electronics

Article

An Image Fusion Method Based on Special Residual Network and Efficient Channel Attention

Yang Li [1], Haitao Yang [2], Jinyu Wang [1], Changgong Zhang [1], Zhengjun Liu [3] and Hang Chen [1,4,*]

[1] School of Space Information, Space Engineering University, Beijing 101416, China
[2] Space Security Research Center, Space Engineering University, Beijing 101416, China
[3] School of Physics, Harbin Institute of Technology, Harbin 150001, China
[4] CNRS, CRAN UMR 7039, Universitéde Lorraine, 54000 Nancy, France
* Correspondence: hangchen@alu.hit.edu.cn

Abstract: This paper presents an image fusion network based on a special residual network and attention mechanism. Compared with the traditional fusion network, the image fusion network has the advantages of an end-to-end network and integrates the feature extraction advantages of the attention mechanism residual network. It overcomes the shortcomings of the traditional network that need complex design rules and manual operation. In this method, hierarchical feature fusion is used to achieve effective fusion. A combined loss function is designed to optimize training results and improve image fusion quality. This paper uses many qualitative and quantitative experimental analyses on different data sets. The results show that, compared with the comparison algorithm, the method in this paper has a stronger retention ability of infrared and visible light information and better indexes. 72% of eleven indexes compared with some images in the public TNO data set are optimal or sub-optimal, and 80% are optimal or suboptimal in the RoadScene data set, which is much higher than other algorithms. The overall fusion effect is more in line with human visual perception.

Keywords: codec network; deep learning; image fusion; attention mechanism

check for updates

Citation: Li, Y.; Yang, H.; Wang, J.; Zhang, C.; Liu, Z.; Chen, H. An Image Fusion Method Based on Special Residual Network and Efficient Channel Attention. *Electronics* **2022**, *11*, 3140. https://doi.org/10.3390/electronics11193140

Academic Editor: Gemma Piella

Received: 8 September 2022
Accepted: 23 September 2022
Published: 30 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Images taken directly can only present the information content of part of the band, considering the limitations of imaging equipment. With the image fusion of multi-band key information, a new image containing more information can be reconstructed. High-quality fused infrared and visible images can provide strong complementary information and have high-value applications in many fields, such as target detection, target tracking, automatic driving, and video surveillance. In recent years, image fusion algorithms have gradually developed into two categories: traditional methods and methods based on deep learning. The most typical case based on conventional methods is image fusion methods based on multi-scale transformation, including discrete wavelet transform [1,2], non-small downsampling shear wave transform [3], and so on. Representation-based learning methods are also widely used in this field, such as sparse representation [4] and joint sparse representation [5]. Subspace-based, saliency-based, and chaotic model-based methods are applied to image fusion. Nevertheless, traditional methods are not ideal due to problems such as complex design rules and weak generalization ability of scenes.

In the field of image fusion, the rapid development of deep learning is also reflected. The mainstream methods include the convolution network-based method, the codec network-based method, and so on. Deep learning includes mainly three elements: deep feature extraction, fusion strategy, and end-to-end training mode. Specifically, depth feature extraction usually consists of a convolution layer [6,7], a designed residual structure [8], a special feature extraction module [9], etc. However, the current network's ability to extract tiny features still needs to be improved. Most networks only transmit the final convolution results to the fusion module, which makes the information of the middle layer

lost. Fusion strategies are used to fuse the extracted feature information. Some networks use manual methods to design fusion strategies [10,11], while others use deep learning methods to design fusion strategies [12,13]. Unlike manually designed fusion strategies, deep learning-based strategies have stronger generalization ability in different scenarios because they can learn from the training data. However, there are still problems such as poor details preservation after fusion and training difficulty for learnable fusion structures.

This paper proposes a new end-to-end fusion framework to address the above problems. A fusion strategy based on convolutional layers is designed, which can follow the network for training. As a result, the depth extraction features can be more effectively utilized, reducing the training difficulty. In addition, in the feature extraction part of the network, a gradient operator feature extractor based on Laplace is designed to replace the work of many convolution layers. An efficient channel attention network (ECANet) is used as the last step of depth feature extraction to reduce the number of network layers and shorten the fusion time. Moreover, aiming at the problem of information loss in the middle layer of feature extraction, this paper adopts the method of feature extraction and feature reconstruction in the network so that different levels of information can be retained in the fused image.

In the aspect of loss function design, this paper designs a combined loss function, which includes strength loss, texture loss, and structure loss. It restricts the retention of pixel intensity features of infrared images, the retention of texture details of visible images by texture loss, and the retention of edge structure features of source images by structure loss.

Finally, the network is verified by comparative experiments on several public infrared and visible fusion data sets, and the corresponding subjective and objective analysis is made. Compared with the comparison algorithm, the fusion results obtained by the proposed method have a better visual effect, and the target definition and background details are more perfect.

2. Related Work

The application of deep learning methods in image fusion is mainly based on convolutional neural networks (CNN), which can be divided into end-to-end networks and non-end-to-end networks according to network types. This section mainly introduces the research results and related work in the field of image fusion, including traditional infrared and visible light fusion methods, coding and decoding models, typical deep learning image processing methods, etc.

2.1. Traditional Image Fusion Algorithms

Traditional image fusion algorithms refer to methods that usually perform spatial or frequency transformation by manually designing fusion methods and feature extraction methods, including the space-based method, wavelet transform method, sparse representation method, multi-scale transform method, and so on. The fusion algorithm consists of feature extraction, fusion, and image reconstruction. Among them, image reconstruction is usually the inverse process of feature extraction. Therefore, the key to the algorithm lies in the innovation and improvement of the feature extraction and fusion methods. There are many space-based methods, including but not limited to independent component analysis (ICA), principal component analysis (PCA), and non-negative matrix factorization (NMF). In terms of the basic principle of these methods, the high-dimensional source image information is first projected and transformed into the low-dimensional space or subspace. The source image information is then operated to eliminate the redundant information in the natural image so that the internal structure of the source image is enriched in the low-dimensional subspace. The method based on sparse representation is widely used in image processing [14–16]. With this method, the source image is sparsely represented by the linear combination of sparse bases in the dictionary to realize the stable representation ability of the image containing information [17]. Sparse representation has disadvantages as obvious

as advantages. The super-complete dictionary can endow this method with good fusion ability in specific fields. Still, the time and difficulty of its construction is a difficult point that cannot be ignored. The method based on the multi-scale decomposition framework is firstly based on the principle that the source image can be decomposed into components of different sizes. Secondly, the multi-scale images will be fused according to the designed fusion rules, and finally, the fused multi-scale images will be inverse transformed to get the fused images. The representative methods of multi-scale decomposition include discrete wavelet transform (DWT), and non-small down-sampled shear wave transform. Other methods applied to the fusion of infrared and visible images also provide a lot of research directions and ideas, including but not limited to salient features based on the principle of human visual perception [18], total variation (TV) [19], entropy [20], fuzzy theory [21] and so on.

2.2. Image Fusion Method Based on Deep Learning

Typical deep learning has attracted wide attention due to its application in image processing and has gradually shown its unique advantages over traditional methods [11,22]. The feature extraction network can obtain richer feature information than the hand-designed extraction algorithm, so the deep learning method occupies a place in the task of image fusion. In this paper, a method based on a convolutional neural network is proposed, which combines different focused images as the input of the CNN network, detects the common focus in the network, and outputs the fused image, thus realizing the effective fusion of multi-focused images. In addition, the CNN model is used to learn the effective focus measurement of image pixels, and the focus map is generated by comparing the high attention areas.

The self-codec method is a deep learning method based on the CNN method. This method usually realizes image feature extraction and feature recovery by training the self-encoder network, and the feature fusion in the middle is carried out by hand-designed rules [23]. In this paper, feature fusion is not carried out, but the infrared image features and visible light features are connected and introduced into the feature recovery layer. A learnable residual network is used for feature fusion, which achieves good results. This paper designs a codec with residual block and uses it to design a new fusion network called DenseFuse [23].

Squeeze-and-Excitation Networks (SENet) attention mechanism can also be used for image processing tasks. SENet introduces a channel attention mechanism to obtain the correlation between channels through global average pooling and full connection layer. In practice, the attention mechanism is widely used in image fusion because of its high efficiency and ease of use. In SESF, SENet is used to improve the encoder's feature extraction ability [22]. In this paper, the attention module is used to extract multi-scale features in the fusion framework and make the network focus on the discrimination area. In the NestFuse network framework [24], cascade spatial attention and channel attention are used as the fusion strategy to realize the fusion of infrared and visible light depth features.

3. Detailed Work

In this section, the construction and training methods of the fusion network in this paper will be introduced in detail. Section 3.1 introduces the main network framework of the method, Section 3.2 introduces two feature extraction modules, and Sections 3.3 and 3.4 introduce the network's training method, the setting of super parameters, and the composition of the loss function.

3.1. Converged Network Framework

The method proposed in this paper is an end-to-end network structure, and its composition is shown in Figure 1. This learning method touts the advantages of synergy and easier global optimal solutions than non-end-to-end networks. The network framework is

based on the classical codec network, including the encoder, fusion part, and decoder part. See Table 1 below for detailed composition.

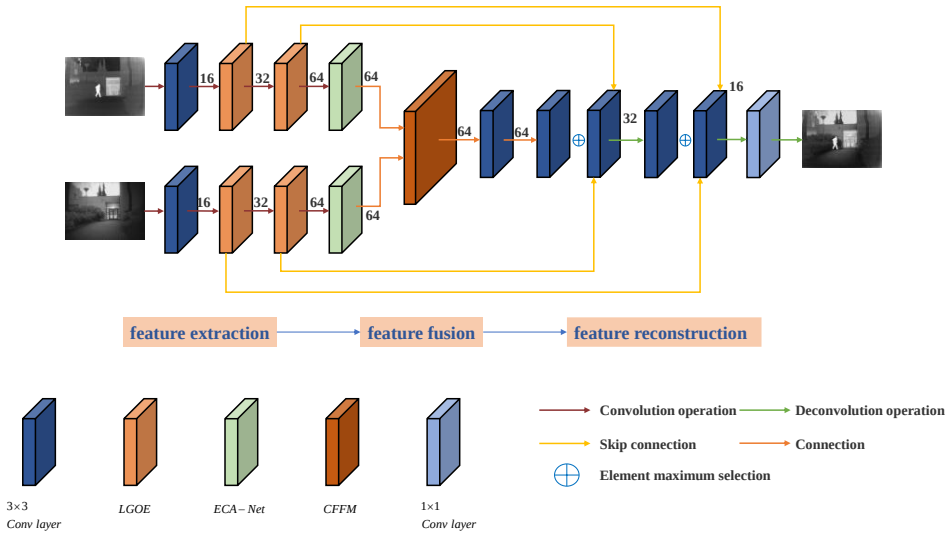


Figure 1. The overall framework of the proposed method.

Table 1. Detailed architecture diagram of the method proposed in this paper.

	Layer	Kernel Size	Stride	Input Channel	Output Channel	Activation
Feature extraction module	Conv ($3 * 3$)	3	1	1	16	Leaky Relu
	Conv (LGOE1)	-	-	16	32	-
	Conv (LGOE2)	-	-	32	64	-
	ECANet	1	1	64	64	Leaky Relu
Feature fusion module	RNF	-	-	64	64	-
Image reconstruction	Conv ($3 * 3$)	3	1	64	64	Leaky Relu
	Conv ($3 * 3$)	3	1	64	32	Leaky Relu
	Conv ($3 * 3$)	3	1	32	16	Leaky Relu
	Conv ($3 * 3$)	3	1	16	1	Leaky Relu

In this paper, the input infrared image and visible image are defined as I_i and I_v, respectively. In the feature extraction module, the visible light intermediate feature extracted by the Laplace gradient operator extractor (LGOE) layer is $t_{vi}^{L_k}$, the infrared feature is $t_{ir}^{L_k}$, and k represents the number of feature extraction layers. The fusion result of the output network is I_f, and all the source images of the input network have been registered.

In the encoder part, this paper uses the trained model to multi-level encode the input source image to extract different scale features, including a specific layer of $3 * 3$ convolution extraction for extracting rough features, two layers of LGOE, and one layer of the ECA attention module. The encoder network used in this paper has the following three advantages. Firstly, the convolution kernel size is $3 * 3$, and the step size is 1, which makes the size of the input image unlimited. Secondly, the use of LGOE can extract depth features and detailed information of images as much as possible to ensure that all salient features are transmitted backward. Thirdly, the channel attention (ECA) module will not increase many parameters and network complexity under improving network performance. The strategy of the above feature extraction module will be introduced in detail below.

In the decoder part, the convolution layer of three $3 * 3$ convolution kernels is selected as the decoding part, which is used to reconstruct the fused image. Such a simple but effec-

tive structure can express the extracted information well, keep the network low complexity, and avoid too low a fusion efficiency and long training time.

The features and advantages of the fusion network in this paper are introduced in general, and the feature extraction and fusion modules are introduced in detail below.

Laplace-based gradient operator extractor

3.2. Feature Extraction Module

The encoder is usually used for feature extraction in image processing tasks, and this basic feature extraction structure is widely used because of its portability and effectiveness. The encoder in this paper consists of a convolution layer, two LGOE layers, and an ECA attention layer.

3.2.1. Laplace-Based Gradient Operator Extractor

The basic convolution layer can realize the selection and backward transmission of image features, but some image details will be lost in the transmission process. To enhance the network's ability to extract fine-grained features from the source image, this paper designs a gradient operator feature extractor based on Laplace. The structure of LGOE is shown in Figure 2. Each LGOE contains three $3 * 3$ convolution layers, with the activation function of Leaky Rectified Linear Unit (LReLU). The dense connection is used in the three-layer convolution to improve the feature extraction ability.

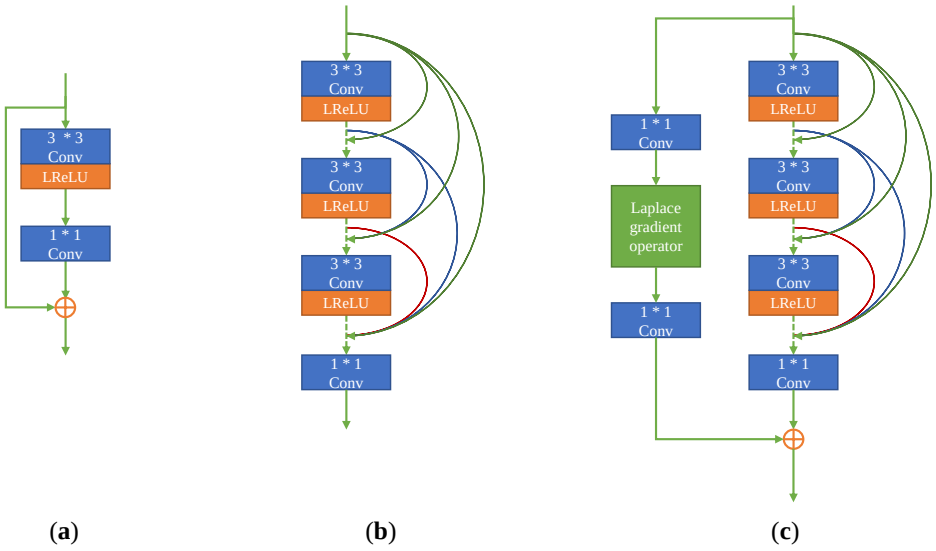


Figure 2. A schematic of three deep learning architectures, (**a**) Residual block; (**b**) Dense block; (**c**) Dense residuals with Laplacian gradient operators.

After dense connection, a regular convolution layer with convolution kernel $1 * 1$ is added to eliminate the difference in the channel number. In the residual part, a special convolution layer is designed according to the Laplacian gradient operator to extract the fine granularity features of the image. Finally, the extracted depth and fine granularity features are fused and output. LGOE is placed in this paper's second and third layers of the feature extraction layer.

Aiming at the advanced semantic features in infrared and visible light, this paper proposes a gradient operator feature extractor based on Laplace. Let the image of the input LGOE be F^i, and its output F^{i+1} can be expressed as follows:

$$F^{i+1} = GRDB\left(F^i\right) = \mathrm{Conv}^n\left(F^i\right) \oplus \mathrm{Conv}\left(\nabla F^i\right) \tag{1}$$

where $\mathrm{Conv}(-)$ represents convolution layer, $\mathrm{Conv}^n(-)$ represents N cascaded convolution layers, and $\oplus$ represents the summation of elements in tensor. ∇ is a gradient operator,

a special convolution operation, and the convolution kernel is the Laplacian convolution kernel. The network can extract fine-grained features by convolution operation between the input features and the high-frequency convolution kernel.

3.2.2. ECAnet

ECAnet [25] is a new lightweight channel attention mechanism type, as shown in Figure 3. This structure uses a local cross-channel interaction strategy without dimensionality reduction and a method of adaptively selecting the size of the one-dimensional convolution kernel. The purpose is to ensure the model's high computational performance and low complexity and achieve noticeable performance improvement with a few additional parameters.

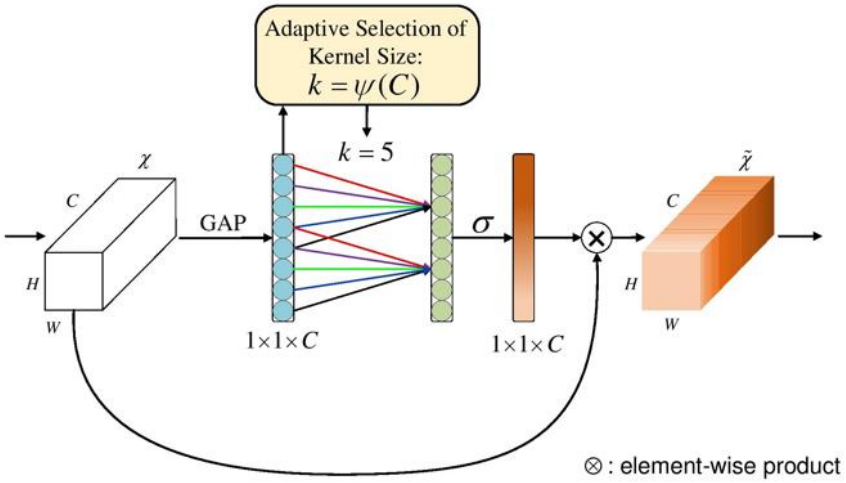


Figure 3. The network structure of ECANet.

ECANet is improved based on SENet [23]. This network is a new image model proposed in 2017 that can improve network performance by modeling the correlation between feature channels, called the channel attention mechanism. SENet consists of compression, excitation, and feature recalibration, and the network mechanism is shown in Figure 4. Compared with SENet, ECANet uses fewer parameters and avoids the negative impact of dimensionality reduction. The structure of ECANet is shown in Figure 3. The main difference between ECANet and SENet is that the former omits the step of dimensionality reduction after GAP. Instead, it uses the channel and its adjacent K channels to obtain cross-channel communication information.

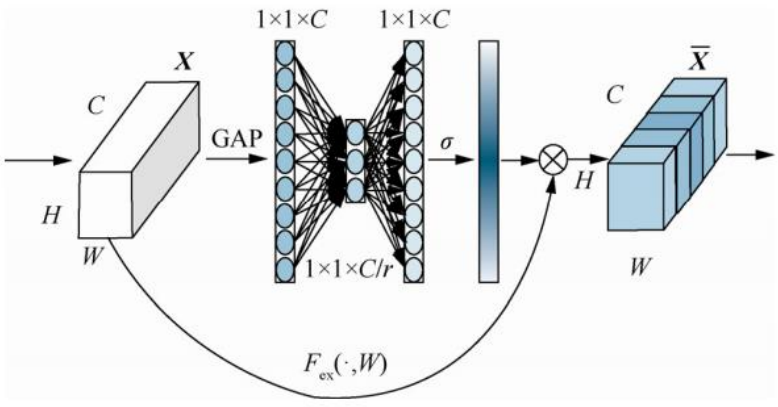


Figure 4. The network structure of SENet, where ⊗ is element-wise product.

Let the dimension of the output of a convolutional layer layer be $\mathcal{X} \in \mathbb{R}^{W \times H \times C}$, where W, H, and C represent the width, height of tensor, and the number of convolution kernels, respectively. The weight of each channel in the SENet network is calculated as follows:

$$\boldsymbol{\omega} = \sigma\Big(f_{\{\mathbf{W}_1,\mathbf{W}_2\}}(g(\mathcal{X}))\Big) \tag{2}$$

$$g(\mathcal{X}) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H} \mathcal{X}_{ij} \tag{3}$$

where $g(\mathcal{X})$ is global average pooling (GAP), σ is Sigmoid function. $\mathbf{y} = g(\mathcal{X}), f_{\{\mathbf{W}_1,\mathbf{W}_2\}}$ has the following relationship with:

$$f_{\{\mathbf{W}_1,\mathbf{W}_2\}}(\mathbf{y}) = \mathbf{W}_2\mathrm{ReLU}(\mathbf{W}_1\mathbf{y}) \tag{4}$$

ReLU stands for the rectified linear unit $\mathbf{W}_1 = C \times (\frac{C}{r}), \mathbf{W}_2 = (\frac{C}{r}) \times C$, but this processing method destroys the direct relationship between channel and weight.

In ECANet, dimension reduction is not required for $\mathbf{y} \in \mathbb{R}^C$ after GAP, so

$$\boldsymbol{\omega} = \sigma(\mathbf{W}\mathbf{y}) \tag{5}$$

where $\mathbf{W}$ is the parameter matrix of $C \times C$. In ECANet, $\mathrm{W_k}$ it is used to indicate the learned channel attention:

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w \end{bmatrix} \tag{6}$$

$\mathrm{W_k}$ includes $k \times C$ parameters, and only y_i and its k adjacent information interactions are considered for the weight y_i:

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w_i^j y_i^j\right), y_i^j \in \Omega_i^k \tag{7}$$

where Ω_i^k represents the set of k adjacent channels o y_i^j. Afterward, all channel weight information is shared to improve performance:

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w^j y_i^j\right), y_i^j \in \Omega_i^k \tag{8}$$

To sum up, this method realizes the information interaction between channels through one-dimensional convolution with convolution kernel size k:

$$\boldsymbol{\omega} = \sigma\Big(\mathrm{C1D}^k(\mathbf{y})\Big) \tag{9}$$

where CID represents one-dimensional convolution, which contains k parameter information. This ability of cross-channel information interaction ensures the performance and efficiency of the network. The value of k is determined in the subsequent experiment.

3.3. Fusion Structure

In the fusion network based on deep learning, most authors choose simple manual strategies for deep feature fusion [6,26,27], such as elementwise-add, elementwise-mean, elementwise-maximum, and elementwise-sum. Moreover, this fusion method only focuses on deep convolution features, ignoring the importance of shallow features in the fusion

network. To solve this problem, this paper uses a network structure based on a convolution layer to fuse deep features and uses the rule for the highest pixel value to fuse shallow features.

The Central fusion network (CFN) used in this paper is a variant of a simple convolution module used in the fusion of deep image features. Its structure is shown in Figure 5. In this network, "Conv1", "Conv2," and "Conv6" are the first layer of fusion features. "Conv1" and "Conv2" are connected in series as the input of "Conv3", and "Conv6" is input as the residual.

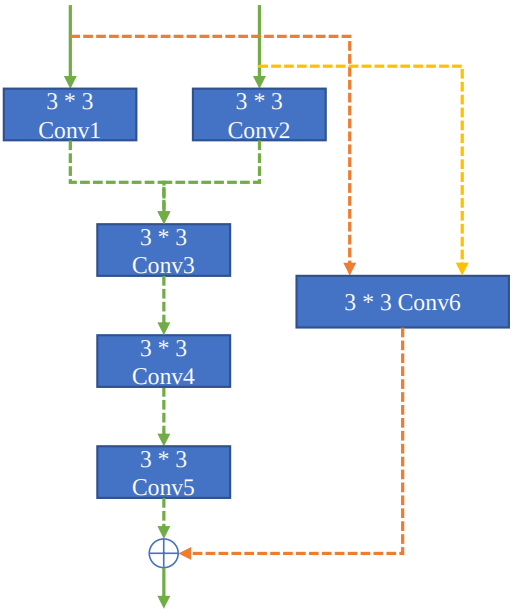


Figure 5. The network structure of CFN.

The structure of CFN is simple and light, so it will not produce too many parameters, and the training difficulty is low. It can be seen from the comparative experiments that CFN has a better effect than other methods commonly used in fusion.

For the fusion of shallow features, the convolution results of the previous part of the encoder are fused and transmitted to the decoder to realize the preservation of shallow features in the fused image. Since each pixel value of the convolved feature map can represent a certain area of the original image and has a high information value, this paper selects the maximum pixel value.

3.4. Loss Function

In the network training stage, we need to train the feature extraction ability of the coding layer and the feature reconstruction ability of the decoding layer at the same time.

Ground truth plays a crucial role in traditional CNN network training, and it is usually used to test the accuracy of the training set's classification of supervised learning techniques. However, the result of the image fusion network in this paper is a fusion image that includes both infrared target information and visible light details, so there is no Ground truth. This paper designs a combined loss function for such a special case to guide the network's training. A loss function L is generated by an appropriate combination strategy to minimize the loss of network training.

The loss function in this paper consists of three parts: strength loss, texture loss, and structure loss. The total loss is as follows:

$$L = L_{\text{int}} + \alpha L_{\text{tex}} + \lambda L_{\text{ssim}} \tag{10}$$

where L_{int} represents strength loss, which restricts the pixel performance of the fusion result, mainly aiming at the target features of infrared images. L_{tex} indicates texture loss, which restricts the embodiment of the detailed information of the fusion result, mainly aiming at the fine-grained background information of visible light images. α and λ is the weight used to balance the losses. The value of α is set to 10 according to the research and λ is set to 0.5 after the research in this paper. For the specific analysis experiment, see the parameter verification section in the chapter on experimental verification. The formula of L_{int} is as follows:

$$L_{\text{int}} = \frac{1}{HW}\left\| I_f - \max(I_{ir}, I_{vi}) \right\|_1 \tag{11}$$

where HW represent the height and width of the image, respectively, $\|\bullet\|_1$ represents the L1 norm, and $\max(I_{ir}, I_{vi})$ means that the maximum value is selected pixel by pixel. The formula of L_{tex} is as follows:

$$L_{\text{tex}} = \frac{1}{HW}\left\| \left|\nabla I_f\right| - \max(|\nabla I_{ir}|, |\nabla I_{vi}|) \right\|_1 \tag{12}$$

SSIM [28] is a typical image evaluation algorithm representing the structural similarity between the standard image and the image to be evaluated. In this paper, SSIM is used to constrain the overall structure of the fused image. If the marked image is O, the image to be evaluated is I, the input infrared image is I_{ir}, and the visible input image is I_{vi}, then the formula of L_{ssim} is as follows:

$$\text{SSIM}(O, I) = \frac{(2\mu_O\mu_I + C_1)(\sigma_{OI} + C_2)}{(\mu_O^2 + \mu_I^2 + C_1)(\mu_O^2 + \mu_I^2 + C_2)} \tag{13}$$

$$L_{\text{ssim}} = \frac{\varepsilon \cdot \left(1 - SSIM\left(I_f, I_{ir}\right)\right) + (1 - \varepsilon) \cdot \left(1 - SSIM\left(I_f, I_{vi}\right)\right)}{H \cdot W} \tag{14}$$

where $\mu_O\mu_I$ indicates the average value of pixels of image O and image I, σ_{OI} represents the covariance between image O and image I, C_1C_2 are constant, set to 0.2 based on a previous experiment [29]. HW are the height and width of the image, respectively, and ε is used to balance the weights of the respective structural similarities between the infrared and visible images.

4. Experimental Verification

4.1. Experimental Environment and Data Preparation

The experimental environment of this paper is as follows: Intel (R) Core (TM) i9-9820X processor, NVIDIA GeForce RTX 3090 GPU, win11 operating system, software environment pytorch 3.6 (Pytorch.org, Warsaw, Poland), network training software platform pycharm, and test algorithm platform matlab2021b (The MathWorks, Inc., Natick, MA, USA).

For the selection of data sets, this paper chooses the RoadScene data set as the training set and test set of this network model, and the TNO data set and some RoadScene data sets as the verification set. One thousand pairs of infrared and visible images in the data set are selected for network training, and 360 pairs of infrared and visible images in the data set are selected as test sets. All the images in the data set have been registered so that this paper can import the data directly into the network. In this paper, the learning rate is set to decline with the number of rounds during training, the initial learning rate is 0.001, and the learning rate decreases by 25% for each round of training. Adam solver is used to optimizing the loss function, and the size of each round is set to 8, and a total of 30 rounds are trained.

For the selection of evaluation indicators, this paper selects a group of commonly used performance evaluation indicators of Fusion I. The index system is mainly divided into four categories, which are evaluation indicators based on information theory, including EN (entropy) [30], MI (mutual information) [31], and PSNR (peak signal to noise ratio) [32]. MSE (mean-square error) based on structural similarity, SF (spatial frequency) [33], SD (standard

deviation) [34], AG (Average Gradient) [35] based on image features, Evaluation index VIF (Visual fidelity) [36] based on human visual perception, and evaluation index based on source image and generated image, mainly including CC (Correlation Coefficient) [37], SCD (Difference Correlation) [38] and Qabf (Gradient-based Fusion Performance) [39]. To ensure the accuracy and consistency of objective evaluation, all comparative experiments are used in this paper.

4.2. Comparative Experiment

4.2.1. Comparison of Loss Function Parameters

There are two self-defined parameters α, λ in the loss function, and the parameters in the loss function control the proportion of different parts of the combined loss function. In this paper, the parameter α is set to 10, according to the paper [40]. To evaluate the optimal value of λ, the value of λ (of this paper is verified by experiments with control variables. Other factors being the same, assign λ values to 0, 0.1, 0.5, 1, 10, and 100, and conduct network training and testing, respectively. The test uses image pairs in the TNO data set, as shown in Figure 6.



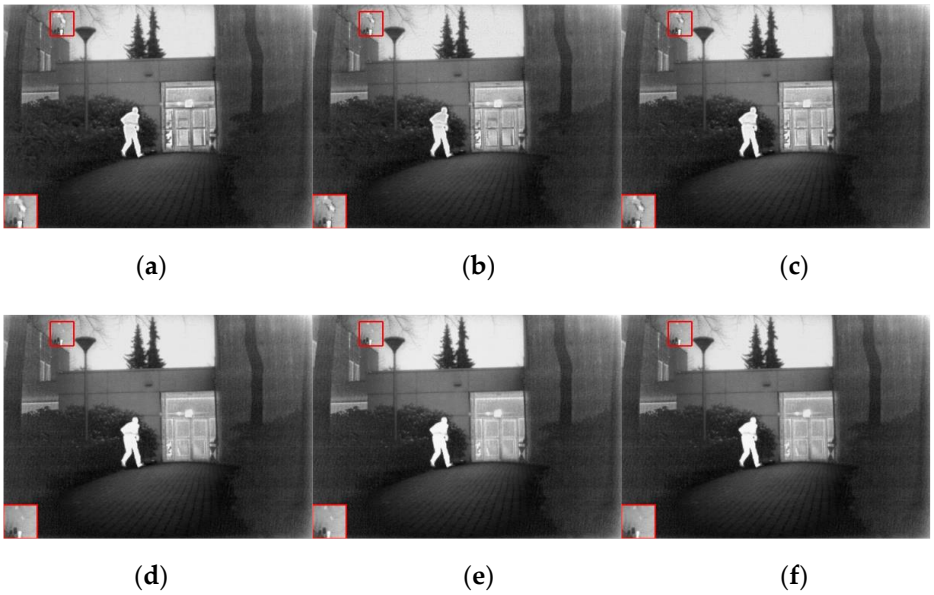
(**a**) (**b**)

Figure 6. The loss function compares the experimental source data, (**a**) Visible image, (**b**) Infrared image.

Figure 7 shows the output results of images with different λ values. Intuitively, as the value increases, the result's overall brightness is gradually improved. When $\lambda = 0.1$, it can be seen from the "glass door" part of the picture that visible light information is reflected more, while the "glass door" part of the picture only reflects infrared features when $\lambda = 1000$. It can be preliminarily considered that with the increase of the λ value, the fusion results will gradually eliminate the visible light features and retain more infrared features. When $\lambda \geq 10$, the infrared features in the red box disappeared, the visible light information of the upper body of the infrared target figure in the picture was erased, so the fusion effect was poor. Table 2 shows the comparison of objective indicators of experimental results, Objectively speaking, the optimal values of evaluation indexes are scattered, but when $\lambda = 0.5$, more indexes reach the highest or second highest point. Based on intuitive and objective comparisons, this paper sets $\lambda = 0.5$ as the value of the loss function.

Table 2. The following is the fusion result of different loss function parameters, red is optimal, and blue is suboptimal.

LGOE	EN	SF	SD	PSNR	MSE	MI	VIF	AG	CC	SCD	Qabf
0.1	7.3250	0.0414	8.8758	62.1439	0.0397	3.9588	1.0213	4.1167	0.5500	1.7407	0.5516
0.5	7.3841	0.0415	8.9677	62.2285	0.0398	4.1862	1.0439	4.2497	0.5517	1.7396	0.5851
1	7.3785	0.0416	8.8833	62.0650	0.0404	4.2851	1.0557	4.1819	0.5550	1.7465	0.5704
10	7.3515	0.0394	8.8256	61.9424	0.0416	5.7972	1.1401	4.0409	0.5493	1.7159	0.5790
100	7.3701	0.0405	8.8921	61.7965	0.0430	5.7946	1.1516	4.1248	0.5499	1.7152	0.5738
1000	7.3588	0.0395	8.8863	61.9661	0.0413	5.8156	1.1421	4.0274	0.5489	1.7116	0.5740



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 7. Experimental results of different loss function parameters, (**a**) $\lambda = 0.1$; (**b**) $\lambda = 0.5$; (**c**) $\lambda = 1$; (**d**) $\lambda = 10$; (**e**) $\lambda = 100$; (**f**) $\lambda = 1000$.

4.2.2. Comparison of Convergence Strategies

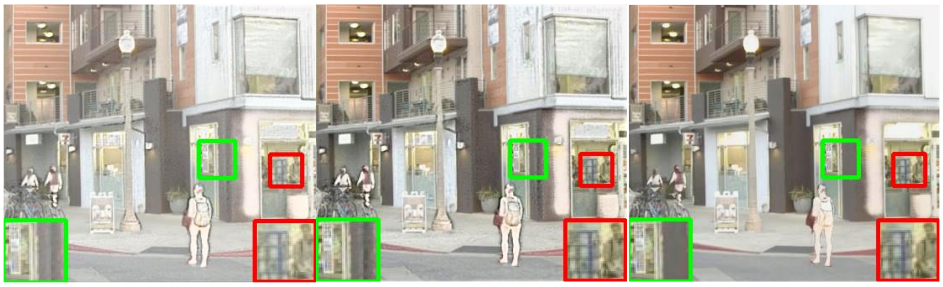
This paper designs a fusion module based on a convolution layer. To verify the superiority of this module over other fusion modules, several commonly used fusion rules are selected, including four methods: maximum selection, tensor dimension superposition, tensor value superposition, weighted average, and a comparative experiment is carried out. In the experiment, an image in the RoadScene dataset is selected as the comparison object, as shown in Figure 8.



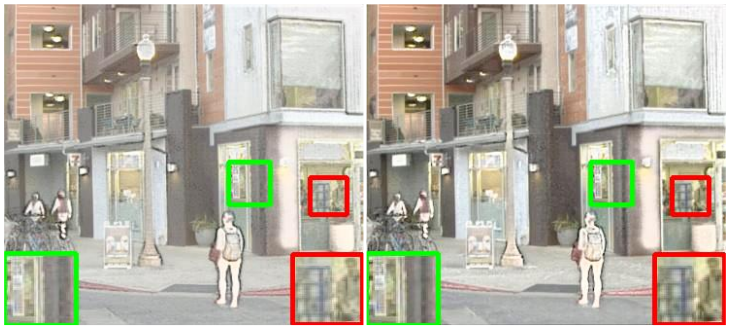
(**a**) (**b**)

Figure 8. The fusion strategy compares the source image of the experiment, (**a**) Visible image, (**b**) Infrared image.

Figure 9 shows the fusion results of replaci ng different fusion modules while keeping other parts consistent. The contrast of the maximum selection method is lower than that of this method, and the overall fusion result has higher brightness and lower contrast. Compared with the method proposed in this paper, numerical addition and tensor dimension superposition erase more infrared details, and the method of tensor dimension superposition looks smoother in the image. The black strip features of the wall with the green box are entirely lost in the results of this method. The results of the weighted average method have some disadvantages, such as low contrast and unclear infrared features.



(**a**) Max (**b**) Add (**c**) Tensor superposition



(**d**) Average (**e**) CFN

Figure 9. Experimental results of different fusion modules, (**a**) Max, (**b**) Add; (**c**) Tensor superposition; (**d**) Average; (**e**) CFN.

From the comparison of indexes, several methods can get the best among different indexes, but the method in this paper can get more optimal values, as shown in Table 3. Combining subjective and objective results, the CFN fusion module proposed in this paper is superior to the contrast method, the fusion results have high contrast, and the infrared and visible features are well reflected.

Table 3. The following are the fusion results of different fusion modules. Red marks are optimal and blue marks are suboptimal.

Fusion Modules	EN	SF	SD	PSNR	MSE	MI	VIF	AG	CC	SCD	Qabf
Max	6.5023	0.0326	8.1126	67.1221	0.0126	4.5716	1.1937	2.8334	0.6514	1.8795	0.6084
Tensor superposition	6.4707	0.0317	8.0400	67.2473	0.0123	5.0429	1.1723	2.6576	0.6511	1.8832	0.5888
Add	6.4867	0.0324	8.0433	67.2571	0.0122	4.8275	1.1853	2.7842	0.6493	1.8711	0.5118
Average	6.4839	0.0326	8.1292	67.2960	0.0121	4.6653	1.1945	2.8082	0.6486	1.8666	0.5949
CFN	6.5078	0.0331	8.0590	67.2213	0.0124	4.6765	1.2026	2.8490	0.6489	1.8811	0.6160

4.2.3. Comparison of LGOE Cascade

This paper uses a two-layer LGOE feature extractor in the encoder for deep feature extraction. To verify whether the number of cascaded layers is optimal, a comparative verification experiment of the LGOE cascade is designed. Under the same other factors, the number of cascaded layers is set to 1, 2, 3, and 4. In the experiment of the 4-level cascade, because the network depth is too deep, it can't converge at the initial stage of training, and the gradient disappears, so it is possible to exclude 4-level and above as the optimal one. Select an image in the experimental RoadScene data set as the contrast object, as shown in Figure 10.

Figure 11 shows the fusion results of a different number of feature extractors. Intuitively, with the number of cascades increasing, the brightness of the fusion results gradually decreases. That is, the proportion of infrared images in the fusion results gradually increases; The brightness of the picture is high, but the contrast is low; The brightness of the

image is dark, and the infrared target features are not prominent; The overall effect of the picture is good, the clarity and brightness are balanced, the infrared target is prominent, and the visible light details are clear.

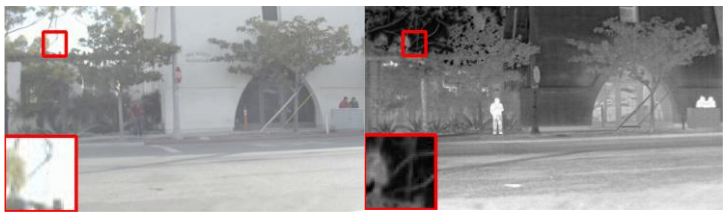


Figure 10. Source image of LGOE cascade comparison experiment.



(**a**) (**b**) (**c**)

Figure 11. Experimental results of a different number of couplets, (**a**) 1, (**b**) 2; (**c**) 3.

From the comparison of objective indicators in Table 4, when the number of cascades is 2, the evaluation indicators of the fusion effect have obtained the optimal values of most indicators. Compared with other cascades, the fusion results have better performance in objective indicators and intuitive performance. Therefore, two LGOE feature extractors are selected in this paper.

Table 4. The following is the fusion result of the number of cascades. Red is optimal.

LGOE	EN	SF	SD	PSNR	MSE	MI	VIF	AG	CC	SCD	Qabf
1	6.951	0.049	9.213	63.041	0.045	4.371	0.955	4.535	0.576	1.690	6.951
2	6.933	0.050	9.178	63.171	0.043	3.816	1.012	4.774	0.581	1.704	6.963
3	6.550	0.048	8.203	63.205	0.038	2.529	0.658	4.398	0.540	1.352	6.550

4.3. Fusion Effect Comparison

To prove the effectiveness and superiority of the network architecture proposed in this paper, six fusion algorithms are selected for comparative experiments, namely, Cross Bilateral Filter (CBF) [41], convolutional neural network (CNN) [42], Proportional Maintenance of Gradient and Intensity (PMGI) [43], FusionGAN [12], IFCNN [29], U2Fuison [44].

CBF is an image fusion algorithm based on cross bilateral filter, which is a classical algorithm of traditional image fusion; PMGI is a general image fusion algorithm based on gradient and intensity ratio, which is the same end-to-end deep learning algorithm as this paper, but the main extracted information is pixel gradient and intensity; CNN method uses the same method of convolution to extract features as this paper, but mainly for multi-focused FusionGAN is a fusion algorithm based on GAN algorithm, the basic principle is to generate the image with infrared features and multi-layer visible gradient by generator, and force the fused image to have more visible texture by discriminator, which is the same end-to-end fusion method as this paper, but is a derivative of GAN algorithm; IFCNN is a convolutional neural network based IFCNN is a general image fusion algorithm based on convolutional neural network, which is a more basic fusion method, only extracting the features of the source image through two convolutional layers and passing them into the fusion layer (including three optional methods of element maximum, element minimum or element average) for fusion, which extracts fewer features compared with this paper;

U2Fuison is an unsupervised image fusion algorithm, using pre-trained VGG-16 as the feature extractor, and the other parts are based on DenceNet.

To prove the generalization ability of the algorithm, this paper verifies it on the following two data sets.

4.3.1. Comparison of Effects in TNO Data Set

To avoid the chance of data, 36 images in the TNO data set are selected as the source image data of the effect comparison experiment, as shown in the figure. At the same time, five images were selected from the used data for display to visually compare the fusion results of different algorithms without occupying too much space. Some details or main objects in the images were marked and enlarged with green highlighted boxes, and the visible and infrared images as input were placed in the first and second rows. The following are the fusion results obtained by different methods, as shown in Figure 12.

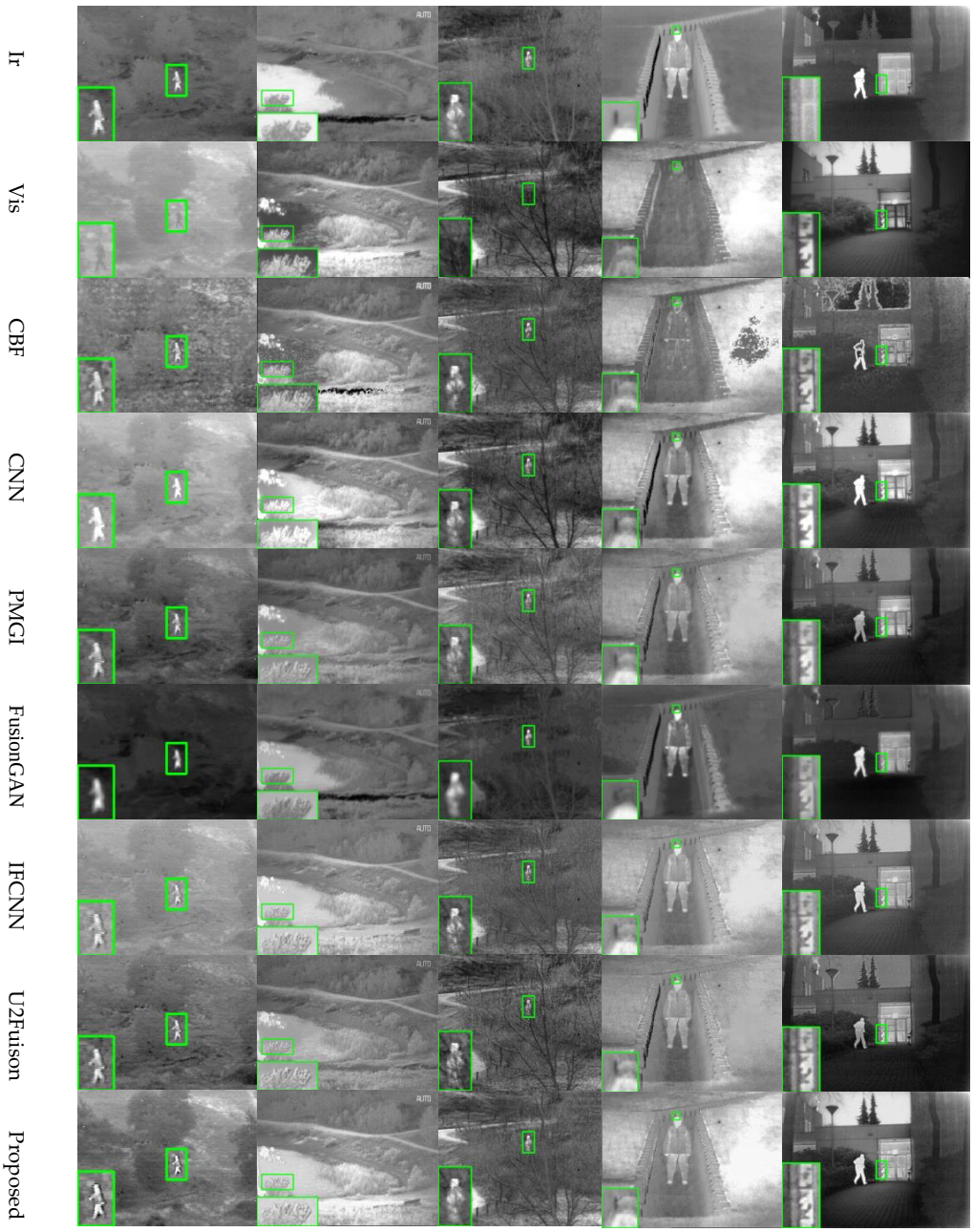


Figure 12. Experimental results of different fusion methods.

From an intuitive qualitative point of view, there are obvious gaps in the fusion results of the seven methods. It can be seen that the fusion results of different methods have different emphases. The fusion result of CBF has a poor visual effect and contains a lot of noises that do not exist (for example, the land in the first image, the noise on the right side in the fourth image, and the target outline in the fifth image), and the detailed information is seriously distorted. The results are similar to those of CNN IFCNN, and the overall visual effect is good. However, the contrast of environmental information in the first image is too low, and the fusion effect of the algorithm on low-quality images is poor. The brightness of grass in the enlarged part of the second image is too high, but there is no infrared feature in the source image. The trench's low brightness feature on the left side in the fourth image completely covers the visible light information.

PMGI U2Fuison has good fusion ability for low-quality source images. The first image has high contrast and full environmental information. In the fifth image, the outline in the glass is unclear, the contrast is low, the overall fusion effect is general, and the sky is gray, so the fusion effect is general. And the visible light information of PMGI in the third image is saved less, which makes fused images less sharp. The overall effect of FusionGAN is poor, and the details are insufficient. For example, the infrared target in the first picture and the infrared target in the fifth picture is not clear in outline, the fusion result is gray, and the contrast is insufficient. Compared with the shortcomings of other methods mentioned above, this paper can achieve a better balance in different quality images, keep clear visible light information in all images, and keep the infrared target bright. The first image has a good fusion contrast, keeping high contrast information, and the second image keeps different brightness for different places, which has a better visual effect based on inheriting the source image information.

The objective evaluation results of indicators are shown in Table 5, with the best results in red font and the second-best in blue font. From the results, the algorithm proposed in this paper has better performance in this data set, and the fusion results can obtain the optimal and suboptimal values of most indexes.

Table 5. The objective experimental results of different fusion methods show that the red font is optimal, and the blue font is suboptimal.

Method	EN	SF	SD	PSNR	MSE	MI	VIF	AG	CC	SCD	Qabf
CBF	6.914	0.054	9.098	62.799	0.045	2.230	0.541	5.487	0.386	1.367	0.406
CNN	7.161	0.043	9.610	62.169	0.052	2.425	0.951	4.165	0.454	1.695	0.538
PMGI	6.919	0.031	9.297	62.523	0.045	2.010	0.829	3.218	0.522	1.785	0.356
FusionGAN	6.583	0.024	8.751	60.921	0.060	2.337	0.659	2.327	0.428	1.381	0.232
IFCNN	6.732	0.045	9.006	63.175	0.040	2.398	0.810	4.428	0.487	1.690	0.505
U2Fuison	6.950	0.044	9.550	62.811	0.042	1.876	0.810	4.710	0.515	1.781	0.423
Proposed	7.090	0.047	9.516	63.197	0.057	3.645	0.973	4.606	0.450	1.625	0.549

Specifically, the fusion results of this algorithm have relatively good information entropy, spatial frequency, standard deviation, mean square error, good gradient characteristics in peak signal-to-noise ratio, mutual information, and visual fidelity. Experimental results show that the proposed algorithm has better fidelity, lower distortion, and artifacts in the TNO data set than the comparison algorithm and has a better fusion effect in general.

4.3.2. Comparison of Effects in RoadScene Data Set

This section tests six comparison algorithms and our algorithm on the RoadScene dataset. The test set is 50 pairs of visible and infrared images selected from the RoadScene data set. Five typical scenes are selected from the results shown in Figure 13, with local details selected by red boxes and enlarged.

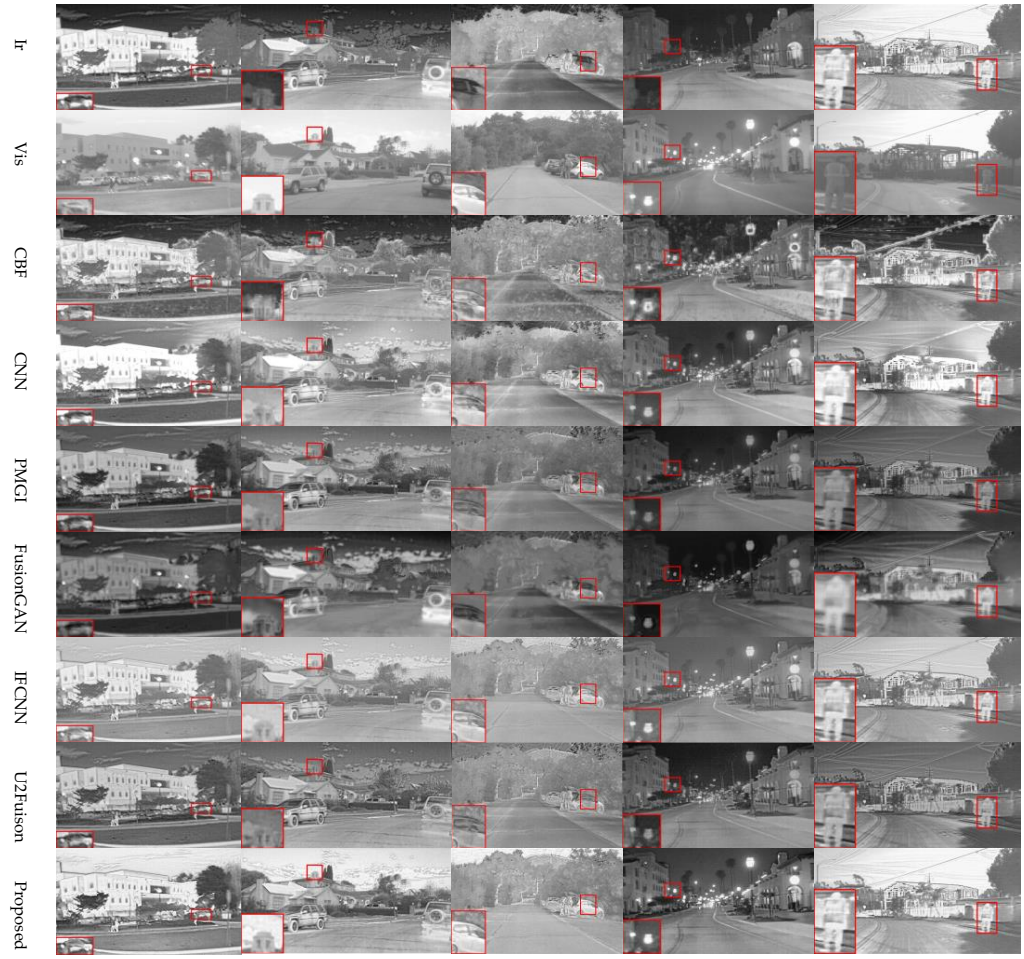


Figure 13. Experimental results of different fusion methods.

In the figure, the infrared image mainly includes the highlighted thermal radiation infrared information, including pedestrians, vehicles, some buildings, etc. The visible image contains a lot of detailed environmental information. From an intuitive point of view, this method is more prominent in infrared target information retention, and the brightness of people's vehicles is significantly higher than that of the environment, which is convenient for human eye recognition and subsequent programmed processing. In terms of visible light detail information retention, the overall result details and textures are more clear, and the brightness intensity distribution is appropriate.

Objective Contrast In this paper, 11 evaluation indexes, such as EN, SF, SD, etc., used above are used. After fusing 50 images in the test set, each image is evaluated separately, and the average value is obtained. The best image is marked with red and the second-best is marked with blue. The results are shown in Table 6. In this paper, the algorithm achieves the best value in spatial frequency, standard deviation, mean square error, average gradient, and difference correlation index, and the second-best value in mutual information, visual fidelity, and gradient-based fusion performance.

Overall, in the test results on the RoadScene data set, compared with other methods, this paper has better information retention ability and better visual effect.

Table 6. The objective experimental results of different fusion methods show that red font is optimal and blue font is suboptimal.

Method	EN↑	SF↑	SD↑	PSNR↑	MSE↑	MI↑	VIF↑	AG↑	CC↑	SCD↑	Qabf↑
CBF	7.532	0.064	10.670	60.651	0.063	3.520	0.595	6.465	0.368	1.148	0.479
CNN	7.244	0.059	10.360	61.403	0.052	2.614	0.828	5.643	0.502	1.583	0.578
PMGI	7.134	0.047	9.761	60.390	0.063	3.115	0.743	4.599	0.558	1.851	0.473
FusionGAN	7.175	0.034	10.305	59.207	0.062	2.962	0.583	3.347	0.442	1.375	0.272
IFCNN	6.973	0.059	10.357	61.535	0.050	3.008	0.720	5.624	0.532	1.589	0.512
U2Fuison	7.197	0.062	9.827	61.404	0.050	2.868	0.700	6.313	0.529	1.769	0.479
Proposed	7.247	0.073	10.840	60.067	0.071	3.200	0.802	6.962	0.471	1.882	0.559

5. Conclusions

This paper presents an image fusion network based on a special residual network and attention mechanism. A Laplacian-based gradient operator feature extractor in the network is designed to extract fine-grained features from the network images. Secondly, the ECA attention mechanism is introduced to guide the extraction of deep features. In the fusion part, a convolution-layer-based fusion module is designed to fuse deep features. A multi-layer feature fusion method is used to fuse shallow features into the decoder using the maximum pixel value rule. Finally, an appropriate combined loss function is used for network training. Experiments on different types of public data sets show that, compared with the traditional comparison algorithm and the comparison algorithm based on deep learning, the fusion results obtained by the proposed method have better visual effects and better retention effects for visible light information based on prominent infrared information. Objectively speaking, the algorithm in this paper can obtain the optimal or suboptimal value in most evaluation indexes. The eleven categories of metrics for partial image comparison in the public TNO dataset are 72% optimal or suboptimal and 80% optimal or suboptimal in the RoadScene dataset, which is much higher than other algorithms. Overall, the results show that this method is superior to the current advanced infrared image and visible image algorithms.

Author Contributions: Data curation, Y.L. and J.W.; Formal analysis, H.Y., C.Z. and H.C.; Funding acquisition, H.C.; Investigation, Z.L.; Resources, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant 62005320, 61975044).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are indebted for the two anonymous reviewers for their helpful comments to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hui, L.; Manjunath, B.S. Multi-sensor image fusion using the wavelet transform. *Graph. Model. Image Process.* **2002**, *57*, 235–245.
2. Li, S.; Yang, B.; Hu, J. Performance comparison of different multi-resolution transforms for image fusion. *Inf. Fusion* **2011**, *12*, 74–84. [CrossRef]
3. Ben, H.A.; He, Y. A multiscale approach to pixel-level image fusion. *Integr. Comput. Aided Eng.* **2005**, *12*, 135–146.
4. Wang, J.; Peng, J.; Feng, X.; He, G.; Fan, J. Fusion method for infrared and visible images by using non-negative sparse representation. *Infrared Phys. Technol.* **2014**, *67*, 477–489. [CrossRef]
5. Zhang, Q.; Fu, Y.; Li, H.; Zou, J. Dictionary learning method for joint sparse representation-based image fusion. *Opt. Eng.* **2013**, *52*, 057006. [CrossRef]

6. Li, H.; Wu, X.-J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef] [PubMed]
7. Li, H.; Wu, X.-J.; Kittler, J. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [CrossRef]
8. Long, Y.; Jia, H.; Zhong, Y.; Jiang, Y.; Jia, Y. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Inf. Fusion* **2021**, *69*, 128–141. [CrossRef]
9. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An Infrared and Visible Image Fusion Network Based on Salient Target Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
10. Yu, L.; Xun, C. Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* **2016**, *23*, 1882–1886.
11. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]
12. Ma, J.; Wei, Y. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2018**, *48*, 11–26. [CrossRef]
13. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.-P. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef] [PubMed]
14. Nejati, M.; Samavi, S.; Shirani, S. Multi-focus image fusion using dictionary-based sparse representation. *Inf. Fusion* **2015**, *25*, 72–84. [CrossRef]
15. Yang, J.; Wright, J. Image Super-Resolution as Sparse Representation of Raw Image Patches. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, USA, 24–26 June 2008.
16. Yang, B.; Li, S. Visual attention guided image fusion with sparse representation. *Opt. Int. J. Light Electron Opt.* **2014**, *125*, 4881–4888. [CrossRef]
17. Zhang, Y.; Liu, R.S. Sparse representation based multisensor image fusion for multi-focus and multi-modality images: A review. *Inf. Fusion* **2018**, *40*, 57–75. [CrossRef]
18. Zhao, J.; Chen, Y.; Feng, H.; Xu, Z.; Li, Q. Infrared image enhancement through saliency feature analysis based on multi-scale decomposition. *Infrared Phys. Technol.* **2014**, *62*, 86–93. [CrossRef]
19. Ma, J.; Chen, C.; Li, C.; Huang, J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion* **2016**, *31*, 100–109. [CrossRef]
20. Zhao, J.; Cui, G.; Gong, X.; Zang, Y.; Tao, S.; Wang, D. Fusion of visible and infrared images using global entropy and gradient constrained regularization. *Infrared Phys. Technol.* **2017**, *81*, 201–209. [CrossRef]
21. Rajkumar, S.; Chandra Mouli, P.V.S.S.R. Infrared and visible image fusion using entropy and neuro-fuzzy concepts. In *ICT and Critical Infrastructure, Proceedings of the 48th Annual Convention of Computer Society of India, Visakhapatnam, India, 13–15 December 2013*; Springer: Cham, Switzerland, 2014; Volume 248, pp. 93–100.
22. Li, H.; Wu, X.-J.; Durrani, T.S. Infrared and visible image fusion with ResNet and zero-phase component analysis. *Infrared Phys. Technol.* **2019**, *102*, 103039. [CrossRef]
23. Jie, H.; Li, S. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 7132–7141.
24. Li, H.; Wu, X.-J.; Durrani, T. NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [CrossRef]
25. Wang, Q.; Wu, B. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, Seattle, WA, USA, 13–19 June 2020.
26. Prabhakar, K.R.; Srikar, V.S.; Babu, R.V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4724–4732. [CrossRef]
27. Yu, Z.A.; Yu, L.B. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118.
28. Zhou, W.; Bovik, A.C. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image. Process.* **2014**, *13*, 600–612.
29. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss Functions for Image Restoration With Neural Networks. *IEEE Trans. Comput. Imaging* **2017**, *3*, 47–57. [CrossRef]
30. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy. *J. Appl. Remote Sens.* **2008**, *2*, 023522.
31. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 313–315. [CrossRef]
32. Jagalingam, P.; Hegde, A.V. A Review of Quality Metrics for Fused Image. *Aquat. Procedia* **2015**, *4*, 133–142. [CrossRef]
33. Eskicioglu, A.; Fisher, P. Image quality measures and their performance. *IEEE Trans. Commun.* **1995**, *43*, 2959–2965. [CrossRef]
34. Rao, Y.-J. In-fibre bragg grating sensors. *Meas. Sci. Technol.* **1997**, *8*, 355. [CrossRef]
35. Cui, G.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.* **2015**, *341*, 199–209. [CrossRef]
36. Sheikh, H.; Bovik, A. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef] [PubMed]
37. Meng, X.; Shen, H.; Li, H.; Zhang, L.; Fu, R. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Inf. Fusion* **2018**, *46*, 102–113. [CrossRef]

38. Aslantas, V.; Bendes, E. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-Inter-Natl. J. Electron. Commun.* **2015**, *69*, 1890–1896. [CrossRef]
39. Xydeas, C.; Petrović, V. Objective image fusion performance measure. *Electron. Lett.* **2000**, *36*, 308. [CrossRef]
40. Tang, L.; Yuan, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
41. Kumar, B.K.S. Image fusion based on pixel significance using cross bilateral filter. *Signal, Image Video Process.* **2015**, *9*, 1193–1204. [CrossRef]
42. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolution Inf. Process.* **2018**, *16*, 1850018. [CrossRef]
43. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.J.; Ma, J. Rethinking the Image Fusion: A Fast Unified Image Fusion Network based on Proportional Maintenance of Gradient and Intensity. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12797–12804. [CrossRef]
44. Xu, H.; Ma, J. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef]