IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 26, 2020, accepted November 6, 2020, date of publication November 16, 2020,
date of current version November 25, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3037770*

# Infrared and Visible Image Fusion Using a Deep Unsupervised Framework With Perceptual Loss

**DONGDONG XU[1], YONGCHENG WANG[1], XIN ZHANG[1,2], NING ZHANG[1,2], AND SIBO YU[1,2]**

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Yongcheng Wang (wangyc@ciomp.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant 11703027.

**ABSTRACT** The fusion of infrared and visible images can utilize the indication characteristics and the textural details of source images to realize the all-weather detection. The deep learning (DL) based fusion solutions can reduce the computational cost and complexity compared with traditional methods since there is no need to design complex feature extraction methods and fusion rules. There are no standard reference images and the publicly available infrared and visible image pairs are scarce. Most supervised DL-based solutions have to take pre-training on other labeled large datasets which may not behave well when testing. The few unsupervised fusion methods can hardly obtain ideal images with good visual impression. In this paper, an infrared and visible image fusion method based on unsupervised convolutional neural network is proposed. When designing the network structure, densely connected convolutional network (DenseNet) is used as the sub-network for feature extraction and reconstruction to ensure that more information of source images can be retained in the fusion images. As to loss function, the perceptual loss is creatively introduced and combined with the structure similarity loss to constrain the updating of weight parameters during the back propagation. The perceptual loss designed helps to improve the visual information fidelity (VIF) of the fusion image effectively. Experimental results show that this method can obtain fusion images with prominent targets and obvious details. Compared with other 7 traditional and deep learning methods, the fusion results of this method are better on objective evaluation and visual observation when taken together.

**INDEX TERMS** Infrared and visible images, deep learning, unsupervised image fusion, densely connected convolutional network, perceptual loss.

## I. INTRODUCTION

Image fusion technology can realize the information synthesis of multi-source images, involving sensor imaging, image preprocessing, image transformation, computer vision, artificial intelligence and other research fields. Image fusion has been widely used in a variety of scenarios. For example, in clinical diagnosis [1], the fusion of important information from CT and MR images can reduce the difficulty of accurate diagnosis. In digital photography, images with different exposure degrees or different focal points can be fused to provide clearer and more realistic digital images. In remote sensing image processing [2]–[4], panchromatic (PAN) images are fused with multi-spectral (MS) images to obtain fusion images with high spatial resolution and rich spectral content.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

In the scenarios of target detection or monitoring [5], infrared and visible images are always fused. The target intensity of infrared image and the details of visible image preserved in the fusion image make it possible to realize the all-weather operation. For a long time, many researchers have devoted to the research of infrared and visible image fusion and made great progress. The fusion of these two kinds of images can realize the integration and balance of image information, and ensure that the target in the fusion image is prominent and the texture is clear. It can enhance the ability to understand different scenes, help to identify the target more accurately and obtain the real-time states [6]. Infrared and visible image fusion has great research value in military field [7], security monitoring, object detection and so on.

After years of development, many traditional infrared and visible image fusion methods have been proposed, including multi-scale transform (MST) method [8]–[10],

sparse representation (SR) method [11], [12], neural network method [13]–[15], hybrid method [16]–[18] and so on. These methods have long been researched and widely used in a variety of scenarios. The steps of these methods usually include image transformation, activity level measurement and fusion rule design. All these steps need to be manually set and operated, and the whole process is relatively complex and has heavy computing burden. In recent years, the application of deep learning in image processing has achieved remarkable results. It has been gradually expanded to target recognition, scene classification, image fusion and other fields. When it is applied to the infrared and visible image fusion [19]–[22], the fusion process is converted to train different deep neural networks. After training, the network can model the complex relationship between the data, and automatically extract the feature information from the source images, and then fuse them. There is no need to design the activity level measurement and fusion rule manually. The fusion process becomes simpler and has strong applicability. The fusion results have obvious targets, rich details and good visual effects.

The publicly available infrared and visible image pairs are insufficient and there are no standard images for reference when training. Some scholars [20] adopt the method of network pre-training to train the encoding/decoding ability of the network on other labeled datasets, and then use the trained network for infrared and visible image fusion. This method increases the amount of calculation. Since the source infrared and visible images are not involved in the training process, it may cause inconformity when testing. These problems limit the application and development of the supervised DL methods for infrared and visible image fusion. If the unsupervised method can be adopted at the moment, the characteristics of infrared and visible images can be directly used. What's more, combined with the corresponding network structure and loss function, there is no need for pre-training and other early network training process, which greatly improves the fusion accuracy and efficiency.

In the scenes of multi-exposure image fusion (MEIF) [23] and multi-focus image fusion (MFIF) [24], the fusion methods of unsupervised deep learning have effectively improved the fusion results. As to unsupervised infrared and visible image fusion, the source image pairs are insufficient and always grayscale images. The designs of the fusion network and the loss function are very important to preserve more useful information and improve the visual effect. In this paper, a kind of infrared and visible image fusion method based on unsupervised convolutional neural network (CNN) with perceptual loss is proposed and the main contributions of our research can be shown as follows:

- An unsupervised CNN model is innovatively proposed and constructed for infrared and visible image fusion. The fusion model does not need any pre-training on other datasets because of the implementation of dataset augmentation. An end-to-end fusion model is realized and there is no need to design complex activity level measurements or fusion strategies.
- For the first time, the perceptual loss is modified and applied for the training of infrared and visible image fusion. By constraining the high-level features of the source images and fusion image extracted by the loss network, the distinguishing features of source images are gradually added to the fusion image. With the join of the perceptual loss, both the performance of the proposed fusion network and the quality of the fusion image get effectively improved.
- In the design of the network architecture, we adopt the multi-layer convolution structure. With a view to the insufficient information of the grayscale source images, a specific form of DenseNet [25] is used as a sub-network for feature extraction and transmission, which can make full use of features of each layer. All these ensure that the decoded fusion image contains multi-scale and multi-level features of the source images.
- The proposed network trained with particular losses can obtain superior fusion images which perform noticeably well on subjective and objective evaluations. Especially, the visual information fidelity gets improved significantly. The correlation between source images and fusion image gets strengthened.

The rest of this paper is organized as follows. In Section 2, some related work about unsupervised learning applied in image fusion, DenseNet, and the perceptual loss are introduced. Section 3 presents the fusion network structure and loss function of our method in detail. Experimental settings, results, fusion evaluations and discussions are provided in Section 4. Finally, Section 5 concludes the paper.

## II. RELATED WORK

### A. IMAGE FUSION BASED ON UNSUPERVISED DL

In the field of image fusion, method based on unsupervised DL was first proposed by Prabhakar [23] *et al.* when realizing the MEIF in 2017. The input multi-exposure images were converted into YCbCr channel data and the CNN was used to fuse the luminance channel to preserve structural details and brightness variations. The no-reference image quality metric structural similarity (SSIM) was designed as the loss function. The network was trained on cropped small image patches without any ground-truth images for supervision. The whole fusion process could be seen as an end-to-end model and there was no need to make pre-training on other datasets. This unsupervised method achieved inspiring fusion results and it could be generalized to MFIF and other fusion tasks. After that, Mustafa [24] *et al.* came up with a deep unsupervised convolutional network for MFIF in which DenseNet was introduced as the feature extraction network. The method could overcome the limitation of unavailability of labeled images for MFIF. In addition, SSIM loss and pixel loss were used as the loss function to train the end-to-end model. Extensive experiments validated the efficiency of the method. Also, the quantitative and qualitative evaluation results were commendable. In [26], another

**TABLE 1. The comparison of different methods.**

| Methods | Traditional | Supervised DL | Unsupervised DL |
|---|---|---|---|
| Realization Form | Multiple (MST/SR…) | Single (Training) | Single (Training) |
| Activity level Measurement | Manual | Auto | Auto |
| Fusion Rules | Manual | Auto | Auto |
| Pre-Training | — | With | Without |
| Efficiency | — | Low | High |
| Loss Function | — | Simple | Complex |
| Network Structure | — | Based on the actual | |

unsupervised deep model for MFIF was proposed. The training dataset and the loss function were similar to [27]. The authors constructed the encoder net through SEDense block which was combined with DenseNet and SeNet [28]. The promising fusion performance demonstrated the advantages of the proposed method. Furthermore, Yan [29] also proposed an unsupervised deep MFIF method trained on cropped source image pairs without any pre-training. This end-to-end model had a stronger ability of extracting features, which was important for image reconstruction. As to infrared and visible image fusion, Ma *et al.* [30] constructed an unsupervised model through GAN which trained on source images and achieved preferable results. Currently, the stumbling blocks in using deep learning for image fusion are lack of sufficient training data and ground-truth images for supervision. Above-mentioned unsupervised image fusion methods can overcome the difficulties to a certain extent by constructing unique DL frameworks and utilizing the no-reference quality metric as loss function. The two issues need to be carefully designed when realizing, especially for infrared and visible image fusion in consideration of the source images in low pixel resolution. Nowadays, the problems still exist and these unsupervised fusion methods have practical significances. The pros and cons of traditional, supervised and unsupervised DL-based infrared and visible image fusion solutions are listed in Table 1.

### B. INTRODUCTION OF DENSENET

In 2017, Huang et al[25] proposed the DenseNet for the first time after the research of the properties of multi-channel features. The design draw lessons from ResNet [31] and Inception [32] network. The DenseNet focuses on feature propagation and feature reuse, and all the middle layer features in the network are applied to reconstruct the output features. They can help to avoid the problem of overfitting caused by the increase of the depth and width of the network. Fig.1 shows the structure of DenseNet.

In the picture above, $x_l$ indicates the feature maps of every layer ($x_0$ is the input), $H_l$ represents a set of continuous

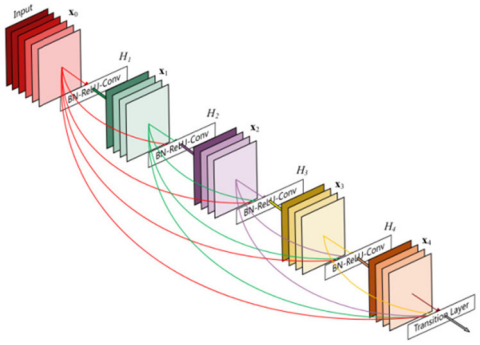


**FIGURE 1. The network structure of DenseNet [25].**

operations that are BN (Batch Normalization), ReLU (Activation Function) and Conv (Convolution). In the $l$th output layer, $x_l$ can be defined as follows:

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \tag{1}$$

$[x_0, x_1, \cdots, x_{l-1}]$ represents the concatenation of feature maps of all former layers in channel dimension. It can be seen that the densely connected network is able to maximize the flow of information. For infrared and visible image fusion, we want to make the best use of the information of the few and low-pixel source images for training. Through DenseNet, the feature information of the front layers is fully utilized and all the information of the middle layers is retained. Moreover, Multi-scale and multi-channel features can be got by DenseNet which are very important for image fusion.

### C. APPLICATION OF PERCEPTUAL LOSS

The perceptual loss was first proposed by Johnson *et al.* [33] for image real-time style transfer and super-resolution. The specific loss is calculated by comparing the high-level feature maps of the ground-truth image extracted through a deep convolutional neural network called loss network with the feature maps of the generated image extracted through the same loss network. The differences of the maps can be converted to loss function to constrain the training process. At last, the high-level feature maps of the ground-truth image and the generated image become similar which means the perceptual features are close to each other. Under normal conditions, if the high-level feature maps of the two images are close, that means the generated image and ground-truth image themselves become similar. The perceptual loss provides new ideas when designing the loss function for image generation or fusion networks. Fig.2 shows the computational process of perceptual loss.

Infrared and visible images respectively stand for different image characteristics. We want to keep all these unique characteristics in the fusion image as much as possible. The perceptual loss can help to guarantee the similarity of high-level features between source images and the fusion image. Combined with other common losses, the performance of the fusion network gets improved obviously.

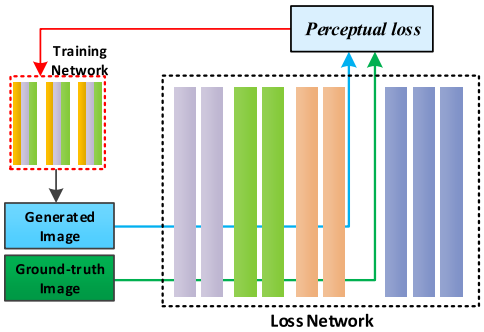


**FIGURE 2.** The computational process of perceptual loss.

### D. COMMON FUSION MODEL

Nowadays, the existing network models for image fusion based on DL mainly consist of six parts, including source inputs, encoder, fusion strategy, decoder, image output and the loss function. Fig.3 shows the process of image fusion. First, the encoder is designed to extract the feature maps of source inputs. Then some fusion strategies are adopted to get the features for reconstruction. Finally, the fusion image is obtained through decoder by feature dimension reduction. The SSIM and the pixel differences between source images and the fused image are always used to optimize the loss.

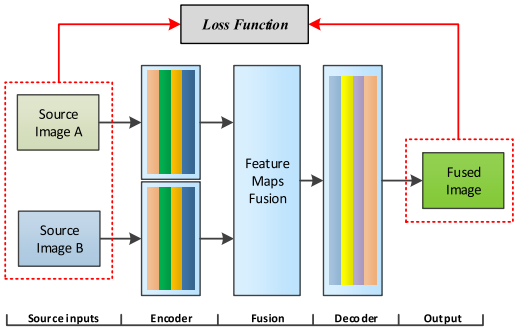


**FIGURE 3.** The network structure of commonly used fusion models.

However, there are mainly two issues that need to be considered when the model is used for infrared and visible image fusion. One is that the structural form of encoders is always the Siamese network [34], which has two branches with the same architectures and weights. It cannot extract and preserve the unique information of the source images obtained by different imaging sensors. Another issue is that pre-training on other large dataset is generally necessary to learn the parameters of encoder and decoder. The features of the multi-modality source images cannot be simulated well.

## III. METHOD

In the following content, the design details of infrared and visible image fusion method based on unsupervised convolutional neural network will be introduced. First, the designed structure of the proposed network is introduced and discussed in detail. Then, we analyze the loss functions which guide the back propagation of the networks. Among the losses, the perceptual loss used for infrared and visible image fusion is originally proposed and calculated. However, as to the SSIM loss, we adopt the commonly used form in the training of image fusion tasks.

### A. PROPOSED FUSION NETWORK

Aiming at the characteristics of infrared and visible images, the fusion network model proposed in this paper has the following improvements compared with the common fusion model.

First, in terms of network input, the source images are no longer separately fed into the siamese-encoder for feature extraction. Instead, the 2-channel network is adopted which has one trunk without branches. By jointly handling the two inputs at the starting point, the 2-channel network owned the highest flexibility [34]. The source images are concatenated as a 2-channel image. The infrared and visible images are in different modalities. The single encoder can synthetically preserve the unique features on the single new input and avoid the problem that the extracted features tend to a certain source image.

Second, the network structure of encoder is promoted from normal convolutional network to DenseNet. This change will help to multi-scale feature transformation and representation, and make it easy for feature transmission and reconstruction. All the features concatenated at the end of the dense block are efficient for reconstructing the fusion image.

Then, there are no specific fusion strategies in the network and the training can be regarded as an end-to-end mapping from input to output. The training and testing process can stay the same without extra design of fusion modules.

Finally, in order to constrain the similarity of high-level features and the low-level features between the source images and the fusion image at the same time, the loss function consisted of SSIM loss and perceptual loss has been designed and optimized to ensure the richness of information preserved in the final fusion image. The SSIM loss can constrain the similarities of the brightness, contrast and structural characteristics. Most basic features can be well preserved in the fusion image. Meanwhile, the perceptual loss can help to guarantee the similarity of high-level features between images. Fig.4 shows the particular network structure.

In the picture above, *Concat* indicates channel concatenation. $k$, $n$, $s$ respectively indicates the kernel size, channel number and the stride. *Conv* and *BN* represent the convolution and batch normalization. ReLu, LReLU and tanh represent three kinds of activation functions. The dense block in the middle consists of five layers with *Conv*, *BN* and *ReLu*. This unsupervised network can output high quality fusion image under the constraint of $L$ (Loss function). The $L$ consists of perceptual loss and SSIM loss. The perceptual loss is calculated through a specific loss network and the SSIM loss is calculated by related algorithms without any network. The $L$ in Fig.4 is simplified to show the whole training process. The concrete contents are introduced in Part C of this Section.

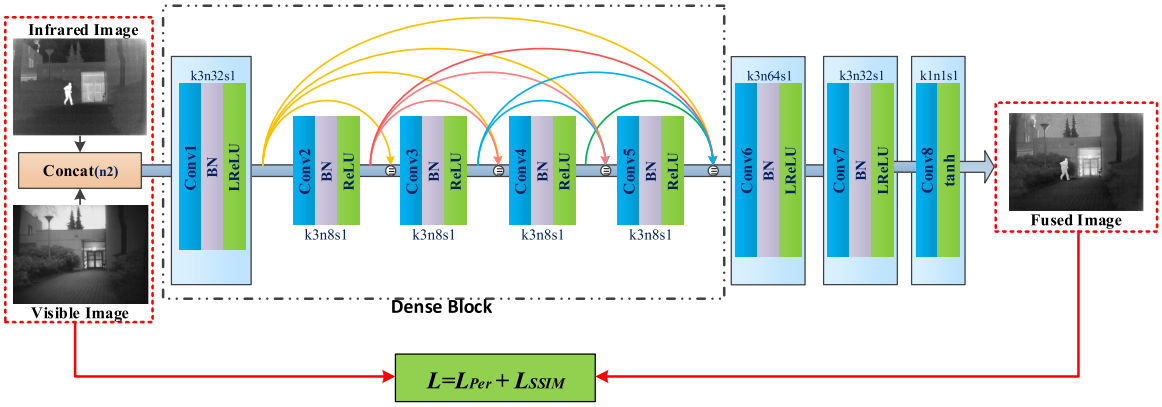


**FIGURE 4.** The network structure of the proposed method.

The settings of the convolution parameters and the functions of DenseNet are explained below.

### 1) THE BASIC SETTINGS

In the proposed network, the kernel sizes are all $3 \times 3$ except the last layer which is $1 \times 1$. In practice, the $3 \times 3$ is the smallest size which can obtain the octagonal information of a pixel. In fact, the number of parameters and computational complexity are greatly reduced by using multiple small convolution kernels. The infrared and visible images are small. The $3 \times 3$ kernels can finely extract and utilize the features. The last convolutional layer adopts single-channel $1 \times 1$ convolution, which can complete the transformation process of linear combination of information between different channels and realize the feature normalization. At the same time, it can complete the dimensionality reduction of multi-channel feature maps, and finally output single-channel fusion image.

The strides are set as 1 for all layers. In convolution operation, the stride of convolution kernel determines the extraction accuracy to some extent. The smaller the stride is, the more comprehensive the extracted features will be and the less the missing information will be. All these are very important for generating the fusion image. If the strides are greater than 1, they may work like pooling operations which are not suitable for image fusion tasks due to the missing of information.

In order to ensure that the size of the fusion image after convolution calculation is consistent with that of the source images, the SAME mode is used for padding in each layer. In addition, since the DenseNet is included in this network, the sizes of the feature maps of each channel are required to remain unchanged to realize the feature superposition at the end of the dense block.

### 2) THE FUNCTIONS OF DENSE BLOCK

The commonly used dense block usually takes the form of "N+N+N+…". Each layer of the block outputs the same number of feature maps and contributes equally to the output of dense block. This may be not applicable to the fusion of infrared and visible images. Because the source images themselves are single-channel grayscale images and the sizes are small. Also, the spatial details are limited and the shallow layers contain more basic information. We want to utilize the information as much as possible through DenseNet. The number of the dense block layers and the output dimension of each layer are adjusted to get desirable experimental results. Finally, the number of dense block layers is adjusted to 5 and performs in the form of "32+8+8+8+8". On the basis of keeping a large number of shallow features, the output of the first layer accounts for half of the channel number of the block. Then the deep features are added. This kind of combination of different channel numbers can effectively improve the quality of the decoded fusion image. The specific functions of dense block for infrared and visible image fusion can be summarized as follows. First, in conventional convolutional neural network, the fusion strategy of feature maps is always the add operation. However, in the densely connected network, feature maps are concatenated through channels, so that features can be used more effectively when integrating the output, and the information preserved becomes richer. Second, through the dense connections, all the features of the middle layers can be retained, the correlation between features is stronger, and the correlation between the fusion image and the source image is guaranteed. Third, this kind of connection and feature transfer in dense block has a certain regularization effect. At the same time, there are fewer channels and fewer parameters in each layer, which can suppress the over-fitting phenomenon to a certain extent. The generalization ability of the network is stronger, which makes the fusion effect better when testing. In addition, the flow of information is increased by multiple skip connections in dense block. Different from the process of forward propagation and back propagation of conventional networks, each layer is directly connected to the input during forward propagation, and losses can be transmitted to the front layer earlier during back propagation. It is conducive to the realization of deep network training. The design of the loss function will be technically introduced in next section.

**TABLE 2.** The structure and parameter information of the image fusion network.

| Name | Layer | Kernel | Stride | Channel (Input) | Features (Input) | Channel (Output) | Features (Output) | Padding | Activation Function |
|---|---|---|---|---|---|---|---|---|---|
| Dense Block | Conv1 | 3*3 | 1 | 2 | Image Pairs | 32 | Net1 | SAME | LReLU |
| | Conv2 | 3*3 | 1 | 32 | Net1 | 8 | Net2 | SAME | ReLU |
| | Concat1 | / | / | 32+8 | Net1/ Net2 | 40 | Concat(Net1,Net2) | / | / |
| | Conv3 | 3*3 | 1 | 40 | Concat(Net1,Net2) | 8 | Net3 | SAME | ReLU |
| | Concat2 | / | / | 40+8 | Concat(Net1,Net2)/ Net3 | 48 | Concat(Net1,Net2, Net3) | / | / |
| | Conv4 | 3*3 | 1 | 48 | Concat(Net1,Net2, Net3) | 8 | Net4 | SAME | ReLU |
| | Concat3 | / | / | 48+8 | Concat(Net1,Net2, Net3)/ Net4 | 56 | Concat(Net1,Net2, Net3, Net4) | / | / |
| | Conv5 | 3*3 | 1 | 56 | Concat(Net1,Net2, Net3, Net4) | 8 | Net5 | SAME | ReLU |
| | Concat4 | / | / | 56+8 | Concat(Net1,Net2, Net3, Net4)/ Net5 | 64 | Concat(Net1,Net2, Net3, Net4,Net5) | / | / |
| Transfer Layer | Conv6 | 3*3 | 1 | 64 | Concat(Net1,Net2, Net3,Net4,Net5) | / | Net6 | / | / |
| Output Layer1 | Conv7 | 3*3 | 1 | 64 | Net6 | 32 | Net7 | SAME | LReLU |
| Output Layer2 | Conv8 | 1*1 | 1 | 32 | Net7 | 1 | Fused Image | SAME | tanh |

#### 3) THE DESCRIPTION OF THE NETWORK STRUCTURE

To clearly describe the structure and parameter information of the image fusion network, the details are listed in Table.2. *Conv* represents the convolution operation, *Net* represents input or output layer of the network, *Concat* represents the channel concatenation.

### B. LOSS FUNCTION

The design of the loss function is of great significance to the performance of the unsupervised convolutional network model. In order to improve the quality of the fusion image objectively and ensure the high similarity between the fusion image and the source images, the perceptual loss is originally introduced to infrared and visible image fusion and combined with frequently-used SSIM loss to achieve optimal constraint between the inputs and the output. The equation (2) gives the definition of the loss function.

$$L = g_1 L_{Per} + g_2 L_{SSIM}, \tag{2}$$

The $L_{Per}$ represents the perceptual loss and the $L_{SSIM}$ represents the SSIM loss. $g_1$ and $g_2$ are ratios which are revised step by step during training.

#### 1) THE PERCEPTUAL LOSS

As described in the related work, perceptual loss performs well in image real-time style transfer and super-resolution. However, it has not been used in infrared and visible image fusion. In order to achieve the training with perceptual loss to improve the effects of the fusion results, the following three main problems need to be solved.

The first problem is the choice of the loss network. Loss network usually refers to a mature and deep neural network trained and verified on a large dataset. They need to have strong capability to extract multi-scale features [35] and adapt to the size of the dataset. Second, there are no standard fusion images for reference. The application of perceptual loss needs to compare the high-level features of standard reference images with the same level features of the fused images generated by the training network. But, in the fusion of infrared and visible images, we cannot obtain the standard images for comparison. Third, the problem of channel correspondence is also need to be considered. In image style transfer or super-resolution reconstruction, the reference image and the output of the network are all single modality images, and it is easy to compare the extracted features. However, in this fusion task, the infrared images and the visible images are obtained by two different sensors, and the output is a single type of fusion image. So it is impossible to extract features directly through the loss network for comparison. In addition, existing loss networks are usually trained based on colorful three-channel (R, G, B) images. Infrared and visible source images and fusion images are all single-channel grayscale images. How to adapt the inputs to loss network also needs to be considered.

In Fig.5, the specific inputs and the computational process of the perceptual loss designed are showed clearly.

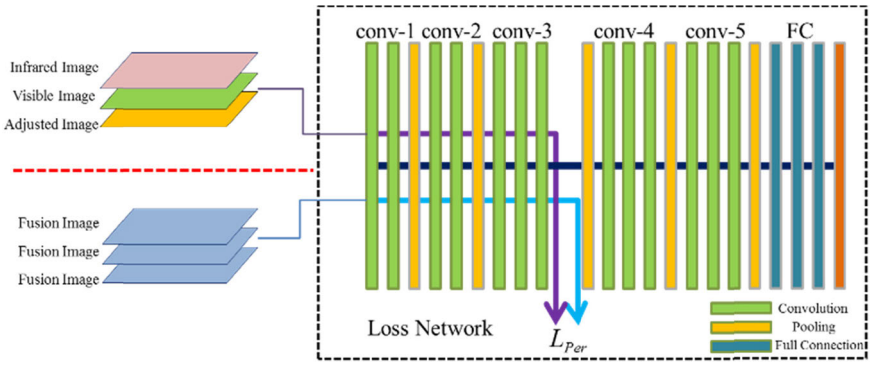


**FIGURE 5.** The perceptual loss used for infrared and visible image fusion.

For the three problems mentioned above, the following solutions are given one by one.

At present, the ready-made models that are saved as parameters and used as feature extraction sub-networks are mainly VGG network [36] (VGG-16, VGG-19) models and residual network (ResNet50, ResNet101, ResNet152) models. They are often used in image classification, target detection, image fusion and other tasks. Considering that the dataset of the source images in this project is small, and the information of infrared and visible gray images is insufficient, the VGG-16 network with fewer layers is more suitable. It can avoid the distortion of high-level features caused by depth extraction, and can improve the training efficiency of the network with fewer parameters.

Although the standard fusion image for reference does not exist, the information of the fusion image must have strong correlations with the input source images. Therefore, in this method, we try to find a way to simulate the standard fusion image. With a view to matching the channels and preserving the information, the idea of the adjusted image is proposed. The two source images and the adjusted image are concatenated to simulate the input. The adjusted image is essentially the weighted average of two source images. The calculation equation is illustrated in (3):

$$F(x, y) = \alpha \cdot I(x, y) + (1 - \alpha) \cdot V(x, y), \tag{3}$$

where $F$ represents the fusion image, $I$ and $V$ represent the infrared image and visible image respectively, $\alpha$ represents the weight, and $(x,y)$ represents the corresponding pixel point.

Finally, we need to solve the problem of channel correspondence. Commonly used loss networks are trained on the datasets of colored visible images. The input of the network is in the form of three channels, namely R, G and B channels of the color image. The first convolutional layer in the network contains 64 convolution kernels of three-channels, and 64 feature maps are calculated as the output. In this method, the specific implementation is showed as follows. Since the infrared, visible and fusion image are all single-channel images, channel concatenation is also adopted when calculating the perceptual loss. When the loss network is used to extract the deep features of the input source images, the single channel infrared image, visible image and adjusted image are cascaded into three channels in the channel dimension to simulate the standard reference image. When the loss network is used to extract the deep features of the fusion image, the same three single channel fusion image obtained during training are cascaded directly into three channels as the input of the loss network. Fig.5 gives the details. The designed calculation method is illustrated in (4):

$$L_{Per}(Y_I, Y_F) = \frac{1}{C_j H_j W_j} \left\| \varphi_j(Y_I) - \varphi_j(Y_F) \right\|_2^2, \tag{4}$$

Among the equation, $\varphi$ indicates the loss network of VGG-16, $Y_I$ represents the combination of source images and adjusted image, $Y_F$ represents the combination of three same fusion images, $j$ represents the $j$th layer of the loss network. $C_j H_j W_j$ represents the $C_j$ feature maps of the $j$th layer with the size of $H_j W_j$ and the total size of the parameters in the layer. $\varphi_j(Y_I)$ and $\varphi_j(Y_F)$ separately represent the output feature maps obtained by the $j$th layer of the loss network. The final loss is calculated with the L2 norm. As we can see from the Fig.5, in the proposed method, the outputs of the third convolutional layer of the third convolutional group called conv-3 are used as the perceptual loss terms for comparison. The number of feature maps outputted by convolutional calculation of this layer is 256, and the size of feature map is determined by the size of the actual input source images. With the join of $L_{Per}$, visually pleasing fusion images are always obtained.

### 2) THE SSIM LOSS

The SSIM loss is the most commonly used loss in unsupervised image fusion tasks as the brightness, contrast and structural characteristics of the images can be comprehensively considered and the spatial structure correlation between images is also considered in the calculation. All these are consistent with the way that the human vision system obtains the structure information in the visual area, and the distortion state of images can be perceived.

In the $L_{SSIM}$, the SSIM between each source image and fusion image is calculated respectively and the weight can be adjusted to adapt different conditions. The $L_{SSIM}$ is calculated

as (5):

$$L_{SSIM} = 1 - (w \cdot SSIM(I, F) + (1 - w) \cdot SSIM(V, F)), \quad (5)$$

where $\omega$ represents the weight and $SSIM$ ($\sim$) just implement the structural similarity operation in [37]. $I$ and $V$ are infrared and visible images, and $F$ is the fusion image.

## IV. EXPERIMENTS

### A. TRAINING AND TESTING OF THE NETWORK

#### 1) DATASET AUGMENTATION

In the field of image fusion, the source images after registration are used as the input and then the final image fusion is realized through feature extraction and transformation. However, limited to imaging equipment and the military use of such images, publicly available infrared and visible image pairs are very scarce. In this experiment, 41 pairs of infrared and visible source images after registration were collected from the TNO [38] dataset. In these images, infrared images generally had obvious intensity contrast, and at the same time, visible images were rich in detail information. The pairs were divided into two parts: 31 for training and 10 for testing. The test images did not participate in the training process. However, if only a few dozen pairs of source images were used for training, it was difficult to get a stable and robust model. Therefore, we decided to amplify the dataset and use the amplified dataset as the basis for network training. Because the original dataset was too small, rotation, stretching and other methods contribute little to the amount of data. The most direct method was to cut the source image into small images of a certain size to expand the number of datasets. Considering the actual fusion effect and the computing capability of GPU, the cropping size was set as $128 \times 128$, and the stride was set as 11. The cropping was done one by one from left to right and from top to bottom on 31 pairs of source images respectively. Finally, 46,209 pairs of small source images were obtained to form a training set. Another 10 pairs of source images were directly used as network input when testing, and 10 fused images were obtained for the comprehensive evaluation of the results. Fig.6 shows the process of image cropping.

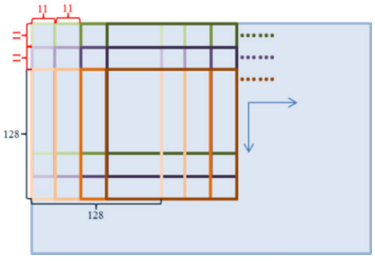


**FIGURE 6.** The process of image cropping.

#### 2) DETAILED SETTINGS OF TRAINING

In addition to the above image cropping settings, the other parameters were set as follows. For training, cropped image pairs were put into the network whose batch size was set as 32 and the learning rate was initialized as $10^{-5}$. The ADAM optimizer was chosen to train the network whose default parameters were set as follows: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. The GPU platform was Intel E5-2680 V3 processor with TITAN V GPU and 64G memory. Except these fixed settings of training, there were some variable parameters in Eq.1, Eq.2 and Eq.3 when calculating the loss functions. $g_1$ and $g_2$ of the total loss $L$ in Eq.1 were both set as 1 at first. With the join of perceptual loss, the visual effects of the fusion images were improved obviously. But when we increased the ratio of $g_1$, image distortion appeared and less texture details were reserved. Then we began to increase the ratio of $g_2$ to constrain the basic structural similarity between the source images and the fusion image. By this time, the quality of the fusion image got better results on subjective and objective evaluation. The final $g_1$ and $g_2$ were set as 1 and 160, and it has been verified by many experiments. The adjusted image was designed to balance the input of the source images and the initial value of $\alpha$ in Eq.2 was set as 0.5. At this time, the hidden objects of infrared images were less prominent in the fusion image. So we began to adjust $\alpha$ to reserve more high-level features of infrared image. Finally, the $\alpha$ was adjusted to 0.6 when calculating the perceptual loss. In order to keep the structural characteristics and the spatial structure correlations between each source image and the fusion image equably, we just followed the original design of the SSIM loss and the $\omega$ in Eq.3 was just set as 0.5. Of course, it worked well.

#### 3) THE CROPPING OF FUSION IMAGE

In order to ensure better stability and robustness of the network, the amplified dataset was used during training, and 10 pairs of source image pairs were directly put into the trained network when testing. Since dense block was introduced in this method, in order to realize the feature concatenation, it was required that the feature sizes of each layer should be consistent. Therefore, padding was introduced in the convolution calculation. However, the filling operation would result in gray pixel blocks at the edges of the fused image, which affected the image quality. In order to eliminate the effect of the gray block, we pre-filled the source image pairs before testing (adding 0 to the periphery). As the size of the input images increased, the size of output fusion image also increased. The gray block only affected the pixels around the image. Therefore, the obtained large fusion image could be cropped to a normal output size effectively by removing the gray block, and the size of the final output image remained the same as the original input. The schematic diagram of testing process is showed as Fig. 7.

### B. COMPREHENSIVE EVALUATION OF FUSION IMAGE

In actual application scenarios, due to different requirements or application purposes, the emphases on features of the fusion image are different. In most cases, there are no standard fusion images for reference and no standard evaluation metrics to measure the absolute quality of fusion results.

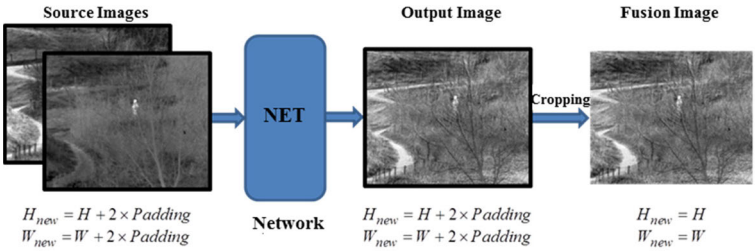


**FIGURE 7.** The process of image fusion and cropping.

At present, researchers generally believe that it is a reasonable way to evaluate the results by using a variety of evaluation metrics [39]. Liu et al. [40] established an evaluation system based on information theory-based metrics, image feature-based metrics, image structural similarity-based metrics and human perception inspired fusion metrics through investigation, which is often used by researchers. In recent years, many researchers have studied how to quantify the ''good'' or ''bad'' of each evaluation metric through the score value or accuracy, and analyzed which metrics are more reliable in evaluation. Petrovic [41] built a dataset specifically designed to evaluate the quality of the fused images. By matching the evaluation results of each metric to the known subjective results of the two fused images in each group, the evaluation accuracy of each metric to the images in the dataset can be calculated, which is called correct ranking (CR). Based on the objective evaluation system proposed by Liu [40], the CR values of the verified metrics and the improved algorithms of the related metrics, this paper starts from the following two aspects to establish an system for comprehensive evaluation to evaluate the results.

The quality evaluation metrics measuring the fusion image itself are introduced as follow. The spatial frequency (SF) [42] indicates the frequency information of the image. The bigger the SF is, the clearer the image is. The standard deviation (SD) [43] evaluates the contrast of the image. The bigger the SD is, the more attractive the image is.

Evaluation metrics measuring the correlation between fusion image and source images are showed below. In the correlation metrics, the correlation coefficient (CC) [44] is selected to evaluate the degree of relevance between images. Structural similarity (MSSIM) [37] is selected to calculate the structural similarity between the source images and the fusion image. The bigger the CC and MSSIM are, the stronger correlations are built between images. Feature mutual information (FMI) [45] is added to evaluate the degree of feature correlation between fusion image and source images from the perspective of information theory. The bigger the FMI is, the more basic information is preserved in the fusion image. The visual information fidelity for fusion (VIFF) [46] is selected to evaluate the fidelity of the fusion image to the information of source images from the perspective of human visual system. The bigger the VIFF is, the better visual effects can be achieved.

Furthermore, subjective evaluation is also used as an auxiliary mean to supplement the objective evaluation result to ensure the comprehensiveness and accuracy of the evaluation.

### C. FUSION RESULTS

In this section, the proposed unsupervised fusion method is compared with 4 traditional fusion methods (DTCWT [47], LP [48], NSCT [49], GFF [50]) and 3 fusion methods based on deep learning (CNN [19], Dense-add/L1 [20], add and L1 represent different fusion strategies, Deepfuse [23]). We used the codes provided by the authors or well-known toolbox to generate the fused images from source image pairs. Notably, as to the Deepfuse, we took the codes provided by [20]. The effectiveness and significance of this method are analyzed by combining subjective evaluation with objective evaluation.

#### 1) SUBJECTIVE EVALUATION

Fig.8, Fig.9 and Fig.10 show the fusion results of each method on the image pairs of ''Bunker'', ''Nato_camp'' and ''Marne_04'' in the test set. (a) and (b) on the left are the source images. (c) to (j) are the fusion results obtained by the comparison methods and the proposed method. As for Dense method, we just give out the better results from the different fusion strategies.

It can be seen from the comparison of the three fusion results above that although the fusion images obtained by the traditional methods of DTCWT, LP and NSCT can reflect the basic information of the two source images, the fusion images are not clear enough and the images always have low contrast, which are not conducive to visual observation. The fusion results obtained by GFF and Dense-L1 methods are convenient for visual observation, but the fusion images tend to retain the information of a certain source image, while the information of another source image is less preserved. The fusion images obtained by CNN method are generally high in contrast and easy to attract people's attention. However, there are always some unsatisfactory areas, such as the edge of eaves and the front window of the car in Fig.10. Dense-add and Deepfuse methods retain more details of the source image, and the fusion image is better, but the contrast of the fusion image is slightly lower.

As for the fusion method proposed in this paper, the fusion image can retain the important features and details of the

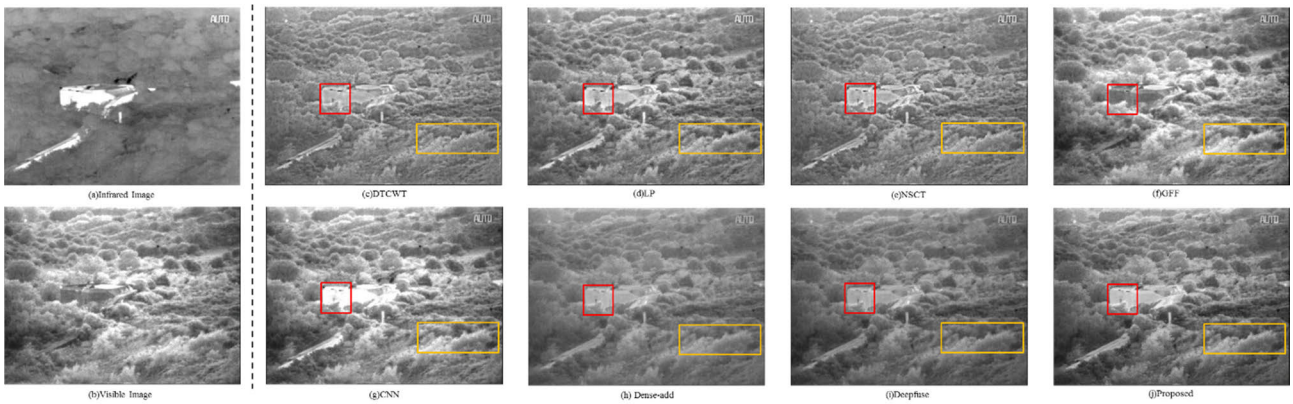


**FIGURE 8.** The comparison of image fusion results of Bunker.

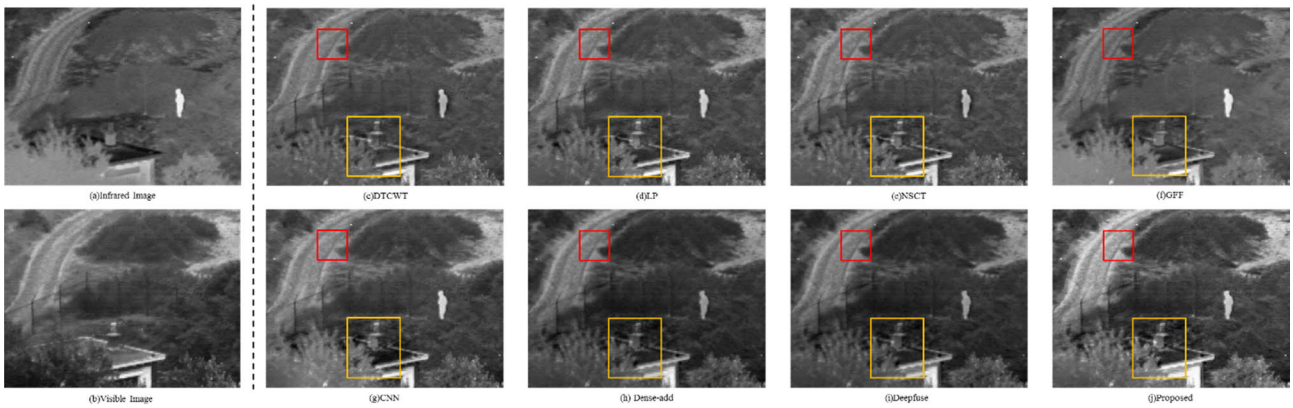


**FIGURE 9.** The comparison of image fusion results of Nato_camp.


**FIGURE 10.** The comparison of image fusion results of Marne_04.

two source images at the same time, and less noise is mixed into the image, which also ensures the correlation between the fusion image and the source images. More importantly, the fusion images have higher visual fidelity. The images look natural and comfortable, and are easier to observe by the human visual system. By zooming in on the three pairs of images above, the advantages of the images fused by the proposed method are prominent. It can be seen that in Fig.8, the outer walls of the bunker are clear and the forest looks exuberant. In Fig.9, the traces of the road are legible and the objects around the chimney are trenchant. In Fig.10, the gap between branches on the left can be seen, and the camouflage on the car can also be distinguished.

**TABLE 3.** The average values of the SIX metrics for 10 fusion images.

| Methods | Traditional Method | | | | DL-Based Method | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | DTCWT | LP | NSCT | GFF | CNN | Dense-add | Dense-L1 | Deepfuse | **Proposed** |
| *SF* | 11.5691 | 11.8508 | 11.6472 | 11.1043 | 11.8958 | 9.2323 | 9.3108 | 9.2475 | **12.2235** |
| *SD* | 29.2046 | 32.7525 | 29.6933 | 40.147 | **48.1136** | 35.6666 | 38.9017 | 36.2734 | 43.2497 |
| *MSSIM* | 0.5579 | 0.5665 | 0.5761 | 0.5653 | 0.5673 | **0.59713** | 0.5798 | 0.5964 | 0.5773 |
| *CC* | 0.5265 | 0.5214 | 0.5312 | 0.4200 | 0.4956 | 0.5503 | 0.5126 | 0.5501 | **0.5635** |
| *FMI* | 0.3876 | 0.3345 | 0.3824 | 0.3896 | 0.3221 | **0.4057** | 0.3623 | 0.4041 | 0.3635 |
| *VIFF* | 0.3589 | 0.4518 | 0.4084 | 0.2913 | 0.5067 | 0.4929 | 0.3619 | 0.5063 | **0.6561** |

#### 2) OBJECTIVE EVALUATION

In this section, the objective and quantitative evaluation of the fusion images are carried out through the comprehensive evaluation system. There are a total of 10 groups of images in the test set, and 6 metrics are used to evaluate the fusion results of each method. The average values of each metric on the fusion images obtained by each method are shown in Table.3 and the larger the values of all metrics, the better.

First of all, a general comparison is made between traditional methods and DL-based methods. It can be seen that among the six evaluation metrics, the values obtained by the DL-based methods are more outstanding, which prove the effectiveness of the DL-based methods in image fusion. There are no complex transformations in DL-based methods and we can improve the fusion results by adjusting the network structure and loss function. They are flexible, easy to implement, and the fusion images are of high quality.

Then, the evaluation results of each method are analyzed. Compared with other methods, the proposed fusion method achieves the best values in three metrics of SF, CC and VIFF, which indicates that the fusion images obtained are rich in spatial details and highly correlated with the source images. The result about VIFF is more outstanding, which indicates that the fusion images have high fidelity of visual information and good visual effect. On the SD metric, the proposed method is only smaller than the CNN method. The reason is that the CNN method not only trains the siamese network to extract features, but also introduces the multi-scale transformation in the traditional method to assist the image decomposition and reconstruction, which greatly improves the image characteristics. In terms of MSSIM metric, the proposed method is smaller than the Dense and Deepfuse method. In the Dense method, a dataset composed of a large number of visible images is used for pre-training. At the same time, densely connected structure is used internally, so the features of structure are well preserved. Also, that method once again achieved the maximum value in FMI metric. For the Deepfuse method, the better performance of MEF SSIM is attributed to its objective function. The proposed method is slightly deficient in FMI metric, partly because it restricts high-level features through perceptual loss, which leads to the neglect of some features of source images and the reduction of mutual information. We can see that the objective evaluation and subjective evaluation are consistent.

### D. MODEL ANALYSES

The fusion method proposed in this paper mainly contains two outstanding points. When designing the network structure, dense block is added to extract and utilize multi-dimensional features. As to the loss function, the perceptual loss is applied to the fusion of infrared and visible images. The following experiments are carried out to analyze the functions of the two designs based on the actual fusion images.

#### 1) THE FUNCTION OF DENSE BLOCK

In the fusion model of this experiment, dense block is replaced by conventional convolutional layers for training. Fig.11 shows the comparison before and after partial changes of the dense block. For the convenience of comparison, although the form of this molecular network changes, the number of input channels (32) and output channels (64) remains unchanged.

After the dense block is changed into the conventional convolution form, although the sizes of the feature maps between layers are unchanged, a lot of parameters are substantially increased in the axial direction. In addition to the addition operation of skip connections, when using dense block network, the calculation of this molecular network involves 12,672 parameters ((32+40+48+56)∗(3∗3)∗8). After modification into conventional convolution, a total of 46,080 parameters (32∗(3∗3)∗(32∗3+64) are involved in sub-network calculation. As can be seen from the comparison between (c) and (d) in Fig.12, the densely connected network can obtain and combine deep image features with fewer parameters through feature propagation and feature reuse, which is conducive to the output of fusion images. Generally, images with rich details can be achieved.

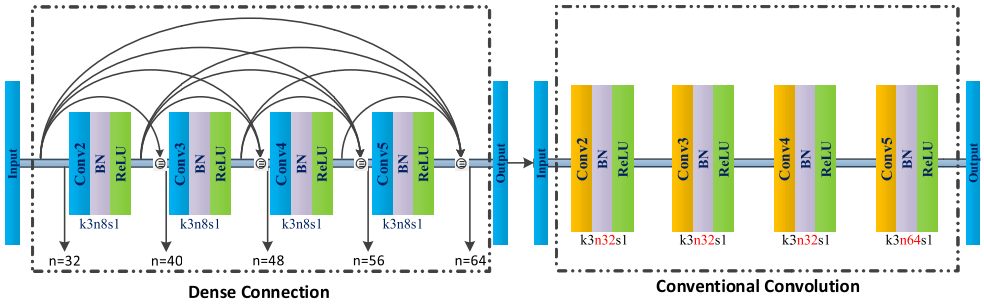


**FIGURE 11.** The change diagram of dense block.


**FIGURE 12.** The comparison of fused images.

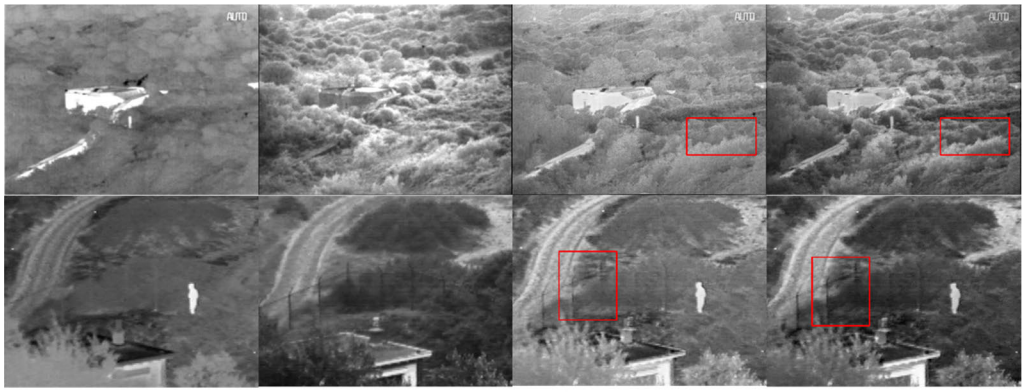


**FIGURE 13.** The comparison of fused images.

We also compare the results obtained by different number of dense block layers. Since the images are similar, we list the values calculated by the objective metrics in Table 4. We can see that the proposed method with 5 layers dense block can get correspondingly better results.

### 2) THE FUNCTION OF PERCEPTUAL LOSS

In this experiment, the structure of the proposed network model remains unchanged during network training, while the perceptual loss is removed, and only the SSIM loss is retained to constrain the image fusion process. The comparison of fusion results with and without perceptual loss is shown in Fig.13.

The perceptual loss mainly constrains the high-level feature similarity between the source images and the fusion image, and then realizes the fusion process. It can be seen from the results above that the pictures acquired by the method in this paper are vivid and have a good visual effect. With the introduction of perceptual loss, the contrast of the fusion image is increased. In the fused images, the visible textures and the infrared targets are comparatively clear. The information preserved in the fusion images is helpful.

**TABLE 4.** The objective evaluation values of different layers.

| | 4 layers | 6 layers | 7 layers | Proposed (5 layers) |
|---|---|---|---|---|
| *SF* | **12.6591** | 12.5341 | 12.2194 | 12.2235 |
| *SD* | 40.5298 | 41.6594 | 39.6599 | **43.2497** |
| *MSSIM* | 0.5724 | 0.5677 | 0.5753 | **0.5773** |
| *CC* | 0.5659 | 0.5658 | **0.5667** | 0.5635 |
| *FMI* | 0.3490 | 0.3477 | 0.3309 | **0.3635** |
| *VIFF* | 0.5984 | 0.6154 | 0.5940 | **0.6561** |

**TABLE 5.** The objective evaluation values of different perceptual loss.

| | VGG-19 | ResNet50 | Proposed (VGG-16) |
|---|---|---|---|
| *SF* | 12.1528 | **12.4327** | 12.2235 |
| *SD* | 42.3984 | 40.9502 | **43.2497** |
| *MSSIM* | 0.5691 | 0.5506 | **0.5773** |
| *CC* | **0.5688** | 0.5604 | 0.5635 |
| *FMI* | **0.3682** | 0.3486 | 0.3635 |
| *VIFF* | 0.6029 | 0.5648 | **0.6561** |

Moreover, other deeper networks for perceptual loss like VGG-19 and ResNet50 have been realized for comparison. We list the values calculated by the objective metrics in Table 5. We can see that the VGG-16 loss is more suitable for the proposed infrared and visible image fusion method.

## V. CONCLUSION

In this paper, an infrared and visible image fusion model based on unsupervised convolutional neural network is proposed and implemented. When designing the network structure, DenseNet is introduced as the feature extraction and transmission sub-network to make full use of the features of each layer. The feature maps concatenated by the block are advantageous for feature reconstruction to obtain the fusion image. The decoded output image contains multi-scale and multi-level features of the source images. The distinct target and detail information in the source images are effectively preserved. The image quality gets evidently improved. For the first time, the training process of infrared and visible image fusion is constrained by the use of perceptual loss. By reducing the difference of high-level features extracted by the loss network between the simulative standard image and the generated image, the perceptual loss can make the fusion image continuously optimized and close to the target image. Applying the perceptual loss to constrain the difference of perceptual features between the source images and the fusion image can guarantee the fusion image contain more information of the source images. The visual effect of the fusion image gets better and it is easier for human visual system to observe. The objective evaluation and subjective evaluation both demonstrate the effectiveness of the proposed method.

We believe that the basic framework of our fusion method can be applied to other image fusion tasks, such as medical image fusion, multi-exposure image fusion and multi-focus image fusion with some changes in the input and loss function.

## REFERENCES


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

[2] Z. Wang, D. Ziou, C. Armenakis, D. Li, and Q. Li, "A comparative analysis of image fusion methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1391–1402, Jun. 2005.

[3] C. Pohl and J. L. Van Genderen, "Review article multisensor image fusion in remote sensing: Concepts, methods and applications," *Int. J. Remote Sens.*, vol. 19, no. 5, pp. 823–854, Jan. 1998.

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

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

[6] S. P. Constantinos, M. S. Pattichis, and E. Micheli-Tzanakou, "Medical imaging fusion applications: An overview," in *Proc. Conf. Rec. 35th Asilomar Conf. Signals, Syst. Comput.*, 2001, pp. 1263–1267.

[7] M. I. Smith, A. N. Ball, and D. Hooper, "Real-time image fusion: A vision aid for helicopter pilotage," *Proc. SPIE*, vol. 4713, pp. 30–41, Feb. 2002.

[8] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Inf. Fusion*, vol. 12, no. 2, pp. 74–84, Apr. 2011.

[9] G. Pajares and J. Manuel de la Cruz, "A wavelet-based image fusion tutorial," *Pattern Recognit.*, vol. 37, no. 9, pp. 1855–1872, Sep. 2004.

[10] Z. Zhang and R. S. Blum, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," *Proc. IEEE*, vol. 87, no. 8, pp. 1315–1326, Aug. 1999.

[11] J. Wang, J. Peng, X. Feng, G. He, and J. Fan, "Fusion method for infrared and visible images by using non-negative sparse representation," *Infr. Phys. Technol.*, vol. 67, pp. 477–489, Nov. 2014.

[12] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3450–3459, Dec. 2012.

[13] R. Eckhorn, H. J. Reitbock, M. Arndt, and P. Dicke, "A neural network for feature linking via synchronous activity: Results from cat visual cortex and from simulations," *Can. J. Microbiol.*, vol. 46, no. 8, pp. 759–763, 1989.

[14] Z. Wang and C. Gong, "A multi-faceted adaptive image fusion algorithm using a multi-wavelet-based matching measure in the PCNN domain," *Appl. Soft Comput.*, vol. 61, pp. 1113–1124, Dec. 2017.

[15] Y. Lin, S. Le, Z. Xin, and Y. Huang, "Infrared and visible image fusion algorithm based on contourlet transform and PCNN," in *Proc. SPIE, Infr. Mater., Devices, Appl.*, vol. 6835, 2008.

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

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

[18] X. Huang, G. Qi, H. Wei, Y. Chai, and J. Sim, "A novel infrared and visible image information fusion method based on phase congruency and image entropy," *Entropy*, vol. 21, no. 12, p. 1135, Nov. 2019.

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

[20] 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.

[21] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039.

[22] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710.
[23] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, p. 3.
[24] H. T. Mustafa, F. Liu, J. Yang, Z. Khan, and Q. Huang, "Dense multi-focus fusion net: A deep unsupervised convolutional network for multi-focus image fusion," in *Proc. 19th. Int. Conf. Artif. Intell. Soft Comput. (ICAISC)*, Zakopane, Poland, Jun. 2019, pp. 153–163.
[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[26] B. Ma, X. Ban, H. Huang, and Y. Zhu, "SESF-fuse: An unsupervised deep model for multi-focus image fusion," 2019, *arXiv:1908.01703*. [Online]. Available: http://arxiv.org/abs/1908.01703
[27] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, vol. 8693, Sep. 2014, pp. 740–755.
[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2017.
[29] X. Yan, S. Z. Gilani, H. Qin, and A. Mian, "Unsupervised deep multifocus image fusion," 2018, *arXiv:1806.07272*. [Online]. Available: http://arxiv.org/abs/1806.07272
[30] 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.
[31] 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.
[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 1–9.
[33] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016.
[34] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.
[35] A. Lucas, S. Lopez-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3312–3327, Jul. 2019.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[37] 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.
[38] (2014). *TNO Image Fusion Dataset*. [Online]. Available: https://figshare.com/articles/TNImageFusionDataset/1008029
[39] G. Schwan and N. Scherer-Negenborn, "An approach to select the appropriate image fusion algorithm for night vision systems," *Proc. SPIE*, vol. 9649, Oct. 2015, Art. no. 964908.
[40] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.
[41] V. Petrović, "Subjective tests for image fusion evaluation and objective metric validation," *Inf. Fusion*, vol. 8, no. 2, pp. 208–216, Apr. 2007.
[42] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, 1995.
[43] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, pp. 355–375, 1997.
[44] M. Deshmukh and U. Bhosale, "Image fusion and image quality assessment of fused images," *Int. J. Image Process.*, vol. 4, no. 5, pp. 484–508, 2010.
[45] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE 8th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2014, pp. 1–3.
[46] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013.
[47] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.
[48] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *Readings Comput. Vis.*, vol. 31, no. 4, pp. 671–679, 1987.
[49] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, Jul. 2009.
[50] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

**DONGDONG XU** received the bachelor's degree from Shandong University, in 2013, the master's degree from the Harbin Institute of Technology, in 2015, and the Ph.D. degree from the Chinese Academy of Sciences, in 2020. He is currently a Research Assistant with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include deep learning, image fusion, and embedded systems.

**YONGCHENG WANG** received the bachelor's degree from Jilin University, in 2003, and the Ph.D. degree from the Chinese Academy of Sciences, in 2010. He is currently a Researcher with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include artificial intelligence, image engineering, and embedded system of space payload.

**XIN ZHANG** received the bachelor's degree from Northeastern University, Qinhuangdao, China, in 2016. She is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. Her research interests include deep learning and hyperspectral image classification.

**NING ZHANG** received the bachelor's degree from Northeastern University, Qinhuangdao, China, in 2017. She is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. Her research interests include image processing, deep learning, and remote sensing image super-resolution.

**SIBO YU** received the master's degree from Harbin Engineering University, Harbin, China, in 2018. He is currently an Algorithm Designer with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests inlcude deep learning and image super resolution reconstruction.