraGAN: Contrastive learning for conditional image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21357–21369.

[34] T. Park *et al.*, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 319–345.

[35] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.

[36] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[37] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 307–361, Feb. 2012.

[38] K. Ma, Z. Duanmu, H. Yeganeh, and Z. Wang, "Multi-exposure image fusion by optimizing a structural similarity index," *IEEE Trans. Comput. Imag.*, vol. 4, no. 1, pp. 60–72, Mar. 2018.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[40] A. Toet. (Apr. 2014). *TNO Image Fusion Dataset*. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[41] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.

[42] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, Mar. 2002.

[43] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, pp. 127–135, Apr. 2013.

[44] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, p. 355, 1997.

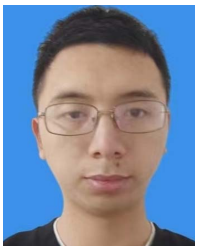
[45] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Opt. Commun.*, vol. 341, pp. 199–209, Apr. 2015.

[46] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[47] X. Luo, Y. Gao, A. Wang, Z. Zhang, and X.-J. Wu, "IFSepR: A general framework for image fusion based on separate representation learning," *IEEE Trans. Multimedia*, early access, Nov. 19, 2021, doi: 10.1109/TMM.2021.3129354.

[48] [Online]. Available: http://www.med.harvard.edu/aanlib/

[49] [Online]. Available: https://github.com/sametaymaz/Multi-focus-Image-Fusion-Dataset

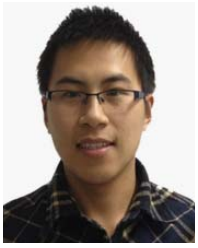


Zhengjie Zhu received the B.E. and master's degrees from the College of Missile Engineering, Rocket Force University of Engineering, Xi'an, China, in 2020, where he is currently pursuing the Ph.D. degree.

His main research interests include computer vision, image processing, and deep learning.

Xiaogang Yang was born in Xi'an, Shaanxi, China, in 1978. He received the Ph.D. degree in control science from the Rocket Force University of Engineering, Xi'an, in 2006.

He is currently a Faculty Member with the Department of Control Engineering, Rocket Force University of Engineering. He is the author of 90 articles and 25 inventions. His research interests include precision guidance and image processing.

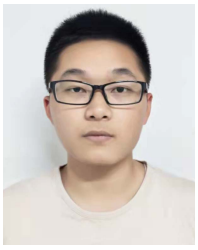


Ruitao Lu received the Ph.D. degree in control science from the National University of Defense Technology, Changsha, China, in 2016.

He is currently a Faculty Member with the Department of Control Engineering, Rocket Force University of Engineering, Xi'an, China. His current research interests include pattern recognition, image processing, and machine learning.

Tong Shen received the M.A.Eng. degree in control science from Chang'an University, Xi'an, China, in 2019.

He is currently a Faculty Member with the Department of Control Engineering, Rocket Force University of Engineering, Xi'an, China. His research interests include pattern recognition, computer vision, and machine learning.



Xueli Xie received the B.E. and master's degrees from the College of Missile Engineering, Rocket Force University of Engineering, Xi'an, China, in 2019, where he is currently pursuing the Ph.D. degree.

His research interests include computer vision, deep learning, and object detection.

Tao Zhang received the B.E. degree from the Wuhan Institute of Technology, Wuhan, China, in 2018, and the master's degree from the College of Missile Engineering, Rocket Force University of Engineering, Xi'an, China, in 2021, where she is currently pursuing the Ph.D. degree.

Her research interests include infrared image processing, visible light remote sensing image processing, object detection, and recognition.