IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2023, accepted 22 July 2023, date of publication 7 August 2023, date of current version 17 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3302702

RESEARCH ARTICLE

An Efficient Network Model for Visible and Infrared Image Fusion

ZHU PAN AND WANQI OUYANG
School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China
Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China
Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China

Corresponding author: Zhu Pan (pandey@wust.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61901310.

ABSTRACT Visible and infrared image fusion (VIF) aims at remodeling an informative and panoramic image for subsequent image processing or human vision. Due to the widespread application in military and civil fields, the VIF technology has achieved considerable development in recent decades. However, the assignment of weights and the selection of fusion rules seriously restrict the performance improvement of most existing fusion algorithms. In response to this issue, an innovative and efficient VIF model based on convolutional neural network (CNN) is proposed in this paper. Firstly, multi-layer convolution kernel is performed on two source images with a multi-scale manner for extracting the salient image features. Secondly, the extracted feature maps are concatenated along the number of channels. Finally, the fusion feature maps are reconstructed to achieve the fusion images. The main innovation of this paper is to adequately preserve meaningful details and adaptively integrate features information driven by source image information in CNN learning model. In addition, in order to adequately train the network model, we generate a large-scale and high-resolution image training dataset based on COCO dataset. Compared with the existing fusion methods, experiment results indicate that the proposed method not only achieves universally outstanding visual quality and objective metrics but also has some advantages in terms of runtime efficiency compared to other neural network algorithms.

INDEX TERMS Convolutional neural network, multi-feature extraction, optimized network, visible and infrared image fusion.

I. INTRODUCTION

For the past few decades, image fusion technology for medical image, multi-focus image, remote sensing image, infrared and visible image [1], [2], [3], [4], etc. has received great attention, because a compositive image can provide more abundant scene information, which is very propitious to special comprehensive analysis and application [5], [6], [7]. Especially, attributing to the wide and important application of infrared systems in military or civil surveillance, the fusion methods for visible and infrared image spring up like bamboo shoots after rain, and evidently improved people's ability to observe or monitor special scenarios [8], [9], [10], [11], [12].

Taking the action domain as the benchmark, the image fusion algorithms can be roughly classified as transform-based and spatial-based methods [13]. In the transform domain, the resource images are decomposed into a certain number of layers with proprietary multi-scale filters. Afterward, the coefficients from different images in the same layer are merged according to some manual fusion rule. Finally, the fusion image is obtained by the inverse transformation. There are two key operations in the transform-based fusion methods: filter setting for reasonably capturing salient feature information and fusion rule selection for reasonable weight coefficient distribution. In order to heighten the performance of fusion algorithm, researchers design

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

or upgrade mass filters to improve information extraction ability, such as the Laplacian pyramid [14], discrete wavelet transform (DWT) [15], curvelet transform [16], dual-tree complex wavelet transform (DTCWT) [17] and non-subsampled contourlet transform(NSCT) [18], etc. Although the fusion performance of these algorithms is observably improved with optimized filter banks, the fixed manual filter, be appropriate to specific structural feature information, still missing some available characteristic information from resource image. In addition, the fusion rule selection is indispensable for assigning appropriate weight to transform coefficients, which can easily lead to feature information loss or distortion, and make the fusion image suffer from low-contrast or blurring [19]. In contrast, the spatial-based algorithms directly integrate the gray information from resource images under a certain standard. These kinds of methods can easily cause image distortion and make the fusion image quality reduced.

In recent years, deep learning has achieved rapid development in the field of image processing, such as image classification, image super-resolution, object identification, and so on [20], [21], [22], [23]. Attributing to its advancement and handleability, deep learning with convolutional neural network (CNN) also yields brilliant results in the image fusion task [24]. Li et al. [25] firstly introduced CNN into image fusion task, and took the lead in applying deep learning for combining visible images with infrared images [26]. They firstly decomposed the source image into basic parts and content of details. Then the basic part was fused by weighted average, and the details is merged by feat of deep neural network. Finally, the two components are accumulated directly to obtain a comprehensive fusion image. Subsequently, multifarious CNN-based fusion models are proposed for special purpose. Li et al. [27] extracted the deep features by ResNet, and used the normalizing deep features with Zero-phase component analysis (ZCA) to acquire the initial weight coefficient. Then, the weight map for integrating image feature information is refined by SoftMax operation. And then they put it separately nest connection and spatial/channel attention models and an end-to-end residual fusion network for infrared and visible images [28], [29], which boost immensely fusion performance of algorithms by improving network structure. Zhang et al. [30] proposed a general image fusion framework based on the convolutional neural network. Inspired by the transform-domain image fusion algorithms, they introduced the concept of multi-scale to convolution kernel and achieved comparable or even better image fusion results. However, the elementwise fusion rules have been utilized to fuse the convolutional features of multiple inputs, which will undoubtedly lead to the loss of some feature information. Tang et al. [31] designed a pixel convolutional neural network for multi-focus image fusion, but the decision map for information integration was obtained by comparing the values of the two score matrixes. This kind of method is effective for multi-focus image fusion seemingly. Ren et al. [32] proposes a novel infrared and visible images fusion method based on improved DenseNet, Max-Relevance and Min-Redundancy and zero phase component analysis. The fusion strategy is be optimized by elaborating activity level maps based related feature processing. Si et al. [33] proposed a dual fusion path generative adversarial network for infrared and visible image fusion, and implemented dual self-attention feature refine module (DSAM) on two fusion paths to refine feature maps in two fusion paths. This kind of targeted design improved distinctly fusion image contrast. Although they can obtain good fusion effects to a certain extent, the complexity of pre-processing and the instability of fusion rules and limited their practicalapplication. Based on the above several fusion cases, most of the current CCN based fusion algorithms is mainly based on the idea of image classification or segmentation to achieve information fusion. However, the characteristic of VI image and multi-focus image have obvious differences, VIF cannot simply be regarded as an image classification task. Besides, with the lack of ground truth for visible and infrared images, the trained network model is limited to retain the enough useful information of source images. In addition, it's pretty obvious that a multitude of of the existing CNN based methods are universally need to design a special feature weight allocation method and fusion rule similar to the transform-based methods. Which is clearly not an easy task for sufficient features fusion, because single manual weight coefficient and feature map integration method is not always effective for complex infrared and visible feature information and inevitably overshadows their image fusion performance. Furthermore, a mass of details may be lost randomly and be difficult to be preserved in the final fusion image due to the pooling process.

Responding to the above problem, an efficient visible and infrared image fusion network model is proposed with ingenious network structure design in this paper. The proposed network structure is shown generally in Figure 1, and the main contributions and innovations of the paper can be summarized as follows:

1) A multi-scale convolutional fusion model with an improved residual block is proposed, which can explicitly integrate deep features without manual weight selection and fusion rule design and have remarkable adaptivity.

2) Compared with the popular training dataset derived from low-resolution images or unrealistic ground truth fusion images, this paper utilizes high-resolution multi-focus images with ground truth images as the training dataset for infrared and visible image fusion, which optimizes the upper bound of the network performance, and conduce to the loss function constrain the network focus on the informative regions of source images effectively to facilitate the retention of the useful information.

3) An effective image reconstruction structure combining affluent skip connections with multi-scale convolutional layers is designed to supplement the lost image details in the

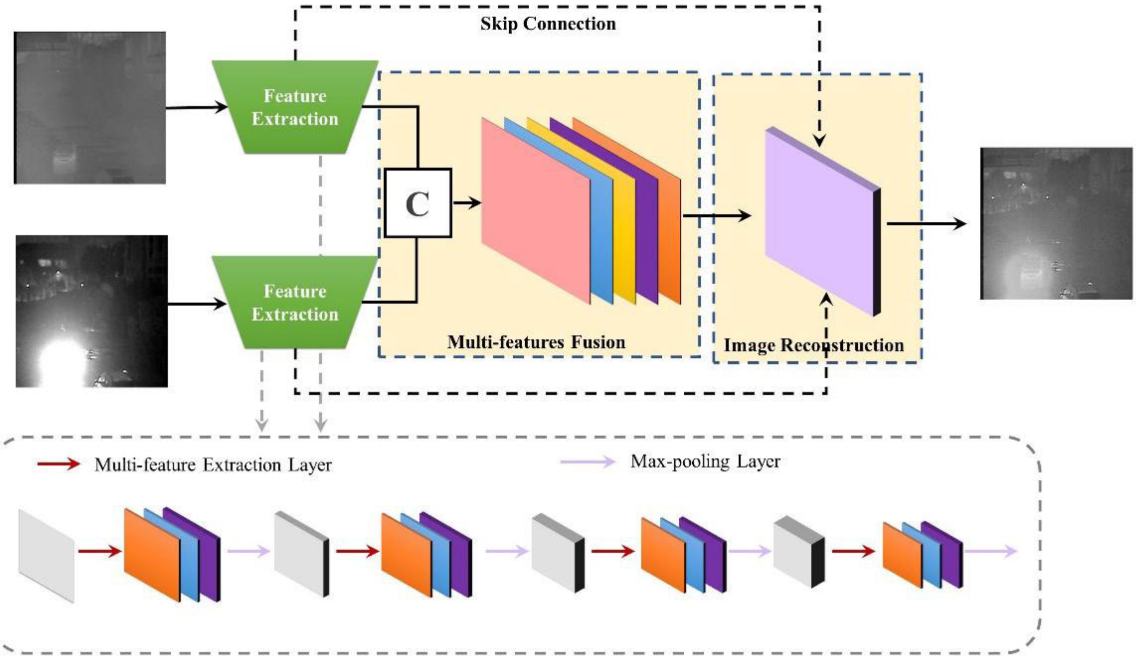


FIGURE 1. The architecture of the proposed network.

pooling process and improve the utilization rate of the image convolution features.

Therefore, the proposed model is a fully convolutional neural network, which is trained in an end-to-end manner without preprocessing, and the complete trained deep CNN properly integrated learn feature extraction, fusion and reconstruction components together to produce reasonable fusion result for visible and infrared images.

The rest of this paper is as follows. In Section II, the related work about deep learning including the image training dataset and the loss function is described. In Section III, according to the experiment results with the reference algorithms, right-minded analysis is presented from both sides of subjective and objective evaluations. Section IV draws the conclusions.

II. RELATED WORK

As we all know that training datasets and loss function selection are directly related to the accuracy of the neural network model. The impact and selection mode of training datasets and loss function will be introduced in this subsection.

A. TRAINING DATASET

On account of that the quality of the training dataset often directly determines the upper bound of the model performance [30], the more training samples and types subsequently the higher the image quality of the training results. For this reason, Tang et al. [31] chose Cifar-10 as the training dataset, which contains 60,000 image blocks of size 32 × 32. Lai selected about 45,000 detail-rich images from ILSVRC 2015 as the original dataset, and these images were then uniformly divided into image blocks of size 128 × 128 for training. In fact, most of the datasets commonly used for neural network training nowadays are composed of small image blocks (32 × 32, 64 × 64) currently. Although small image blocks are beneficial to improving model training time, the model performance also is greatly restricted for the reason that their resolution is low. Therefore, For the sake of promoting the performance of neural model, the image block size is dataset to 256 × 256 in this paper.

Furthermore, because of lacking ground-truth fusion images, searching for appropriate training dataset for VIF is challenging. In order to supervise the image fusion models preferably, Zhang et al. [30] used multi-focused images as the training dataset and obtained desirable fusion images. Homoplastically, Fang et al. [34] selected multiple modal images as training datasets and also achieved satisfactory results in the field of VIF. The different out-of-focus ways and sample richness of multi-focused images can improve the stability of network structure [30]. Therefore, the training dataset for multi-mode image fusion can be confirmed by handling felicitously natural images with ground-truth fusion images, such as multi-focused images. As described above, reasonably segmented multi-focus image, which is more easily generated and has ground-truth fusion images, is chosen as the training dataset in this paper. The multi-focused images are produced by employing 18,800 images deriving from the COCO dataset [35].

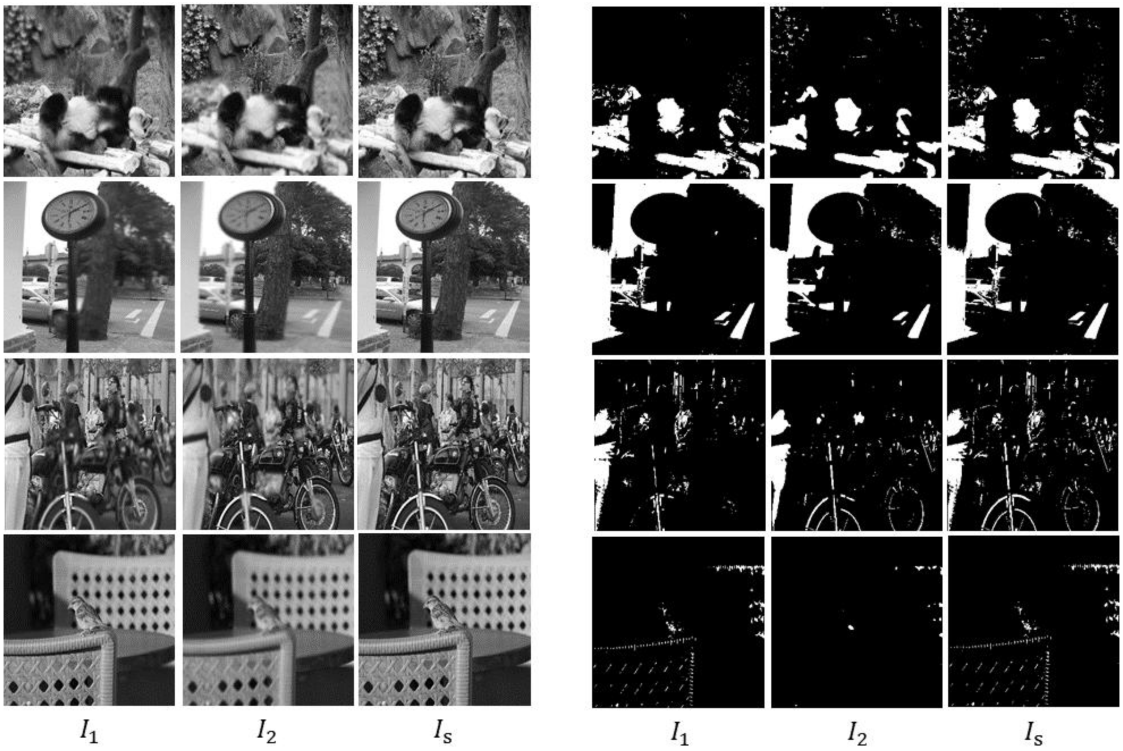


FIGURE 2. **Several image training datasets. The source image is on the left. The right side is the homologous binary image. I_1 and I_2 are multi-focused images, and I_s is the ground truth fusion image.**

The specific steps generating multi-focus images are shown as follows:

Step 1: The complete blurred image I_g is generated by randomly blurring source image I_s with a gaussian filter, which can be expressed as:

$$I_g = G * I_s \tag{1}$$

here '$*$' denotes convolution operation, G denotes Gaussian kernel, and the random kernel radius is from 0 to 30 pixels according to (2).

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

here σ can express the standard deviation of gaussian filter.

Step 2: The edge information of source image I_s are acquired by Otsu algorithm [36]. The algorithm acquires the best threshold of the image by the inter-class variance method, and distinguishes the background of the image from the target edge information. Then the edge information is expanded into region blocks by morphological expansion, and the focus map If is gained. The equation of Otsu algorithm to get the optimal threshold value is denoted as follows.

$$T = w_0 w_1 (u_0 - u_1)^2 \tag{3}$$

T is the optimal image threshold, w_0 represents the proportion of target points to the image, and u_0 represents the average gray value of the target points. w_1 represents the proportion of background points to the image, and u_1 represents the average gray value of the background points.

Step 3: A pair of multi-focused images are generated based on source image $\boldsymbol{I_s}$, blurred image $\boldsymbol{I_g}$ and focus map $\boldsymbol{I_f}$. The focus maps I_1 and I_2 can be fixed according to (4).

$$\begin{cases} I_1 = I_s \bullet I_f + I_g \bullet (1 - I_f) \\ I_2 = I_s \bullet (1 - I_f) + I_g \bullet I_f \end{cases} \tag{4}$$

where **1** represents a matrix, whose size is consistent with source image and all values are 1. '•' denotes dot product operation between matrices.

Because the focus area is more random in this paper, these generated focused images are more natural compared to those that are synthesized by partial data as a whole. Figure 2 shows several datasets of multi-focus images and their ground-truth fusion images, the training dataset acquired by the above method possesses two advantages over other manners: (1) higher image resolution; (2) more diverse blurring styles.

B. LOSS FUNCTION

The aim of image fusion is to reasonably combine salient feature information from source images into an informative and comprehensive image. However, frequent difference prediction between output data and real data is executed by using loss function in the network training process, more likely to lead to unexpected information loss in regression when employing illogical loss function for a certain type of image. Therefore, before implementing deep learning on visible and infrared images, it is necessary to ascertain appropriate loss functions to optimize the parameters of the neural network model for grabbing more abundant textural features from source images.

Mean squared error (MSE) is universally used as loss function to adjust the model predictions close to the truth output in various natural network algorithms. However, it causes a common problem that fusion results in fewer details or is too smooth for visible and infrared images [25]. In view of that infrared image and visible image acquired from the same scenario containing lots of similar structural information, we choose structural similarity loss (SSIM) [37] as loss function to optimize the parameters of the natural network.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

where x is the real image, y is the predicted image, μ_x, μ_y is mean, σ_x, σ_y is variance, and σ_{xy} is covariance. C_1=$(Lk_1)^2$, C_2=$(Lk_2)^2$ are stable constants. L is the dynamic range of pixel values, $k_1 = 0.01$, $k_2 = 0.03$.

when the two images are converged, *SSIM* gets close to 1. conversely, *SSIM* is near to 0. Thus, the loss function is defined as follow.

$$I_{SSIM} = 1 - SSIM(x, y) \quad (6)$$

C. CNN FOR IMAGE FUSION

Through literature review, it is found that the existing CNN image fusion technology has many applications in infrared and visible image [27], [28], [29], but it is more aimed at multi-focus image fusion [20], [21], [24], [25], [31], [35]. The main reason for this phenomenon is CNN is easily applied to image classification or segmentation task in image analysis based on its special convolutional characteristic. It's well known that the key point in multi-focus image fusion methods is seeking the optimized focus measure (FM,) which can be regarded as a classification problem that discriminates focused and defocused maps. Enlightened on this ideal, some researchers used the convolution property of CNN to learn the effective FM for elaborate focus map, and greatly improved multi-focus image fusion performance. For example, Liu et al. [25] introduced CNN as a sorting task to fuse multi-focus images at the first time. Tang et al. subsequently learned a CNN model joined activity level measurement and fusion rule to combine multi-focus images [31]. In order to refining the focus map without post-processing, Zhang et al. designed an end-to-end fully convolutional neural network and achieved state-of-the-art results [30]. Amin et al. integrate three CNNs models to construct the optimized segmented decision map for multi-focus image fusion [35]. Du and Gao [20] introduced segmentation ideal to construct multi-focus image fusion model. Due to the multi-focus image has obvious blurred and clear areas, the above convolutional neural networks based various methods have get astounding achievements in multi-focus image fusion. However, there are obvious differences between infrared and visible images deriving from their imaging modes. For instance, visible images mainly exhibit the rich details and high spatial resolution but weaken momentous target even silently. Whereas infrared images highlight salient target from background but lack texture details. In addition, these features usually overlap in different areas between infrared and visible images. Therefore, the fusion task for infrared and visible image can't just be seen as a simple image classification or segmentation.

According to the above reasons, some deep learning methods suitable for infrared and visible image fusion are studied on account of network structure design and image features analysis. For example, Li et al. [27] extracted more deep features by residual network, and used the normalizing deep features with Zero-phase component analysis (ZCA) to acquire the initial weight coefficient. Then, the weight map for integrating image feature information is refined by SoftMax operation. And then they put it separately nest connection and spatial/channel attention models and an end-to-end residual fusion network for infrared and visible images [28], [29], which boost immensely fusion performance of algorithms by improving network structure. Jian et al. [38] overcome the information redundancy by a symmetric encoder-decoder block network but the middle layer information is ignored. In order to retain significant infrared targets, Ma et al. [39] proposed an image fusion network based on the salient target detection, and the target regions could be marked by the salient target mask similar to a classifier. Inspired by the transform-domain image fusion algorithms, Zhang et al. [30] introduced the concept of multi-scale to convolutional neural network and achieved comparable or even better image fusion results. However, the elementwise fusion rules have been utilized to fuse the convolutional features of multiple inputs, which will undoubtedly lead to the loss of some feature information. Ren et al. [32] proposes a novel infrared and visible images fusion method based on improved DenseNet, Max-Relevance and Min-Redundancy and zero phase component analysis. The fusion strategy is be optimized by elaborating activity level maps based related feature processing. Si et al. [33] proposed a dual fusion path generative adversarial network for infrared and visible image fusion, and implemented dual self-attention feature refine module (DSAM) on two fusion paths to refine feature maps in two fusion paths. This kind of the instability of fusion rules and limited their practicalapplication.

Based on the above-mentioned representation, the current CCN-based fusion model for infrared and visible image is universally needs to design a special feature weight allocation method and fusion rule similar to the transform-based methods. Which is clearly not an easy task for sufficient

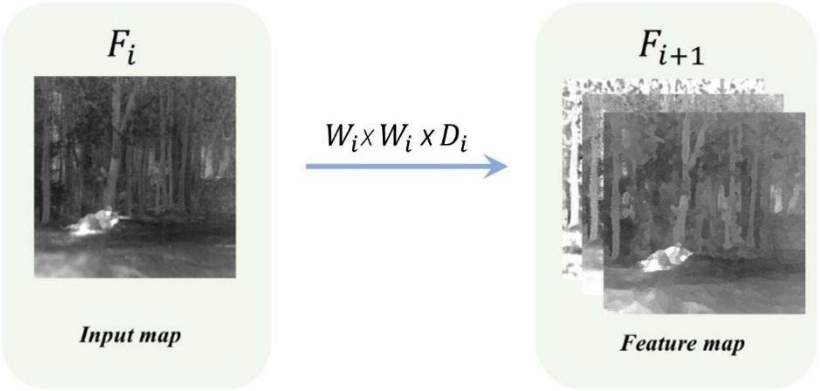


FIGURE 3. General convolutional layer structure. F_i is the input image, F_{i+1} is the output image. W_i is the convolutional kernel size, and D_i is the number of channels.

features fusion, because single manual weight and fusion rule is not always effective for complex infrared and visible feature information and inevitably overshadows their image fusion performance. Meanwhile, some feature information is easily lost because these network models lack sufficient feature extraction and retention ability. Therefore, we propose a new CNN-based fusion method (MLCNN), which introduce improved residual block to multi-scale convolutional fusion model for confirming the weight map and fusion mode adaptivity, and combine skip connection with multi-scale convolutional layer adequately to supplement the lost image details in the pooling process and improve the utilization rate of the image convolution feature information. Based on this, The proposed model can fully demonstrate the data mining capabilities of convolutional neural networks to extract enough deep features and preserve more meaningful details in model training, simultaneously realize the integration of depth features without manual fusion rule adaptively.

III. PROPOSED FUSION METHOD

As reported in previous literatures, multi-layers convolutional filters own superior ability to traditional multi-scale filters in feature information extraction [40]. Exhilarating, weight coefficients for integrating source images can be acquired and optimized adaptively by convolutional filters. Oppositely, weigh coefficients only be fixed stiffly through pre-set fusion rule in transformation domain. Therefore, inspired by the idea of multi-scales decomposition and the resounding success of IFCNN, an efficient visible and infrared image fusion model based on multi-layers convolutional neural network (abbreviated as MLCNN) is proposed, which is end-to-end fully convolutional structures without preprocessing and has great adaptability for determining weight coefficients. Similar to the image fusion process based on multi-scales decomposition, MLCNN can be divided into three components roughly according to the role of each part: multi-scales feature extraction strategy, feature fusion strategy and image reconstruction strategy, as shown in Figure 1. The specific details of each strategy in MLCNN are explained in subsequent subsections.

A. MULTI-SCALES FEATURE EXTRACTION STRATEGY

As shown in Figure 3, The convolutional layer, which is the core of CNN, can pick out the feature information of image with the help of training dataset. Therefore, reasonable and appropriate convolution kernel (CK) is critical for feature extraction. It is interesting to note that CK with small size is sensitive to low-frequency and small detail information, and CK with large size is favorable for capturing high frequency and large detail information [41]. According to the above facts, a multi-features extraction block (MFE), multiple sizes CK are inserted in one convolutional layer, is introduced in this paper. The specific structure of CK is shown simplistically in Figure 4.

In order to extract the low and high features dividually and specifically, the sizes of CK are set as 3 × 3, 5 × 5 and 7 × 7 in our network model, respectively. The feature maps of individual input source images are subsequently concatenated along the number of channels. Longitudinal well-known, the CK with large size need the network to train more parameters, which means more time and slowdown algorithm speed. For the sake of improving execution efficiency, the convolutions with sizes 7 × 7 and 5 × 5 are converted into three connected 3 × 3 convolutions and two connected 3 × 3 convolutions severally, which can greatly reduce the number of parameters and speed up the network training efficiency [42]. For more refined extraction of image features, four MFEs are incorporated, and one bias-corrected linear unit (ReLU) layer is added after each convolution, which increases the nonlinear relationship between the layers and reduces the dependence between parameters in this paper. In consideration of the fact that the input training images are large, a max-pooling layer of size 2 × 2 is carried out after each MFE to decrease training parameters by simplifying image size to optimize the model performance.

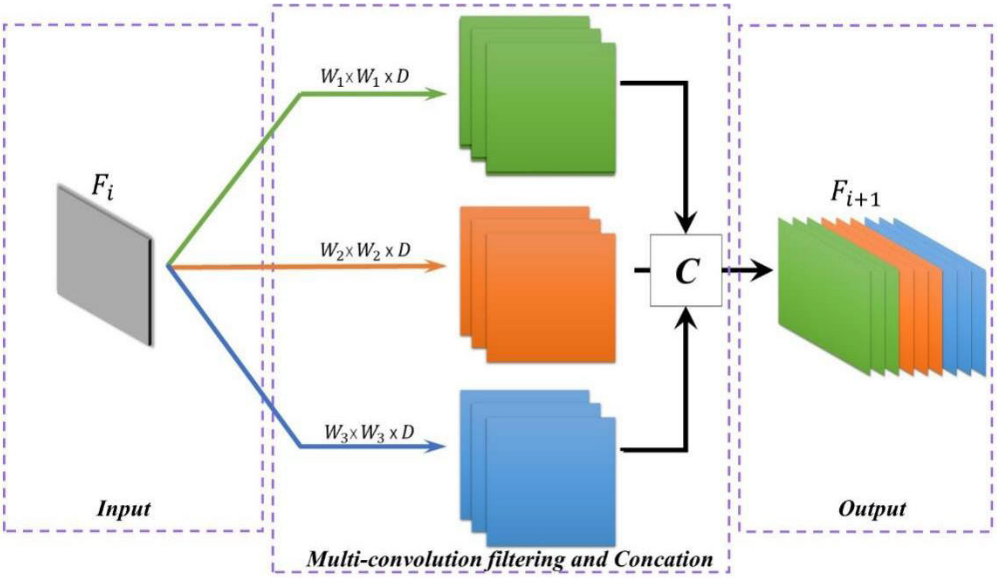

F_i
$W_1 \times W_1 \times D$
$W_2 \times W_2 \times D$
$W_3 \times W_3 \times D$
C
F_{i+1}
Input
Output
Multi-convolution filtering and Concation

FIGURE 4. Multi-features extraction structure. F_i is the input image, F_{i+1} is the output image. W_1, W_2, W_3 are different size convolution kernels, and D_i is the number of channels.

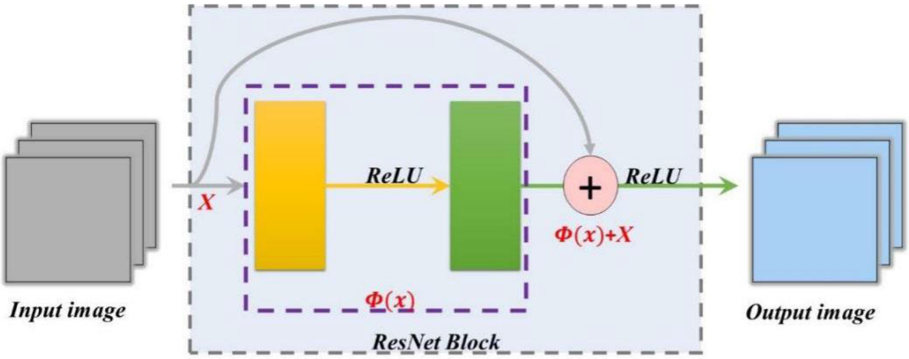


FIGURE 5. Structure of the residual block.

B. MULTI-SCALES FEATURE FUSION STRATEGY

After the features exaction operation is complete, two columns of multi-dimensional and multi-scale features from infrared and visible images are identified, respectively. Although the two sub-networks own similar architecture, their corresponded feature maps are different. Therefore, it requires a reasonable method to integrate the sub-networks with convolutional features of two images. In the field of CNN based information fusion, researchers usually adopt the below two tactics to integrate these convolutional features: (1) the same layer convolutional features from different images are firstly concatenated along the channel dimension, and then the convolutional features after stacking of dimensions are consolidated by a proper convolution, (2) the same layer convolutional features from different images are straightway confirmed by the elementwise fusion rules (such as elementwise-maximum, elementwise-sum and elementwise-mean) [30]. Although the elementwise fusion rules are used widely in CNN based information fusion, they may lead inevitably to submerge or smooth some useful important features from source image, which makes the fusion image appear halo or jitter [26]. In view of the diversity of image backgrounds and details, the adaptivity of this tactic is not ensured. Hence, to prevent the artificially selected fusion strategy from degrading the performance and adaptiveness of the proposed model potentially, concatenation method is utilized to integrate the extractive convolutional features in this paper.

When reducing the dimension of the feature map after series connection, it will cause partial feature information being lost or overwhelmed. In order to refrain from the above problems in the training process, ResNet network [27], as shown in Figure 5, is employed in this paper. The introduction of residual blocks can achieve stable cross-channel information fusion to a certain extent, which is in favor of reducing information loss. Furthermore, with the help of ResNet network, the input information can be directly flowed from any low layer to high layer in forwarding propagation,

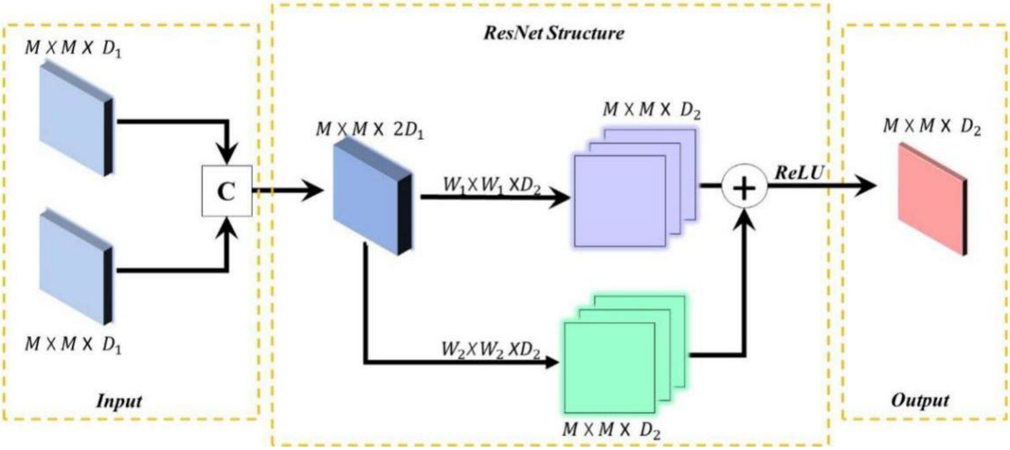


FIGURE 6. Feature fusion structure. *M* is the image size, W_1, W_2 are different size convolution kernels, D_1 and D_2 are the number of image channels, respectively.

which can be beneficial to avert network degradation. Meanwhile, the error information can be directly transferred to the lower layer without any intermediate weight matrix transformation in backpropagation, which can avail against the gradient disappearance problem heavily. To sum up, it can be concluded that ResNet network makes the forward and backward propagation more unhindered and makes the ability to capture deep feature information stronger.

The specific feature fusion structure with ResNet network is shown in Figure 6. a module similar to the residual block structure is added to the multi-feature fusion strategy, which not only can be well to avoid the problem of feature detail loss, but also can deal with the problem of network gradient disappearance. The structure of the residual block is expressed as:

$$\Phi_{i+1} = g((W_i + 1) * \Phi_i + b_i) \tag{7}$$

where W_i and b_i denote convolution kernel and weight of the i-layer, respectively. $\boldsymbol{\Phi_i}$ is the output feature map of the i-convolution-layer, and $g(\cdot)$ represents the activation function. '$*$' *denotes* the convolution operation.

C. IMAGE RECONSTRUCTION STRATEGY

The image reconstruction strategy as shown in Figure 7 consists of four trainable convolutional layers. In consideration of the fact that the max-pooling layer can reduce the size of the image during the training process of multi-scales feature extraction, an up-sampling operation, gradually restore the pooling layer to the source image size, is performed on the fusion layers by means of the transposed convolution layer. Transposed convolution layer equations are derived as (8)–(10), shown at the bottom of the next page.

X, Y represent the input and output image matrices (square matrix), m and n represent the matrix scales, and m=$n/2$. $\boldsymbol{K}$ represents the convolution kernel parameters of the transposed convolution layer. $\boldsymbol{C}$ represents the sparse matrix of $\boldsymbol{K}$ and $\boldsymbol{C^T}$ represents the matrix transpose.

The transposed convolutional layer can only restore the source size of the output image, but cannot recover the image pixel values. To solve such problems, the network structure is optimized by skipping connection linking the multi-scales feature extraction layer with the reconstruction layer, which is conducive to supplementing missing details in the pooling process and reserves the edge information from the source image, as well as avoiding gradient disappearance.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENT PREPARATION

In order to validate adequately the effectiveness of the proposed convolutional neural network model, twelve groups of infrared and visible images as shown in Figure 8 are used to test the proposed algorithm. These tested images are acquired in different experimental environments, which can be sufficiently used to demonstrate the stability and adaptability of the proposed algorithm. Meanwhile, substantial subjective and objective analyses are given with eight state-of-art referenced image fusion algorithms. These comparison algorithms respectively are deep learning (DLF) [26], residual neural networks (ResNet) [27], RfnNet [28], NestNet [29], convolutional neural networks (CNN) [41], guided filtering (GFF) [43], gradient transformation and variance minimization (GTF) [44], anisotropic diffusion (ADF) [45], multi-resolution singular value decomposition (MSVD) [46], salience-based method (TIF) [47], and hybrid model (VSMWL) [48]. The parameter setting of all reference algorithms is strictly consistent with the original literature. All algorithms used in this paper are executed on the same computer with Intel i5-1035G1 CPU (1 GHz) and 2 GB GPU. The proposed fusion model for short MLCNN is achieved by Pytorch 1.8.1 based on Python 3.9.4. 18800 pairs of multi-focused images are trained with an image size of 256 × 256 and a batch size of 32 in the training process, and the learning rate was set to 0.001 using the Adam optimizer [39].

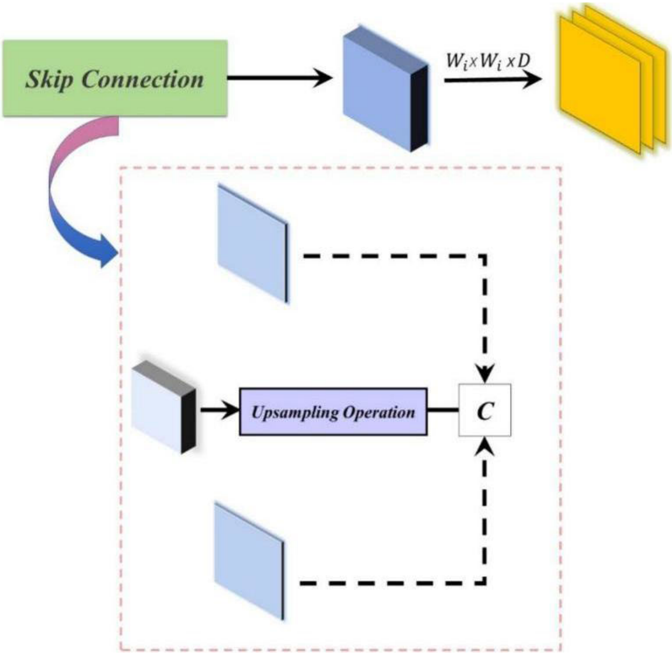


FIGURE 7. Image reconstruction strategy. D is the number of image channels. W_i is the convolution kernel size.

B. OBJECTIVE EVALUATION METRICS

Objective evaluation is important measure to evaluate image fusion quality besides subjective visual analysis, it can effectively make quantitative comparisons based on the characteristics of fusion images. At present, plentiful objective evaluation criteria have been proposed in allusion to different types of image quality analysis. In consideration of the fact that ground-truth fusion image for the visible and infrared image fusion task does not exist, in order to reveal details and other characteristic information of the fusion images and verify the performance of the proposed fusion model, five objective image metrics, such as average gradient (AG), information entropy (IE), space infrequency (SF), edge information retention ($Q^{AB/F}$) and Piella [44], [49] are adopted to reveal the quality of various fusion results. The larger evaluated values of the above five metrics illustrate that the corresponding fusion results contain more valuable information.

C. RESULTS AND DISCUSSION

Limited to the paper length, the fusion results of four groups of images with obvious feature differences as shown in Figure 9, “Car” (a1 and a2), “Human” (b1 and b2), “Wilderness” (c1 and c2), and “Factory” (d1 and d2) under the algorithm mentioned above will be discussed in detail in this section. According to the fusion results acquired from various algorithms, this paper presents a comparative analysis from the perspective of visual effects and objective

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nn} \end{pmatrix}_{n\times n} \quad K = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \tag{8}$$

$$C = \begin{bmatrix} w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} \end{bmatrix} \tag{9}$$

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mm} \end{pmatrix}_{m\times m} \quad C^T \times \begin{bmatrix} y_{11} \\ \vdots \\ \vdots \\ y_{mm} \end{bmatrix} = \begin{bmatrix} x_{11} \\ \vdots \\ \vdots \\ x_{nn} \end{bmatrix} \tag{10}$$

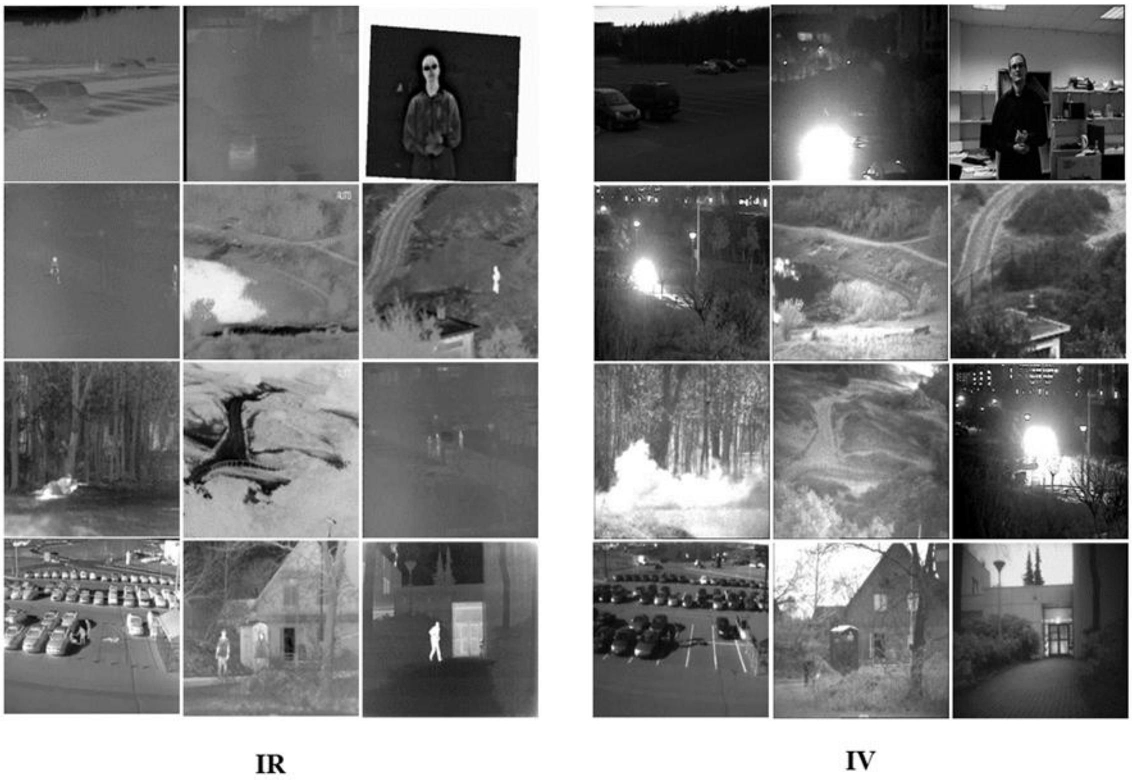


FIGURE 8. Test images. IR indicates infrared images and IV indicates visible images.

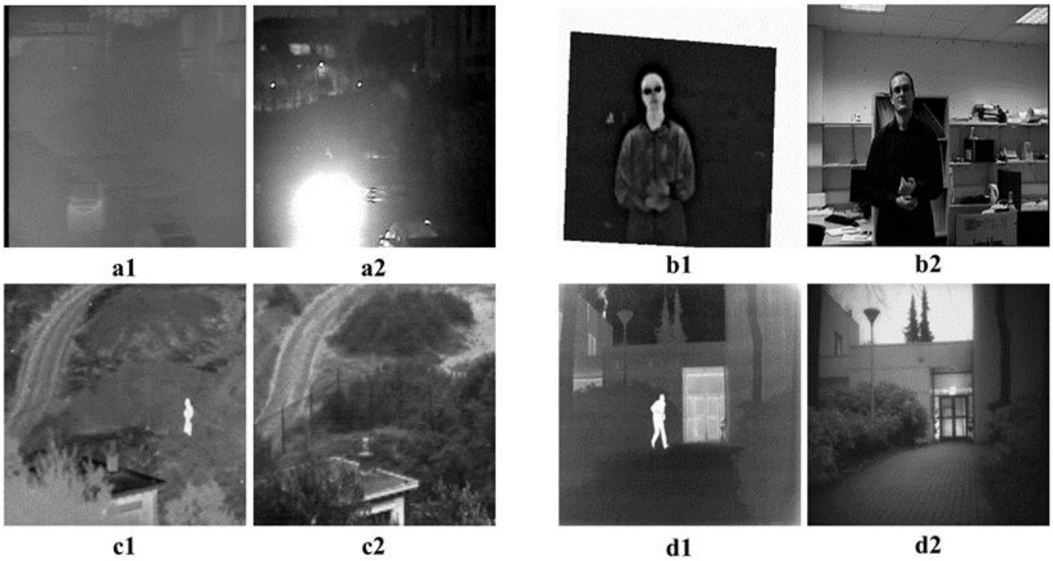


FIGURE 9. Visible and infrared original image. a1 is the infrared image "Car"; a2 is the visible image "Car"; b1 is the infrared image "Human"; b2 is the visible image " Human "; c1 is the infrared image " Wilderness "; c2 is the visible image " Wilderness "; d1 is the infrared image "Factory"; d2 is the visible image "Factory."

evaluation. The best values, the second-best values and the third-best values are indicated in bold, red and italic and blue and italic in Table 1- Table 5, respectively.

The fusion results of twelve algorithms are shown respectively in Figure10-13 to validate the effectiveness of the proposed CNN-based method. Firstly, we compare the

TABLE 1. Comparison for "Car" images fusion.

Methods	AG	IE	SF	$Q^{AB/F}$	Piella
ADF	2.1883	6.8872	7.9733	0.3780	0.7891
CNN	2.8337	**7.6155**	*12.0159*	0.4495	*0.8550*
DLF	1.6610	6.8836	7.0614	0.3275	0.7820
GFF	*2.8178*	*7.5776*	*11.8627*	0.4504	**0.8580**
GTF	2.1648	5.6073	10.2152	0.3634	0.7101
TIF	*3.6260*	6.9788	11.3634	**0.5746**	*0.8308*
VSMWL	*3.0638*	7.0961	8.7826	*0.4836*	0.8186
MSVD	1.9117	6.9434	10.6169	0.3236	0.7440
ResNet	1.6548	6.9573	6.6944	0.3100	0.7895
NestNet	2.1171	6.6930	9.3174	0.5024	0.7755
RfnNet	1.7234	*7.3201*	4.9738	0.3508	0.7875
MLCNN	**3.6718**	7.0129	**14.8318**	*0.4769*	0.7245

TABLE 2. Comparison for "Human" images fusion.

Methods	AG	IE	SF	$Q^{AB/F}$	Piella
ADF	5.6494	7.3395	20.9517	0.4560	0.7039
CNN	7.4086	6.7905	**30.9665**	0.5607	0.6171
DLF	5.0604	7.3104	19.5996	0.4576	0.7118
GFF	*7.9078*	*7.5540*	*29.6399*	*0.5922*	*0.7278*
GTF	6.2181	6.3024	25.6765	0.4906	0.6529
TIF	**8.8612**	**7.5729**	25.9355	0.5456	0.7063
VSMWL	6.2099	7.0900	17.3010	0.4008	0.6934
MSVD	3.8512	6.8841	17.8932	0.1984	0.6039
ResNet	4.7735	7.3184	17.5955	0.3992	0.7028
NestNet	5.5057	7.3277	23.3394	*0.5873*	*0.7485*
RfnNet	3.6173	7.2686	10.2403	0.2360	0.6269
MLCNN	*7.4602*	*7.4229*	*28.1599*	**0.6064**	**0.7535**

TABLE 3. Comparison for "Wilderness" images fusion.

Methods	AG	IE	SF	$Q^{AB/F}$	Piella
ADF	3.7033	6.2868	8.4285	0.4023	0.7389
CNN	*4.9683*	**7.1058**	*11.6710*	*0.4546*	0.7679
DLF	2.9608	6.2398	6.7696	0.3224	0.7681
GFF	4.2957	6.5102	10.6094	*0.4529*	*0.7736*
GTF	3.8206	6.6546	8.9461	0.3859	0.6917
TIF	**5.3440**	6.5346	10.5581	0.4291	0.7725
VSMWL	*5.3439*	6.7307	10.6725	0.4478	*0.7822*
MSVD	4.0434	6.8406	10.9806	0.3443	0.7252
ResNet	2.9167	6.2360	6.5717	0.3142	0.7665
NestNet	4.7099	*6.9810*	*11.1478*	**0.4754**	0.7077
RfnNet	3.7702	*6.8665*	8.0330	0.4037	0.7269
MLCNN	4.6200	6.4265	**11.7358**	0.4095	**0.7971**

TABLE 4. Comparison for "Factory" images fusion.

Methods	AG	IE	SF	$Q^{AB/F}$	Piella
ADF	2.3664	6.4797	7.1348	0.3079	0.7581
CNN	4.3038	**7.2346**	*13.1618*	*0.4619*	0.7126
DLF	2.4116	6.4870	7.4865	0.3136	0.7631
GFF	3.9155	6.8706	12.0853	*0.4729*	*0.7658*
GTF	2.8214	6.9261	9.0672	0.3379	0.7200
TIF	*4.8981*	6.6880	10.8244	0.4300	*0.7703*
VSMWL	**5.1597**	6.9362	11.8005	0.4087	0.6265
MSVD	3.5836	6.6609	12.3723	0.3702	0.7464
ResNet	2.3350	6.4718	7.0496	0.2944	0.7594
NestNet	4.1474	*7.2061*	*12.6184*	**0.5466**	0.7102
RfnNet	2.5653	*7.0164*	7.0685	0.3183	0.6256
MLCNN	*4.5861*	6.8445	**13.6943**	0.4446	**0.7742**

fusion performance of different algorithms from the perspective of visual effects and objective evaluation in Figure10. Figure10 exhibits the difference between the proposed algorithm and the reference algorithms. It is intuitive that

TABLE 5. Comparison for images fusion.

Methods	AG	IE	SF	$Q^{AB/F}$	Piella
ADF	5.6318	6.6717	15.8472	0.4318	0.6853
CNN	6.1190	**7.1700**	**18.9763**	*0.5286*	0.7430
DLF	3.7282	6.5560	10.9008	0.3676	0.7228
GFF	*6.1067*	*7.1226*	*18.1517*	**0.5362**	*0.7591*
GTF	5.0479	6.6087	15.4952	0.4272	0.6668
TIF	**7.0533**	6.7872	16.8684	0.5122	*0.7612*
VSMWL	6.1135	6.8087	14.5090	0.4602	0.7425
MSVD	4.1523	6.7881	14.1155	0.3132	0.6825
ResNet	3.6404	6.5637	10.3437	0.3590	0.7210
NestNet	4.8493	6.8768	15.8425	*0.5168*	0.7328
RfnNet	3.4759	*7.0031*	8.8055	0.3565	0.6962
MLCNN	*6.1431*	6.7328	*18.6722*	0.4779	**0.7677**

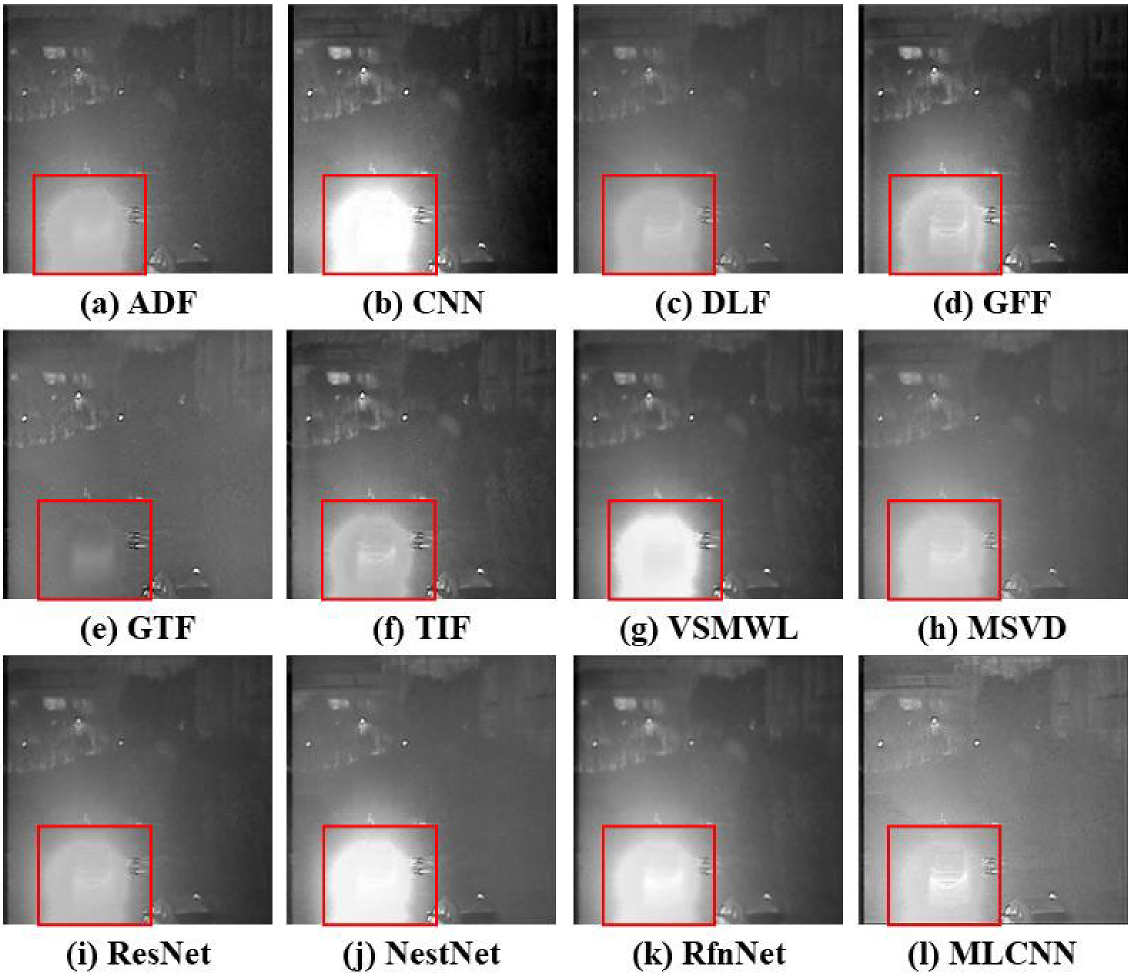

(a) ADF (b) CNN (c) DLF (d) GFF
(e) GTF (f) TIF (g) VSMWL (h) MSVD
(i) ResNet (j) NestNet (k) RfnNet (l) MLCNN

FIGURE 10. Visible and infrared fusion results of the "Car."

the proposed method owns the best visual effect with harmonized visible details and infrared features. Conversely, there are some visible details lost or infrared target annihilation in Figure10 (a)-(e) and (g)-(k). Taking the car in the red box, for example, the car's infrared signature is preserved reasonably in the visible bright light Figure10 (f) and (j). However, the auto-target has almost disappeared in Figure10 (a), (b), (c), (e), (g), (i), (j) and (k), and Figure10 (d), (h), and (i) lost lots of visible details and have obvious halo phenomena in some areas which severely affects human vision.

The values in Table 1 corresponding to various fusion results in Figure 10 can fairly reveal the fusion performance of various algorithms from objective perspective cooperated with subjective visuals. It can be observed that the proposed method is significantly better than the reference algorithms

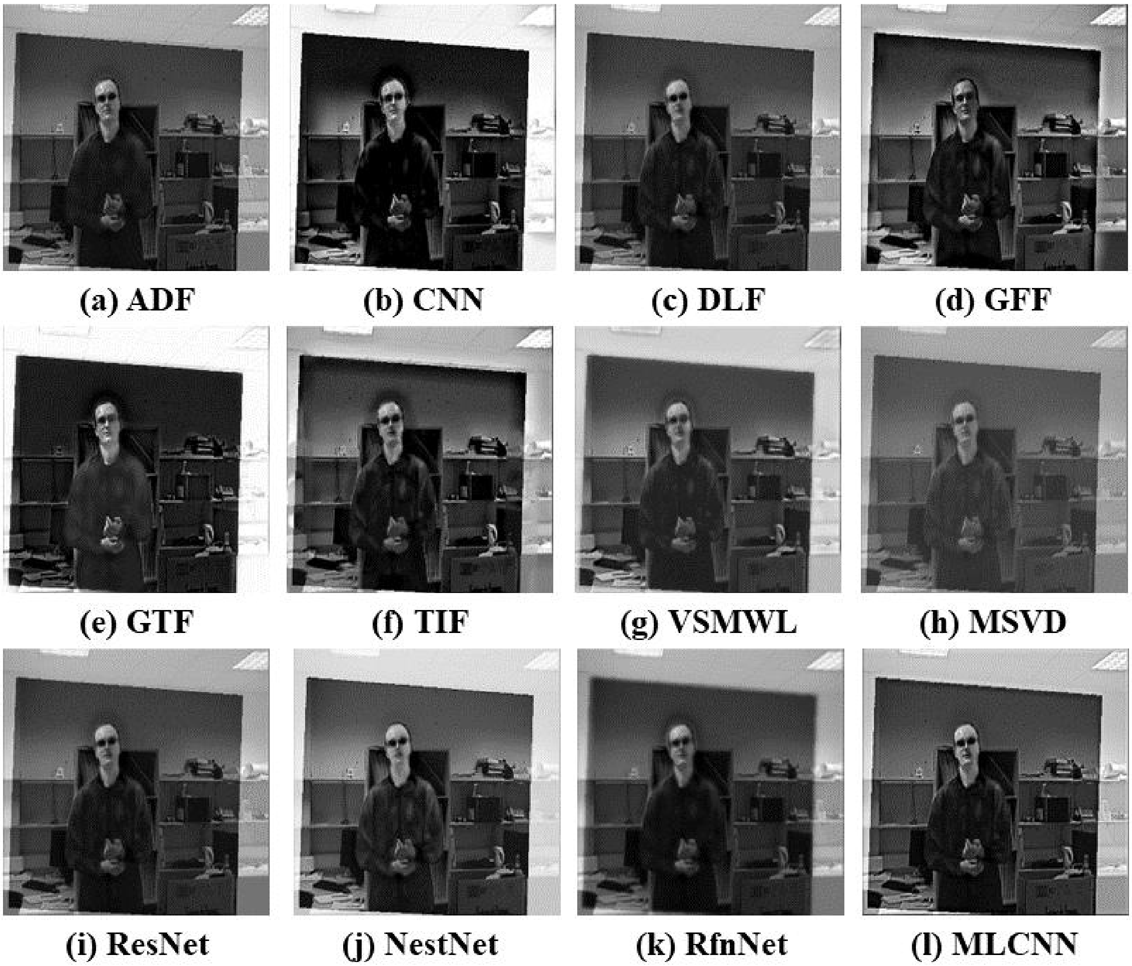


FIGURE 11. Visible and infrared fusion results of "Human."

in terms of AG, and SF values, which indicates that the fusion result of the proposed method possesses rich texture details compared to other algorithms. Although the model has fewer IE, Piella and $Q^{AB/F}$ values than CNN, GFF and TIF, the values that the proposed method achieves are acceptable, which denotes that appropriate information fidelity and evident edge information. By reason of the foregoing, the proposed algorithm acquires better fusion performance for visible and infrared image fusion of "Car" in a comprehensive perspective.

The fusion results of various algorithms with "Human" as resource images are displayed in Figure11. The perfect fusion results should include distinctive human features from the infrared images and clear background details from the visible images. Figure11 (a) (c), (f) and (i) show that ADF, DLF, TIF and ResNet can integrate effectively the available information from resource images to a reasonable extent, but there is a distinct halo around the edge of the person, and the infrared signature tends to dim compared to the source image. Figure11(b), (e) and (j) make clear that is invalid when processing background information, which causes some visible details to be obscured. Although GFF and VSMWL can merge the mutual information among the original images in Figure11(d) and (g), it is visually unnatural due to information distortion in some areas. Obviously, the fusion results presented in Figure11(h) tend to be rayless on account that lots of visible details and infrared futures are equalization. The fusion result based on RfnNet shows dim visible background and also arise distinct halo around the edge of the person, so the visual effects were severely affected. On the whole, the proposed method provide the most observable fusion result, with abundant visible details and infrared features as shown in Figure11(l).

As can be noticed from the objective metrics in Table 2, the proposed method achieved excellent values although some values are lower than others ostensibly. For example, TIF gain greater evaluation value than the proposed method in AG and IE, and CNN get the best SF value, which is mainly caused by the inconsistent fluctuation of image gray level as shown in Figure11. Although the homogeneous values of the proposed method are not optimal, they are obviously higher than those of other reference methods including five kinds of

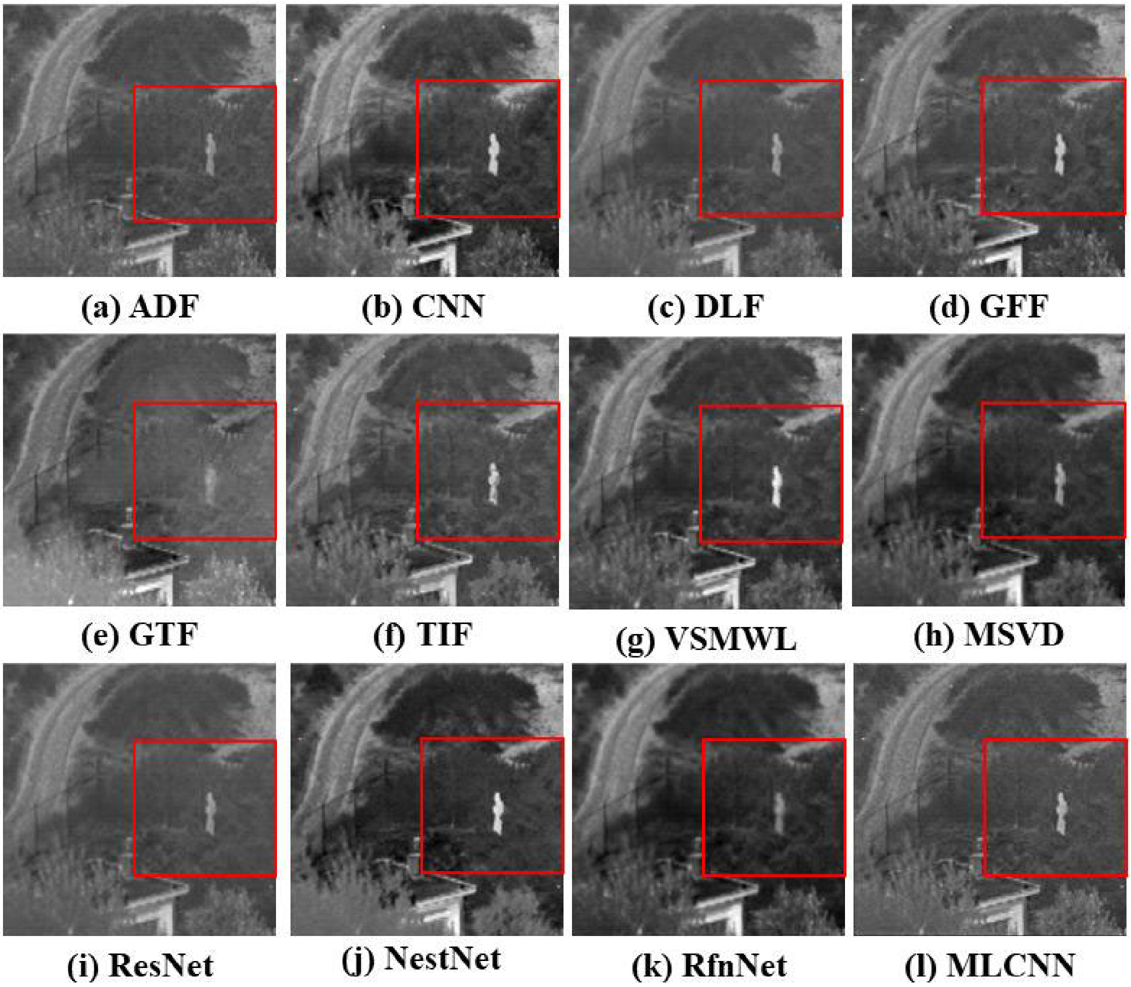


FIGURE 12. Visible and infrared fusion results of " Wilderness. "

classical neural network algorithms. It exhibits that the fusion result of the proposed method is equipped with rich detailed information from resource images. In addition, the proposed method dedicates the best values of $Q^{AB/F}$ compared with all of the reference algorithms, which illustrates that the proposed method can preserve the edge information of the target from source images well. Meanwhile, the Piella value in Table 2 reflects that the fusion result of the proposed method is highly correlated with original images and gets minimal brightness distortion and contrast distortion. In summary, it is evident that the proposed model has clear advantage over the reference algorithm in terms of "Human" VIF.

Figure12 shows the fusion results with various methods on " Wilderness ". The source visible image has abundant texture information such as grass piles, trees and houses, and the matched infrared image reveals prominent target. As shown in the red box of various fusion results, infrared targets are almost lost and visible details are weakened badly in Figure12(e). Although the visible details are reserved to a certain extent in Figure12(a), (c), (i) and (k), the infrared target tends to be dim. In general, the proposed method properly integrates the information from the source images, whose visual effect is as good as the results in Figure12(b), (d), (f), (g) and (j).

Similar to the objective values in Table 1 and Table 2, the proposed method in Table 3 achieved accredited evaluation with subjective vision although some values seem to be lower than other reference methods. The proposed algorithm seizes the best Piella value and SF value, which indicates that the fusion image has the highest correlation with the source images and the optimal brightness and contrast. TIF and CNN obtains the maximum value in AG and IE, which is basically consistent with visual detail perception as shown in Figure12(b) and (f). Although the values of AG, IE and $Q^{AB/F}$ in NestNet are better than that in the proposed method, the image fusion performance of the proposed method is perfectly acceptable. Therefore, it is shown that the proposed model has excellent information integration performance in terms of " Wilderness" VIF in general.

The final specified comparison experiment, as shown in Figure13, takes the visible and infrared images of the "Factory" as the object. It is obvious in terms of visual perception that the proposed method acquires the splendid fusion effect. Figure13(a), (c), (h) and (i) miss some details

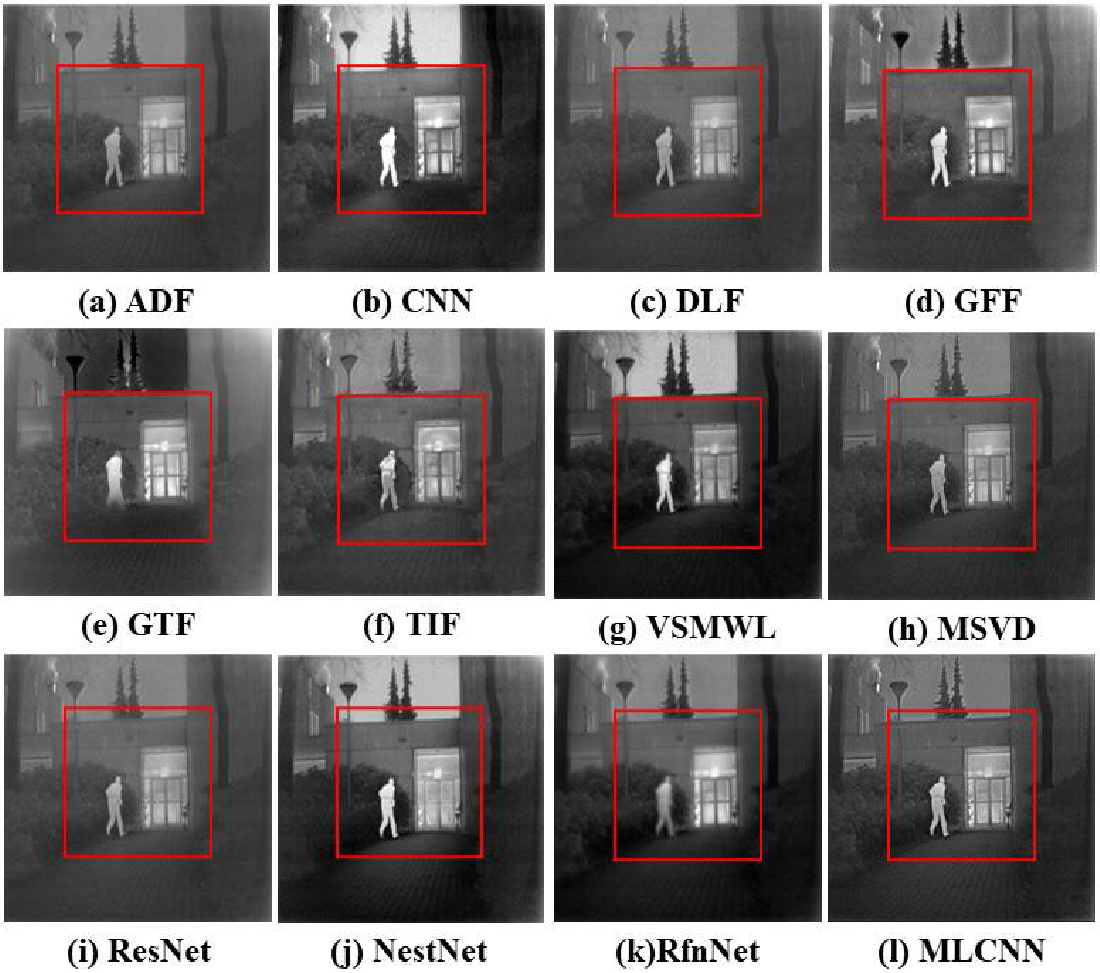

(a) ADF (b) CNN (c) DLF (d) GFF (e) GTF (f) TIF (g) VSMWL (h) MSVD (i) ResNet (j) NestNet (k)RfnNet (l) MLCNN

FIGURE 13. Visible and infrared fusion image of the "Factory". Red boxes indicate highlighted areas of detail.

and suppress the infrared signature. The original infrared information is well preserved in Figure13(b), (g) and (j), but their visible detail background information is tendency to dim which results in inharmonious visual effects. Apparently, serious fusion failure occurs in Figure13(d), (e) and (k), these fusion images result in the loss of important information and severe unnatural distortions. As for Figure 13 (f), the infrared information of the person in the red box appears to display a non-uniform distribution. In conclusion, the proposed method gains the optimal visual perception, which supplies abundant visible details and remarkable infrared features unaffectedly as shown in Figure13(l).

As shown in Table 4, The proposed algorithm still seizes the best Piella value and SF value, which indicates that the fusion image is highly correlated with the source images and its edge information is more abundant. Although NestNet and VSMWL acquire better value in $Q^{AB/F}$ and AG than the proposed method, it is mainly caused by unreasonable visible information loss as shown in Figure12(j) and (g). A similar situation exists in DLF, GFF and GTF. As a whole, the proposed algorithm has better capabilities and more obvious advantages in the VIF of "Factory".

To more sufficiently analyze the performance of various algorithms, Figure14 represents the fusion results of eight groups of tested images for further comparison. It is fully illustrative that the proposed method displays excellent and acceptable fusion results compared to other reference methods, which reflects the proposed algorithm has excellent fusion performance and substantial stability. Meanwhile, the average objective metrics for the twelve groups of fusion results using different fusion strategies are listed in Table 5 and Figure15. The values marked in bold are the best values in all evaluation criteria. Similar to the objective values from Table 1 to Table 4, the proposed algorithm gains the best Piella value, secondary SF value and AG value, which declares that the proposed method can hold the correlation between fusion result and source images, and can maintain the brightness and contrast of the corresponding fusion result. Although the partial reference algorithms obtain better objective values, this phenomenon is mainly caused by the incongruity and irrationality in their fusion images, such as the visual perception of the fusion images in CNN, GFF, and RfnNet, etc. Therefore, it can be summarized that the proposed model has excellent fusion performance and

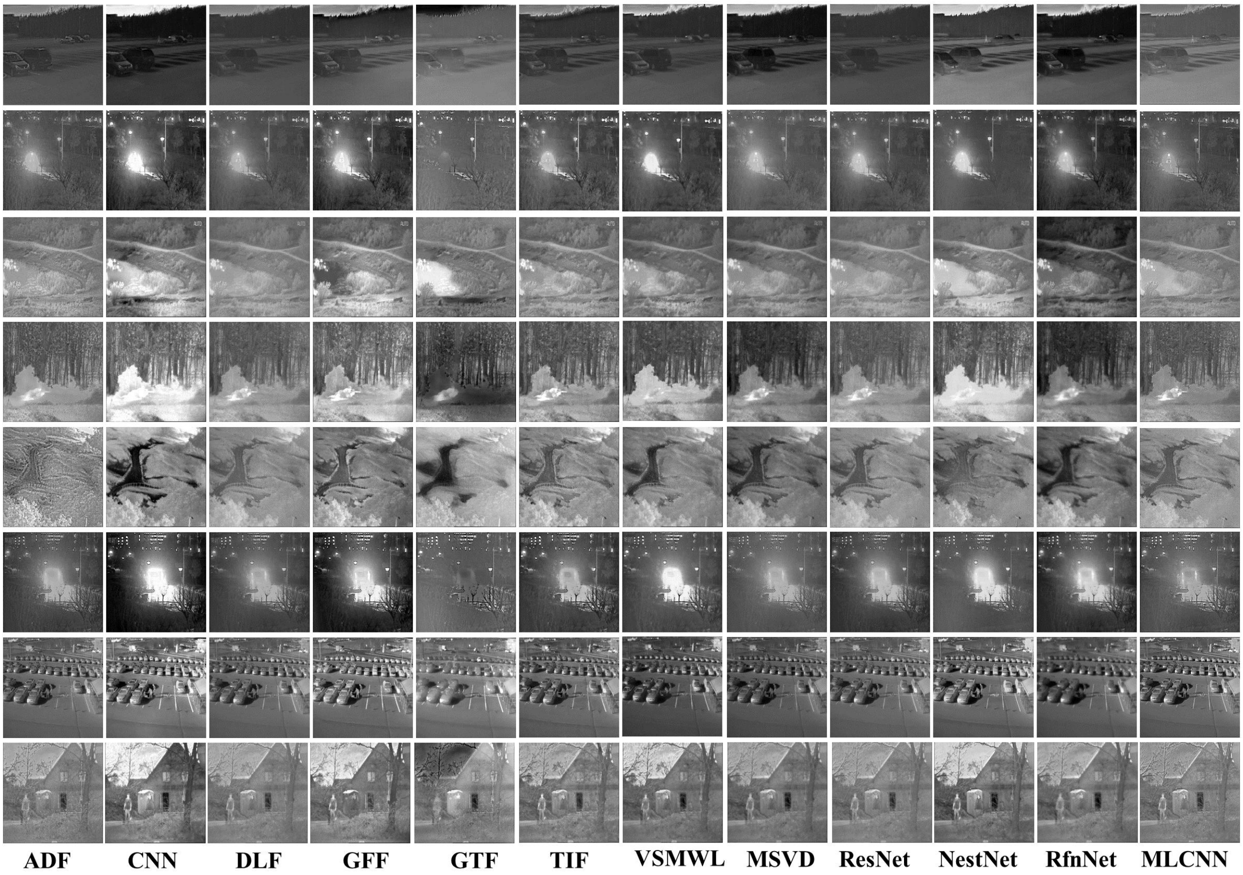


FIGURE 14. Fusion results of all tested images.

TABLE 6. Total running time (RT) based on testing images (unit: seconds).

Methods	RT
ADF	2.56
CNN	119.07
DLF	68.33
GFF	**2.49**
GTF	5.64
TIF	2.74
VSMWL	22.07
MSVD	2.962
ResNet	33.18
NestNet	3.77
RfnNet	5.36
MLCNN	3.53

TABLE 7. Number of network model parameters (NP) (unit: MB).

Methods	NP
CNN	1.54
DLF	510
ResNet	91.5
NestNet	10.4
RfnNet	18.2
MLCNN	103

comprehensive strong ability of information integration for the fusion problem of infrared and visible images on the whole.

D. ANALYSES OF COMPUTATIONAL COMPLEXITY

In addition to the visual analysis and objective evaluation metrics discussion, the running time (RT) is an important indicator for evaluating algorithm performance. the average running time of each algorithm is given in Table 6, the shorter the time the better the algorithm. Obviously, GFF, ADF, TIF, and MSVD obtain better running efficiency, which is mainly due to their multi-scale structures. However, they do not guarantee high-quality visual effects and fusion metrics. In another aspect, the proposed method acquires significantly shorter running times than other reference algorithms, especially algorithms based on neural networks like CNN and ResNet, etc. To sum up, the running time of the proposed method is acceptable and progressive.

To further explore the performance of the fusion algorithm, this paper compares the model parameters of the neural network based fusion algorithms with the values shown in Table 7. Clearly, it can observe that the proposed algorithm requires more parameters to be trained compared to reference algorithms. However, the complexity of the model parameters does not affect the efficiency of the algorithm when combined with RT, indicating that the proposed algorithm is efficient.

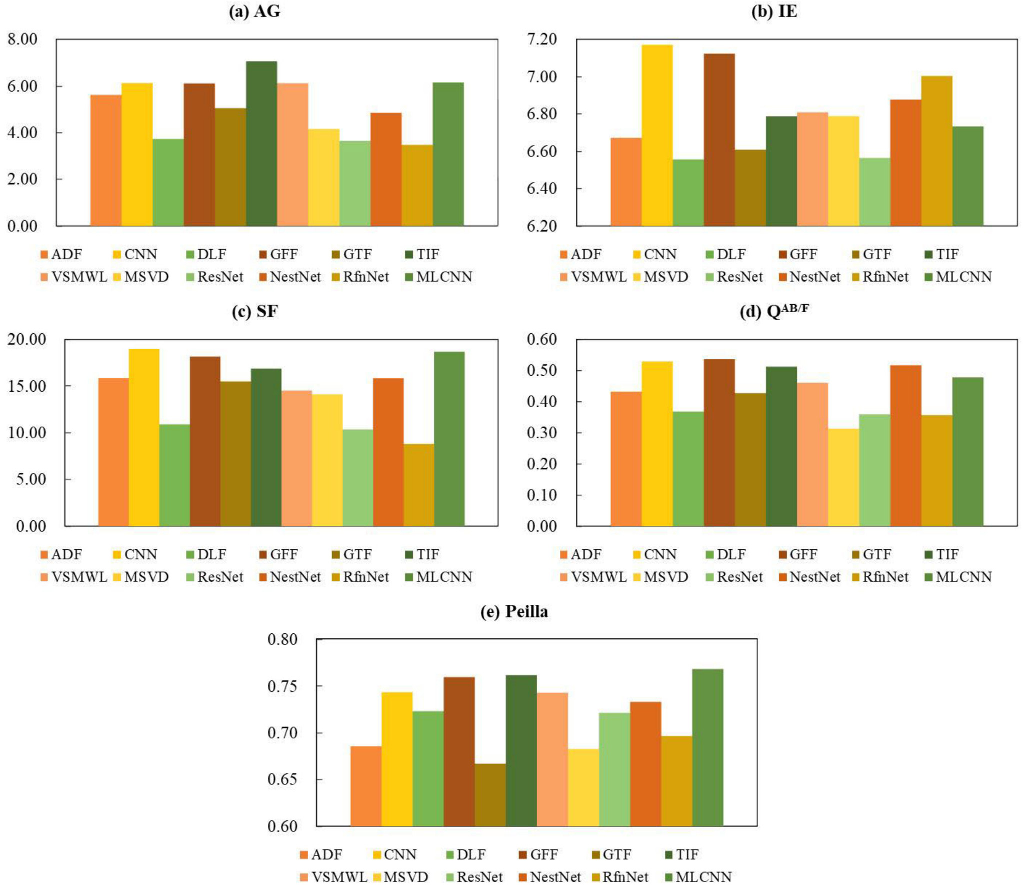


FIGURE 15. Quantitative comparison using mean value of each metric (a) IE, (b) IE, (c) SF, (d) $Q^{AB/F}$, and (e) Piella.

V. CONCLUSION

In this paper, a novel and efficient visible and infrared image fusion network model based on CNN is proposed. The model has three main advantages compared with current CNN based VIF methods: (1) A multi-scale convolutional fusion model with an improved residual block is proposed, which can explicitly integrate deep features without manual weight selection and fusion rule design. (2) In order to better train the proposed model, this paper uses the COCO dataset to reasonably generate a training dataset by means of high-resolution large-scale multi-focus images with ground-truth fusion images. It is significant to optimize the image fusion model in regression, and conduce to the loss function constrain the network focus on the informative regions of source images effectively to facilitate the retention of the useful information. (3) An effective image reconstruction structure combining affluent skip connections with multi-scale convolutional layers is designed to supplement the lost image details in the pooling process. The model is fully convolutional, so it can be trained in an end-to-end manner without a pre-processing process. It has been verified by numerous experiments that the proposed model owns progressive execution performance for infrared and visible image fusion problems compared with the current neural networks-based and popular multi-scale transformation-based methods.

REFERENCES

[1] J. Ji, Y. Zhang, Z. Lin, Y. Li, C. Wang, Y. Hu, F. Huang, and J. Yao, “End to end infrared and visible image fusion with texture details and contrast information,” *IEEE Access*, vol. 10, pp. 92410–92425, 2022.

[2] H. Adeel, M. M. Riaz, and S. S. Ali, “De-fencing and multi-focus fusion using Markov random field and image inpainting,” *IEEE Access*, vol. 10, pp. 35992–36005, 2022.

[3] D. Lei, M. Bai, L. Zhang, and W. Li, “Convolution neural network with edge structure loss for spatiotemporal remote sensing image fusion,” *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1015–1036, Feb. 2022.

[4] L. Ren, Z. Pan, J. Cao, J. Liao, and Y. Wang, “Infrared and visible image fusion based on weighted variance guided filter and image contrast enhancement,” *Infr. Phys. Technol.*, vol. 114, May 2021, Art. no. 103662.

[5] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He, “A survey of infrared and visual image fusion methods,” *Infr. Phys. Technol.*, vol. 85, pp. 478–501, Sep. 2017.

[6] Z. Liu, E. Blasch, and V. John, “Statistical comparison of image fusion algorithms: Recommendations,” *Inf. Fusion*, vol. 36, pp. 251–260, Jul. 2017.

[7] 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.

[8] D. Xu, Y. Wang, X. Zhang, N. Zhang, and S. Yu, “Infrared and visible image fusion using a deep unsupervised framework with perceptual loss,” *IEEE Access*, vol. 8, pp. 206445–206458, 2020.

[9] R. Chen, S. Liu, Z. Miao, and F. Li, “GFSNet: Generalization-friendly Siamese network for thermal infrared object tracking,” *Infr. Phys. Technol.*, vol. 123, Jun. 2022, Art. no. 104190.

[10] X. Liu, J. Li, X. Yang, and H. Huo, “Infrared and visible image fusion based on cross-modal extraction strategy,” *Infr. Phys. Technol.*, vol. 124, Aug. 2022, Art. no. 104205.

[11] Q. Pan, L. Zhao, S. Chen, and X. Li, "Fusion of low-quality visible and infrared images based on multi-level latent low-rank representation joint with Retinex enhancement and multi-visual weight information," *IEEE Access*, vol. 10, pp. 2140–2153, 2022.

[12] D. Zhu, W. Zhan, Y. Jiang, X. Xu, and R. Guo, "MIFFuse: A multi-level feature fusion network for infrared and visible images," *IEEE Access*, vol. 9, pp. 130778–130792, 2021.

[13] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Inf. Fusion*, vol. 35, pp. 81–101, May 2017.

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

[15] X. Wang, J. Yin, K. Zhang, S. Li, and J. Yan, "Infrared weak-small targets fusion based on latent low-rank representation and DWT," *IEEE Access*, vol. 7, pp. 112681–112692, 2019.

[16] Y. Yang, S. Tong, S. Huang, P. Lin, and Y. Fang, "A hybrid method for multi-focus image fusion based on fast discrete curvelet transform," *IEEE Access*, vol. 5, pp. 14898–14913, 2017.

[17] M. Asikuzzaman, H. Mareen, N. Moustafa, K. R. Choo, and M. R. Pickering, "Blind camcording-resistant video watermarking in the DTCWT and SVD domain," *IEEE Access*, vol. 10, pp. 15681–15698, 2022.

[18] P. Singh, M. Diwakar, V. Singh, S. Kadry, and J. Kim, "A new local structural similarity fusion-based thresholding method for homomorphic ultrasound image despeckling in NSCT domain," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 7, Jul. 2023, Art. no. 101607.

[19] 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.

[20] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15750–15761, 2017.

[21] R. Lai, Y. Li, J. Guan, and A. Xiong, "Multi-scale visual attention deep convolutional neural network for multi-focus image fusion," *IEEE Access*, vol. 7, pp. 114385–114399, 2019.

[22] T. Yao, Y. Luo, J. Hu, H. Xie, and Q. Hu, "Infrared image super-resolution via discriminative dictionary and deep residual network," *Infr. Phys. Technol.*, vol. 107, Jun. 2020, Art. no. 103314.

[23] Z. Qu, S.-Y. Wang, L. Liu, and D.-Y. Zhou, "Visual cross-image fusion using deep neural networks for image edge detection," *IEEE Access*, vol. 7, pp. 57604–57615, 2019.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Inf. Sci.*, vols. 433–434, pp. 125–141, Apr. 2018.

[32] K. Ren, D. Zhang, M. Wan, X. Miao, G. Gu, and Q. Chen, "An infrared and visible image fusion method based on improved DenseNet and mRMR-ZCA," *Infr. Phys. Technol.*, vol. 115, Jun. 2021, Art. no. 103707.

[33] S. Yi, J. Li, and X. Yuan, "DFPGAN: Dual fusion path generative adversarial network for infrared and visible image fusion," *Infr. Phys. Technol.*, vol. 119, Dec. 2021, Art. no. 103947.

[34] A. Fang, X. Zhao, J. Yang, B. Qin, and Y. Zhang, "A light-weight, efficient, and general cross-modal image fusion network," *Neurocomputing*, vol. 463, pp. 198–211, Nov. 2021.

[35] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "Ensemble of CNN for multi-focus image fusion," *Inf. Fusion*, vol. 51, pp. 201–214, Nov. 2019.

[36] Y. Liu, J. Sun, H. Yu, Y. Wang, and X. Zhou, "An improved grey wolf optimizer based on differential evolution and OTSU algorithm," *Appl. Sci.*, vol. 10, no. 18, p. 6343, Sep. 2020.

[37] L. Li, Z. Xia, H. Han, G. He, F. Roli, and X. Feng, "Infrared and visible image fusion using a shallow CNN and structural similarity constraint," *IET Image Process.*, vol. 14, no. 14, pp. 3562–3571, Dec. 2020.

[38] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[39] 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.

[40] H. Yan, X. Yu, Y. Zhang, S. Zhang, X. Zhao, and L. Zhang, "Single image depth estimation with normal guided scale invariant deep convolutional fields," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 80–92, Jan. 2019.

[41] 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.

[42] K. Wang, D. J. Dou, Q. Kemao, J. Di, and J. Zhao, "Y-Net: A one-to-two deep learning framework for digital holographic reconstruction," *Opt. Lett.*, vol. 44, pp. 4765–4768, Oct. 2019.

[43] 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.

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

[45] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform," *IEEE Sensors J.*, vol. 16, no. 1, pp. 203–209, Jan. 2016.

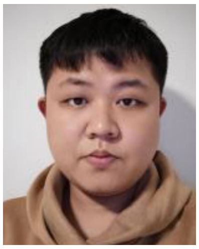
[46] V. P. S. Naidu, "Image fusion technique using multi-resolution singular value decomposition," *Defence Sci. J.*, vol. 61, no. 5, p. 479, Sep. 2011.

[47] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infr. Phys. Technol.*, vol. 76, pp. 52–64, May 2016.

[48] 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.

[49] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. Int. Conf. Image Process.*, vol. 3, Sep. 2003, p. 173.

ZHU PAN received the Ph.D. degree in optical engineering from Tianjin University, in 2017. He is mainly engaged in research in the fields of photoelectric detection, image acquisition, and processing. As the project leader, he has undertaken a national project and a provincial research projects. He has published more than seven SCI papers of *Infrared Physics and Technology*, *Measure*, *Optics & Laser Technology*, and *Optical Review*.



WANQI OUYANG received the bachelor's degree in optoelectronic information science and engineering from the Hubei University of Science and Technology, in 2019. He is currently pursuing the master's degree in instrument science and technology with the Wuhan University of Science and Technology. He has participated in the writing of several journals/conference papers. His research interests include image processing and deep learning algorithms.