applied sciences

MDPI

Article

Lightweight Infrared and Visible Image Fusion Based on Nested Connections and Res2Net

Yi Peng [1,2], Xinyue Tu [1,2] and Qingqing Yang [1,2,*]

[1] College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650031, China

[2] Yunnan Key Laboratory of Computer Science, Kunming University of Science and Technology, Kunming 650500, China

[*] Correspondence: kust_txy@163.com

Abstract: Image fusion is a pivotal image-processing technology designed to merge multiple images from various sensors or imaging modalities into a single composite image. This process enhances and extracts the information contained across the images, resulting in a final image that is more informative and of superior quality. This paper introduces a novel method for infrared and visible image fusion, utilizing nested connections and frequency-domain decomposition techniques to effectively solve the problem of lost image detail features. By incorporating depthwise separable convolution technology, the method reduces the computational complexity and model size, thereby increasing computational efficiency. A multi-scale residual fusion network, R2FN (Res2Net Fusion Network), has been designed to replace traditional manually designed fusion strategies, enabling the network to better preserve detail information in the image while improving the quality of the fused image. Moreover, a new loss function is proposed, which is aimed at enhancing important feature information while preserving more significant features. Experimental results on public datasets indicate that the method not only retains the detail information of visible-light images but also highlights the significant features of infrared images while maintaining a minimal number of parameters.

Keywords: image fusion; infrared image; multi-scale features; lightweight model; attention mode

check for updates

Citation: Peng, Y.; Tu, X.; Yang, Q. Lightweight Infrared and Visible Image Fusion Based on Nested Connections and Res2Net. *Appl. Sci.* **2024**, *14*, 4589. https://doi.org/10.3390/app14114589

Academic Editor: Antonio Fernández-Caballero

Received: 18 March 2024
Revised: 25 April 2024
Accepted: 20 May 2024
Published: 27 May 2024

1. Introduction

Image fusion is a process whereby images from diverse sensors or modalities are amalgamated into a singular image. This crucial phase in image processing is designed to extract and enhance information from multiple sources, resulting in a composite image that is both more informative and of superior quality. Based on variations in imaging devices, image fusion tasks are primarily classified into three categories: multimodal image fusion, remote sensing image fusion, and digital photography image fusion [1]. Among these, the fusion of infrared and visible images, as a branch of multimodal image fusion, has been a topic of considerable interest. Visible images, which are captured following the imaging principles of the human eye, are rich in color and texture information, but they are susceptible to obstruction and cannot function in low-light or nighttime conditions. Infrared images emphasize the thermal distribution characteristics of targets and are suitable for low-light and adverse weather conditions, but they lack color and texture information. Therefore, the fusion of these two types of images can result in a composite image that possesses enriched information and enhanced visual perception, providing significant assistance in various application areas, such as target detection [2], medical diagnosis [3], and remote sensing localization [4]. Currently, infrared and visible image fusion methods are typically categorized into two types based on the representation learning techniques employed in their algorithms: traditional image fusion algorithms and those based on deep learning.

Most traditional fusion algorithms primarily utilize signal processing techniques to accomplish fusion tasks. Traditional image fusion techniques commonly employ multi-scale transformations and methods based on sparse or low-rank representations (SRs/LRRs). Among these, methods based on multi-scale transformations, such as the discrete wavelet transform [5], contourlet transform [6], and shearlet transform [7], are notable. The strength of these methods stems from their capability to extract feature information in the frequency domain, which is unattainable in the spatial domain, which helps enhance the performance of fusion algorithms. However, the efficacy of these algorithms is primarily contingent upon the multi-scale transformation operation, which complicates the task of identifying a suitable transformation applicable to diverse image types. On the other hand, the transformation between spatial and frequency domains not only elevates computational complexity but also results in the loss of crucial image features.

Methods based on sparse/low-rank representations (SRs/LRRs), such as SR-HOG [8], DDL [9], JSR [10], and DLRR [11], are applied directly to source images in the spatial domain to extract features, thereby minimizing the loss of feature information typically incurred by transformations between spatial and frequency domains. However, when the source images contain complex information, the performance of these algorithms can drastically decrease.

In recent years, the rapid advancement of deep learning technology has ushered in new opportunities for the fusion of infrared and visible images. The introduction of convolutional neural networks and attention mechanisms has broadened the scope of image fusion, surpassing traditional algorithms and giving rise to a variety of deep learning-based methods. Li et al. [12] introduced the pretrained deep learning network VGG into the image fusion model, significantly improving fusion performance compared to traditional fusion networks. However, since deep learning processing was only added to a few branches and the pretrained structure was not specifically designed for fusion tasks, the features extracted may not necessarily contain complementary information for infrared and visible images.

To accomplish feature extraction and image reconstruction, the autoencoder-based image fusion method first pretrains an autoencoder on a sizable dataset. Then, for image fusion, a hand-crafted fusion approach is used to combine deep features extracted from various source images. Li et al. [13] proposed a novel autoencoder-based image fusion network called DenseFuse, which adopts the network structure of DenseNet [14] to fully extract image features, enriching the extracted features with more abundant information. Then, a designed fusion method is applied for feature-level fusion, and finally, four convolutional layers are used in feature reconstruction to create fused pictures. During the training phase, a large-scale dataset is used to train the autoencoder designed for the fusion task, facilitating the improvement of feature extraction adaptability in various scenarios. However, this network structure is relatively simple and cannot extract multi-scale deep features. To address this issue, many improved autoencoder image fusion algorithms based on the DenseFuse framework have emerged. In 2019, Song et al. [15] proposed the MSDNet algorithm, which extracts multi-scale features and fuses data across all scales by adding convolutional kernels of varying sizes after the encoder. However, while introducing multi-scale features enriches the information of deep features, it also makes the overall fusion network more complex, increasing the computational complexity of the model. Subsequently, Li et al. [16] further improved the network structure based on DenseFuse and proposed the NestFuse image fusion method, which utilizes nest connections to construct the decoder network structure and achieves multi-scale feature extraction. Nevertheless, this model still requires the manual design of fusion methods and cannot perform fusion specifically for the unique information of infrared and visible images.

Although autoencoder-based image fusion methods significantly improve fusion performance compared to traditional methods, they lack specific datasets for multimodal images, resulting in limitations in their expressive power when dealing with complex multimodal images. With the emergence of more multimodal datasets, a plethora of end-to-end fusion methods have emerged, incorporating end-to-end training, a fusion strategy, and deep feature extraction as the three main fusion process components. In 2017, Prabhakar et al. [17] introduced the DeepFuse model, which was the first to apply an end-to-end network to image fusion. However, its very simplistic network topology results in information loss, as it just uses the final layer's output. Ma et al. [18] utilized GAN [19] to fuse infrared and visible images. They achieved this by creating an adversarial relationship between the two types of data. However, FusionGAN utilizes content loss and discriminator loss as loss functions, resulting in fused images with fewer texture details. Subsequently, Ma et al. [20] improved upon FusionGAN with FusionGANv2, introducing novel loss functions such as detail loss and target edge-enhancement loss to preserve the detailed information of target edges. To solve many kinds of image fusion challenges, Zhang et al. [21] manually assembled a multi-focus image dataset and used a CNN that had already been trained. However, the network's results were constrained when used for other image fusion tasks because it was trained on a dataset of images with several foci. Following this, Xu et al. [22] proposed a unified image fusion method that maintains the adaptive similarity between the fusion results and source images by leveraging adaptiveness. However, the loss function employed in U2Fusion, designed solely around gradient-based adaptiveness, fails to fully capture the significance of source images across various fusion subtasks. For instance, in the fusion task of infrared and visible-light images, compared to infrared images, visible-light images exhibit more texture details and dominant gradient clues, resulting in fusion results biased toward visible-light images. Li et al. [23], building upon NestFuse, designed a fusion network called the residual fusion network (RFN) to replace manually designed fusion strategies. By employing a two-stage training method, the RFN retains detailed information and salient features in the fusion features, significantly enhancing the fusion performance of the network. However, the complexity of the network structure leads to a large number of model parameters.

To address the issues present in the aforementioned image fusion networks, we propose a lightweight multi-scale infrared and visible image fusion method based on nested connections and Res2Net. This innovative combination not only enhances the feature extraction process, thereby substantially enhancing the quality of the fused images, but also ensures low computational complexity. Compared to existing techniques, the nested connection structure that we introduce can integrate multi-scale information more deeply, a facet often overlooked in traditional image fusion methods. Furthermore, by designing the multi-scale residual fusion network R2FN to replace traditional manually designed fusion strategies, our method can effectively highlight key information in the images, thereby enhancing their expressiveness while preserving image details. The introduction of depthwise separable convolution significantly reduces computational complexity and memory requirements, making the algorithm suitable for resource-constrained mobile devices. The main contributions of our algorithm are summarized as follows:

1. We employ the frequency-domain decomposition technique to split the source image into detail and base layers, allowing the network to operate on the image with greater precision.
2. We incorporate depthwise separable convolution into the infrared and visible-light image fusion network. In comparison to existing classical fusion methods, our network achieves the lowest number of parameters without compromising performance.
3. We propose a multi-scale residual fusion module (R2FN) to replace manually designed fusion strategies, enabling the effective fusion of features across multiple scales.
4. We design a new loss function that preserves detail information while enhancing salient target features.
5. We conduct experiments on the TNO dataset to test the proposed fusion method. Comparative analysis with existing classical image fusion algorithms demonstrates that our method achieves optimal performance in these fusion tasks.

2. Related Works

2.1. Res2Net

To improve the multi-scale representation capabilities of CNNs, GAO et al. [24] proposed a novel multi-scale backbone architecture called Res2Net, which is utilized for object detection, class activation mapping, and salient object detection. This method divides the input features into multiple branches, with each branch responsible for extracting features at different scales. These branches are connected together in a manner similar to residual connections, enhancing the scale representation capability of features. The framework of Res2Net is illustrated in Figure 1.

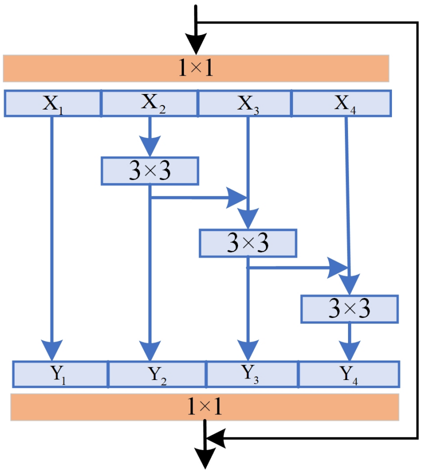


Figure 1. Architecture of Res2net.

The internal connectivity of Res2Net is similar to that of ResNet [25], with the distinction that, in Res2Net, the 3 × 3 convolutions are decoupled. The input features are segmented into multiple groups, each of which is processed by a corresponding set of filters to extract features. Subsequently, the output features of the preceding group are combined with the input features of the subsequent group and processed by the next group of filters. This process is iterated multiple times until all groups of features have undergone processing. Finally, the feature maps from all groups are concatenated and subjected to a set of 1 × 1 filters and then concatenated with the original features to derive the final result. Through this approach, Res2Net enhances the network's performance and representation capability by increasing the effective receptive field and generating multi-scale feature representations.

2.2. Depthwise Separable Convolution

Depthwise separable convolution (DSC), first proposed by Sifre et al. [26], gained widespread recognition when it was introduced in the MobileNet model by the Google team in 2017 [27]. The fundamental concept of MobileNet involves significantly reducing computational complexity and model size by employing DSC.

Depthwise separable convolution comprises two processes: Depthwise Convolution (DW) and Pointwise Convolution (PW). In DW, the number of convolutional kernels matches that of the input channels, thereby establishing a one-to-one correlation between channels and kernels. Consequently, in DW, the number of output feature maps matches that of the input channels. PW then convolves the output feature maps from DW with convolutional kernels, ensuring that each output feature map integrates information from all input feature maps. The schematic diagram of depthwise separable convolution is illustrated in Figure 2.

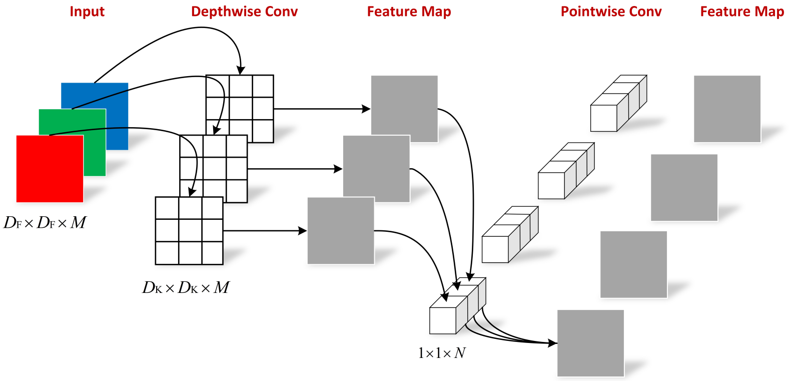


Figure 2. Depth-separable convolution schematic.

In the figure, the input image size is denoted by $D_F \times D_F \times M$ and is convolved with convolution kernels of size $D_K \times D_K \times M$ to obtain M-channel feature maps. Then, the M-channel feature maps are input to convolution kernels of size $1 \times 1 \times N$, resulting in N-channel feature maps. The computational complexity of the entire process is as follows:

$$D_F \cdot D_F \cdot M \cdot D_K \cdot D_K + D_F \cdot D_F \cdot M \cdot N \tag{1}$$

If the input image with the size $D_F \times D_F \times M$ is convolved using regular convolution with kernels of size $D_K \times D_K \times M$ to obtain the same feature maps as in the above process, the computational complexity is as follows:

$$D_K \cdot D_K \cdot D_F \cdot D_F \cdot M \cdot N \tag{2}$$

The ratio between (1) and (2) is

$$\frac{1}{{D_K}^2} + \frac{1}{N} \tag{3}$$

In feature extraction, the commonly chosen kernel size is 3 × 3. Therefore, theoretically, depthwise separable convolution reduces computation by a factor of 8–9 compared to regular convolution.

MobileNetv3, introduced by [28], incorporates depthwise separable convolution, inverted residual blocks, and Squeeze-and-Excitation (SE) modules [29]. The input feature map is initially expanded through convolutional layers to extract additional features. Subsequently, Depthwise Convolution (DW) is applied, followed by the SE module to adjust the weights of each channel, thereby enhancing the model's accuracy. Finally, downsampling is performed through convolutional layers. When the number of input and output features matches, shortcut connections are utilized by the Bottleneck (Bneck). The structure of the Bneck network is depicted in Figure 3.

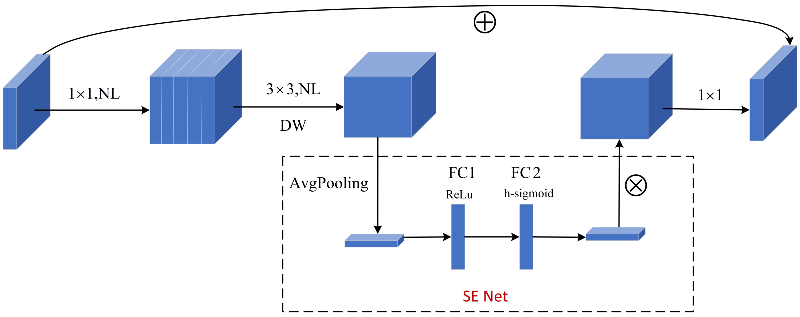


Figure 3. Bneck of MobileNetv3 structure.

3. Approach

3.1. An Overview of the Proposed Method

We propose a lightweight multi-scale infrared and visible image fusion method. The method first divides the original visible-light and infrared images into base and detail layers using mutually guided image filtering (muGIF) [30], which allows for extracting more hierarchical representations in high-frequency and low-frequency domains. The base layer encompasses information such as image content and spatial structure, whereas texture and local shape information are contained within the detail layer. Subsequently, sub-images at the same hierarchical level are input into the image fusion network for fusion. Finally, the fused images from both high-frequency and low-frequency components are merged to derive the final fused image. The flowchart of the proposed algorithm is illustrated in Figure 4.

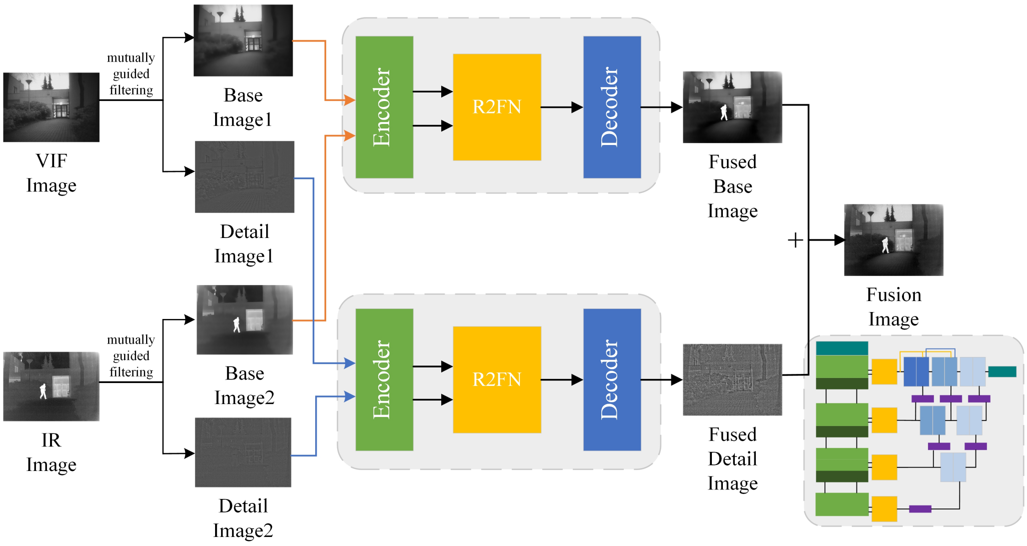


Figure 4. Two-layer fusion framework of the proposed method.

The decomposition process of the source image primarily consists of two steps. Firstly, the image's base layer is obtained through the muGIF method, which can be calculated using (4).

$$I_{\text{base}} = \text{muGIF}(I_{\text{i}}, \alpha, T) \tag{4}$$

Here, I_{base} represents the base layer image, muGIF denotes the mutually guided image filtering operation, I_{i} stands for the source image, α is the parameter controlling the extent of texture removal, and T represents the number of iterations. We set α to 0.003 and T to 3.

After extracting the base layer, the detail layer image is obtained through the operation in (5):

$$I_{\text{detail}} = I_{\text{i}} - I_{\text{base}} \tag{5}$$

When fusing sub-images at the same hierarchical level, we propose a lightweight multi-scale infrared and visible fusion network. Taking the fusion process of the base layer as an example, its architecture is illustrated in Figure 5. We draw inspiration from the RFN-Net's network structure, where the fusion network consists of encoder, fusion, and decoder modules. We introduce depthwise separable convolution into the encoder and decoder networks, replacing conventional convolutions in the original network to address the issue of the relatively large parameter size. The encoder module of the encoder network comprises two improved bneck layers and a max-pooling layer. Through this combination, the encoder can extract multi-scale depth features with a smaller computational cost. The multi-scale fusion network R2FN is employed to integrate multimodal depth features extracted at each scale. The fused features are then input into the decoder with a nested connection structure. The advantage of this structure is its ability to avoid information loss from previous layers during convolution operations, thereby fully utilizing multi-scale features for image reconstruction.

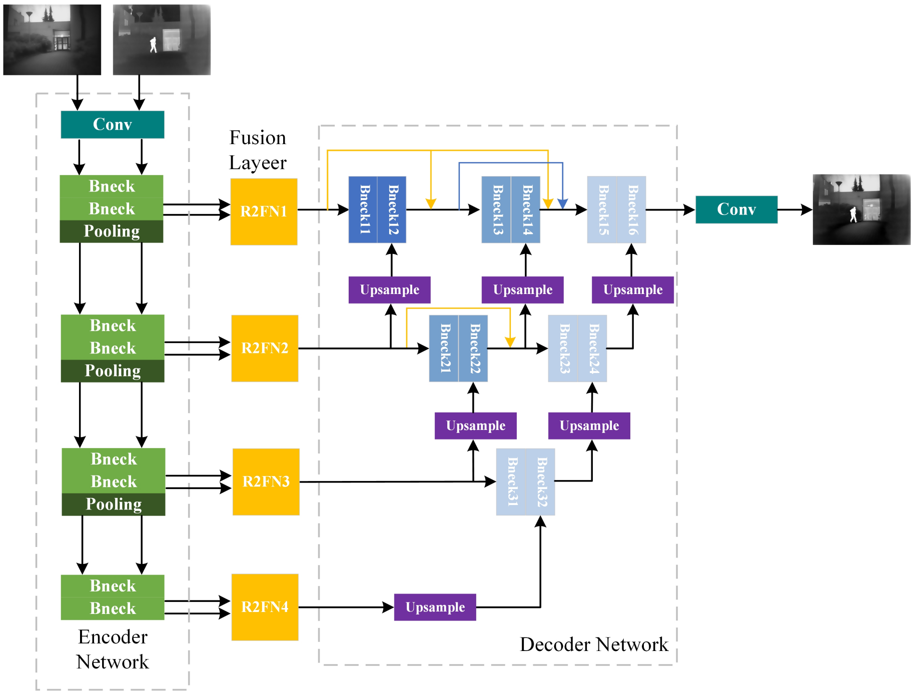


Figure 5. Framework of proposed method.

We adopted the Bottleneck (Bneck) structure from MobileNetv3. DSC is capable of reducing the number of model parameters and computational complexity, but it might also adversely affect the convolution's capacity to extract features. Therefore, the attention module CBAM [31] is introduced to enhance the model's attention concentration ability and improve the information-processing mechanism, effectively improving the quality of image fusion and the overall performance of the model. CBAM, as a lightweight and versatile attention mechanism, can be easily added to the convolutional layers of any network at a minimal cost. CBAM applies attention mechanisms simultaneously in both the channel and spatial dimensions, enhancing the model's accuracy. Additionally, the parameter size of the improved network has been further decreased. The enhanced Bneck structure is illustrated in Figure 6.

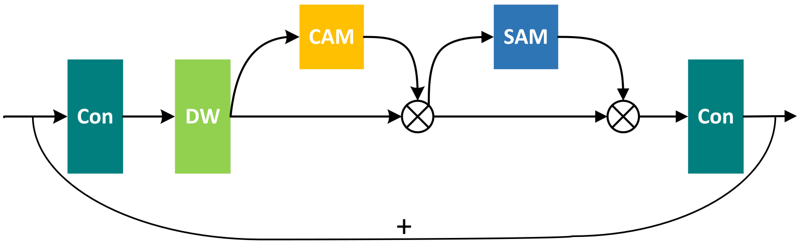


Figure 6. The structure of the enhanced Bneck.

3.2. Fusion Network

The fusion network R2FN, tailored for the dual-modal image fusion task, is designed based on the Res2Net architecture. In the fusion network, the parameters of R2FN vary across different layers. The structure of the R2FN network is illustrated in Figure 7.

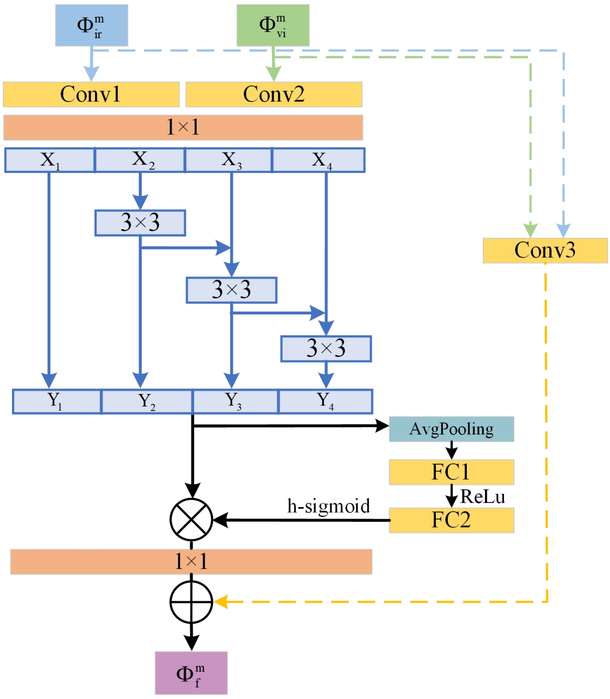


Figure 7. The architecture of R2FN.

In the figure, Φ^{m}_{ir} and Φ^{m}_{vi} represent the infrared and visible depth features extracted by the encoder network, respectively. Initially, the outputs of Conv1 and Conv2 are concatenated and fed into a 1 × 1 convolution for channel transformation. The feature maps are subsequently partitioned into s subsets, each characterized by an identical spatial size and 1/s of the total channel count, with s serving as the scale control parameter. The first subset, X_1, remains unchanged and is directly propagated to Y_1, while the remaining subsets undergo 3 × 3 convolution operations before being added to the next feature subset. Subsequently, the acquired feature maps are fed into the SE module to modify the weights of each channel, thereby improving the model's accuracy. Here, ReLU and h-sigmoid are sequentially utilized as activation functions. Finally, the results from the 1 × 1 convolution are added to the fusion convolutional layer Conv3 to obtain the final outcome C. In our experiments, we selected s = 4 as the scale control parameter.

4. Training Strategy

During the training phase, our image fusion network needs to possess superior performance based on two key factors: one is the feature extraction capability of the encoder network and the feature reconstruction capability of the decoder network, and the other is the capability of R2FN to extract dual-mode multi-scale features. Therefore, a two-stage training method is adopted in this study. Firstly, the encoder network and the decoder network are trained as a whole, with the objective of reconstructing the network input. Then, R2FN is trained using multimodal images, with the parameters of the encoder and decoder obtained from the first stage being fixed during this phase.

4.1. Training in the First Stage

We are essentially training an autoencoder network to recreate the input images during the training phase because the fusion layer of the network is dropped. The formulation of the loss function is a critical factor that impacts the quality of image fusion outcomes. In the case of the autoencoder-based image fusion network, the loss function calculates the loss between the reconstructed image and the source image to supervise the learning process. While simultaneously imposing constraints on the output image to maintain consistency in texture details with the input image, our goal is for it to share greater structural and intensity distribution similarity with the source image. Considering these factors, we introduce similarity loss L_{sim} and gradient loss L_{grad} to formulate the total loss function L, defined as follows:

$$L = L_{\text{sim}} + \lambda L_{\text{grad}} \tag{6}$$

where L_{sim} is used to retain important information from the source image. It limits the resemblance between the fusion result and the source image, ensuring that the fusion result retains the essential features of the source image to the greatest extent possible, thereby enhancing the quality and perceptual effect of the fusion result. L_{grad} is used to constrain the fusion result to maintain consistent gradient information and texture features with the source image. λ represents the balance parameter, which is used to adjust the balance between the loss terms L_{sim} and L_{grad}, keeping different loss terms on the same scale. This enables the encoder and decoder to balance information from different modalities when dealing with infrared and visible images.

To determine L_{sim}, we utilize two metrics, SSIM and MSE, to comprehensively assess the similarity of fusion results. The SSIM is the most commonly used metric to assess the similarity between two images. It assesses their similarity by comparing the brightness, contrast, and structural information of the two images. The SSIM values range from - 1 to 1, with a value closer to 1 indicating a higher similarity between the two images. To minimize the loss, we use the dissimilarity between the two images to represent the structural similarity loss L_{ssim}, which is calculated using (7):

$$L_{\text{ssim}} = 1 - \text{SSIM}(X, Y) \tag{7}$$

where X represents the output image, Y represents the input image, and $\text{SSIM}(\cdot)$ represents the structural similarity operation. It is worth noting that the SSIM primarily focuses on changes in contrast and structure, with weaker constraints on intensity distribution differences. Therefore, we introduce MSE as a supplement. MSE is a metric that measures the error between two images. Using MSE as the loss function ensures that the distribution of pixel intensities in the input and output images are similar in image fusion tasks. L_{mse} can be calculated using (8):

$$L_{\text{mse}} = \frac{1}{HW} \sum_i \sum_j (X_{i,j} - Y_{i,j})^2 \tag{8}$$

where H and W represent the height and width of the image, respectively. X represents the output image, and Y represents the input image. i, j represent the pixel values in row i and column j. Since the scales of L_{ssim} and L_{mse} are different, we introduce a balance parameter, μ, to control the balance between the two terms. The final expression for L_{sim} is as shown in (9):

$$L_{\text{sim}} = \mu L_{\text{ssim}} + L_{\text{mse}} \tag{9}$$

We utilize gradient operators to compute the gradients of both the input image and the output image, followed by the calculation of the Euclidean distance between them. Gradient operators can compute the gradient values of each pixel in the image, representing the rate of color change at that pixel. Therefore, the gradient loss ensures that the output image has similar texture details to the input image, thereby improving the quality of the fusion result. L_{grad} can be calculated using (10):

$$L_{\text{grad}} = \frac{1}{HW} \sum_i \sum_j \sqrt{(\nabla X_{i,j} - \nabla Y_{i,j})^2} \tag{10}$$

where $\nabla(\cdot)$ represents the gradient operator, which can calculate the gradient values for each pixel in the image, and $\sqrt{(\nabla X_{i,j} - \nabla Y_{i,j})^2}$ represents the Euclidean distance between the input image and the output image at pixel (i, j). A smaller L_{sim} indicates that the texture details in the output image are more similar to those in the input image, leading to higher fusion result quality.

4.2. Training in the Second Stage

In the fusion layer, the multimodal multi-scale feature fusion module R2FN is designed to replace the manually designed fusion strategies typically used in autoencoder-based fusion networks. In the second stage of training, the focus is on training the R2FN module to enhance its capability of extracting multimodal multi-scale features. The R2FN module is trained using multimodal images, aiming to optimize its performance in effectively fusing multimodal multi-scale features. The parameters of the encoder and decoder obtained from the first stage are kept fixed to ensure consistency in the features extracted and reconstructed by these networks. Subsequently, a loss function tailored for R2FN is designed to train the multi-scale depth feature fusion network.

The fixed encoder network is employed to extract multi-scale features from the source images, with the features at each scale being fused by the corresponding R2FN. The fused multi-scale features are then used as inputs to the decoder network to reconstruct the fused image. We define a loss function L_{R2FN} as the training loss for R2FN. L_{R2FN} consists of two components: detail loss (L_{detail}) and feature enhancement loss (L_{feature}), defined as follows:

$$L_{R2FN} = \beta L_{\text{detail}} + L_{\text{feature}} \tag{11}$$

where β represents the balancing parameter between L_{detail} and L_{feature}.

In infrared and visible image fusion networks, the visible image typically contributes texture details in the background. Therefore, we define the detail preservation loss by computing the structural similarity loss of the visible-light image. It is defined as follows:

$$L_{\text{detail}} = 1 - \text{SSIM}(O, I_{vi}) \tag{12}$$

In infrared images, more salient object features are typically present. Therefore, a feature enhancement loss function is designed to enhance salient feature information. It is defined as follows:

$$L_{\text{feature}} = \sum_{m=1}^{M} \omega_1(m) \cdot \left[\left(\phi_f^m - \phi_{ir}^m \right)^2 \cdot \omega_{ir} + \left(\phi_f^m - \phi_{vi}^m \right)^2 \cdot \omega_{vi} \right] \tag{13}$$

where M represents the number of multi-scale features obtained through downsampling, $\omega_1(m)$ represents the balancing parameter for the m-th multi-scale feature, and ω_{ir} and ω_{vi}, respectively, represent the balancing parameters controlling the ratio of visible-light depth features and infrared features.

5. Experiments and Results Analysis

5.1. Dataset and Experimental Environment

In order to verify the efficacy of our method, in the training stage, we selected 80,000 images from the MS-COCO dataset [32] as the first-stage training set and utilized the KAIST dataset [33] as the second-stage training set. In the first-stage training, the balancing parameter λ between the similarity loss and the gradient loss in the loss function was set to 1, and μ was set to 100. In the second-stage training, β was set to 500, ω_{ir} was set to 5, and ω_{vi} was set to 3. The model training parameters were set as follows: epochs = 20; batch size = 4.

During the testing phase, to verify the effectiveness of our method, images were selected from the publicly available infrared and visible-light dataset TNO [34] for experimentation. Six sets of images were chosen for comparative analysis. These images, rich in detail and texture, are suitable for assessing the quality of image fusion. Having been widely used in previous studies, they provide a benchmark for comparing our results with existing methods.

Our experiments were conducted on a system running Windows 11 with hardware specifications that include an Intel(R) Core(TM) i5-12400F 2.50 GHz processor. The model was run on an NVIDIA GeForce RTX 3060 GPU. The software environment for the experiments included Python 3.8.3, PyTorch 1.10.1, CUDA 11.3, and the PyCharm 2020.1 IDE.

5.2. Evaluation Metrics

To objectively evaluate the algorithm, we selected Entropy, the Standard Deviation, and the Structural Similarity Measure (SSIM) as objective evaluation metrics to assess the amount of edge, texture, and contrast information in the fused images. Mutual Information, Feature Mutual Information using Discrete Cosine Transform (FMIdct), Feature Mutual Information using Wavelet Transform (FMIw), and Visual Information Fidelity were used to evaluate the distortion, noise, and artifacts caused by fusion, similarity, and the transfer of complementary information between the fused and source images. We also used the “params" metric to evaluate the model’s size.

EN is used to assess the information content of fused images, with higher values indicating richer content, and is crucial for evaluating fusion effectiveness. The SD measures pixel dispersion, reflecting image contrast, which is important for enhancing visibility and details. The SSIM evaluates the structural similarity between the fused and original images, with high values showing the effective preservation of visual features. MI assesses the degree of information correlation between the fusion result and original images, indicating the preservation of original data. FMIdct and FMIw assess the Mutual Information of discrete cosine and wavelet features, respectively, reflecting the algorithm’s ability to retain significant original features. VIF, which assesses the visual quality of the fusion result, shows that higher values indicate greater fidelity to human visual perception, representing better quality. The model’s size and computational complexity are critical for practical applications, with models having fewer parameters being easier to deploy in resource-limited environments, reducing energy and operational costs.

In comparative experiments, we selected six classical image fusion algorithms as benchmarks: DeepFuse (CNN-based fusion), DenseFuse (DenseNet-based fusion with autoencoders), NestFuse (fusion with nested connections and spatial/channel attention models), FusionGAN (GAN-based fusion), U2Fusion (end-to-end unsupervised fusion), and IFCNN (fusion using multiple fusion strategies).

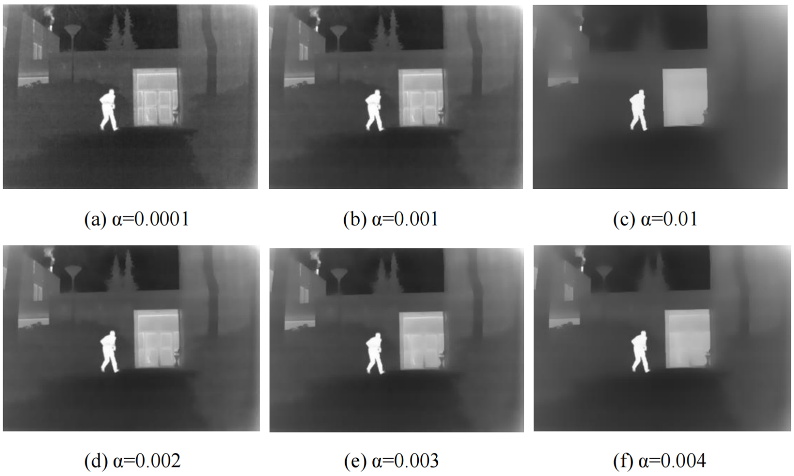
5.3. Ablation Study

To validate the optimization effects of our various strategies and assess the effectiveness of the proposed methods, we designed and conducted ablation experiments. These experiments aimed to further evaluate the impact of improved techniques on the performance of image fusion.

5.3.1. Frequency-Domain Decomposition

In this section, an analysis is conducted on the frequency-domain decomposition module, and the influence of different parameters of α in the guided filtering operation (muGIF) on the network is examined.

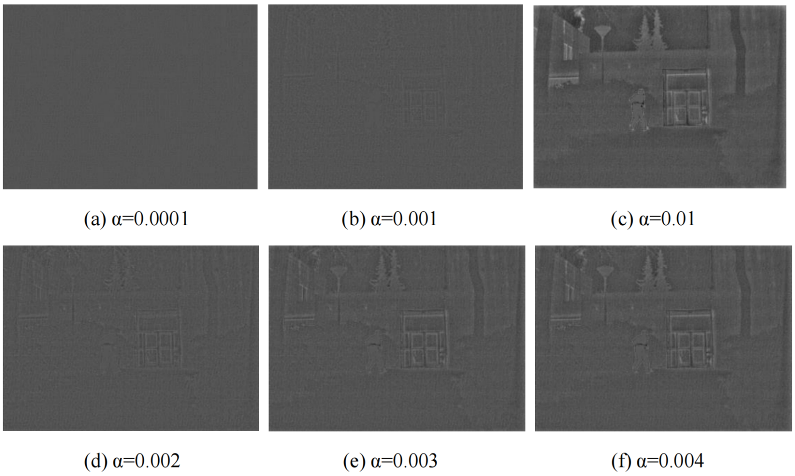
As discussed in Section 3.1, mutually guided image filtering is employed during frequency-domain decomposition to decompose the input source image into base and detail layers. The quality of filtering critically impacts the final image fusion performance. Therefore, in the experiments of this section, α is set to 0.0001, 0.001, 0.01, 0.002, 0.003, and 0.004 to analyze its influence on the filtering effect. (Only the experimental results of infrared image frequency-domain decomposition are listed in this paper, and it is observed that the trends in the decomposition effects of visible-light images are consistent with those of infrared images.) The experimental results are shown in Figures 8 and 9.



(a) α=0.0001 (b) α=0.001 (c) α=0.01

(d) α=0.002 (e) α=0.003 (f) α=0.004

Figure 8. The infrared base layer images corresponding to different α values of mutually guided filtering.



(a) α=0.0001 (b) α=0.001 (c) α=0.01

(d) α=0.002 (e) α=0.003 (f) α=0.004

Figure 9. The infrared detail layer images corresponding to different α values of mutually guided filtering.

From the figure, it can be observed that, with the increase in α, the base layer gradually becomes smoother, as larger α values result in the removal of more high-frequency details. When α increases to a certain extent, such as α = 0.01, the detail layer loses too much high-frequency information, leading to less prominent image details. At α = 0.003, the base layer demonstrates moderate smoothness, removing an appropriate amount of high-frequency detail information while still retaining sufficient structural information. The detail layer preserves more texture and edge information without excessive smoothing, indicating that the filter can better distinguish between the base content and detail content. Therefore, we select 0.003 as the value of α.

After determining the value of α, to verify the impact of the frequency-domain decomposition module on the performance of image fusion, we conducted ablation experiments targeting this module using the same image fusion network. The experimental results are presented in Table 1.

Table 1. The average values of objective metrics obtained without using the frequency-domain decomposition module and with the frequency-domain decomposition module included.

	EN	**SD**	**SSIM**	**MI**	**FMI$_{dct}$**	**FMI$_{w}$**	**VIF**
Exp.1	6.7597	42.4632	0.7348	14.0251	0.3783	0.4132	0.6531
Exp.2	7.0876	45.9123	0.7564	14.2398	0.3975	0.4371	0.7360

In the table, Experiment 1 corresponds to the fusion results without the frequency-domain decomposition module, while Experiment 2 corresponds to the fusion results with the inclusion of the frequency-domain decomposition module. Clearly, after incorporating the frequency-domain decomposition module, the numerical values of various evaluation metrics improved, validating the effectiveness of this module in enhancing the performance of the image fusion network.

5.3.2. The Loss Function during the First Stage of Training

The first stage of training focuses on the feature extraction capability of the encoder and the reconstruction ability of the decoder, independent of the fusion layer. To balance the magnitudes of different loss terms in the loss function, we introduce the balancing parameters λ and μ, where λ is used to balance the magnitude difference between L_{sim} and L_{grad}, and μ is used to balance the magnitude difference between L_{ssim} and L_{mse}. We assessed the average values of objective metrics under different magnitude combinations to validate the optimal combination of balancing parameters. The experimental results are presented in Table 2. The optimal values are highlighted in bold font.

Table 2. The average values of objective metrics obtained by setting different balancing parameters for the loss functions during the first stage of training.

μ	λ	**EN**	**SD**	**SSIM**	**MI**	**FMI$_{dct}$**	**FMI$_{w}$**	**VIF**
1	0.1	6.6176	41.9542	0.7021	13.4094	0.3125	0.4003	0.6629
	1	6.6268	42.1055	0.7113	13.4598	0.3317	0.4206	0.6831
	10	6.6341	42.2104	0.7168	13.4783	0.3354	0.4239	0.6876
10	0.1	6.6194	41.9845	0.7081	13.4297	0.3177	0.4056	0.6687
	1	6.6331	42.1588	0.7170	13.4722	0.3380	0.4261	0.6888
	10	6.6404	42.2558	0.7195	13.4984	0.3411	0.4291	0.6922
100	0.1	6.6430	42.2992	0.7129	13.4419	0.3233	0.4119	0.6742
	1	**6.6659**	**42.6404**	0.7186	**13.8028**	**0.3511**	**0.4387**	**0.7065**
	10	6.6593	42.4887	0.7217	13.5089	0.3461	0.4340	0.6969
1000	0.1	6.6465	42.3545	0.7082	13.4724	0.3294	0.4166	0.6805
	1	6.6567	42.5179	0.7192	13.5304	0.3394	0.4263	0.6899
	10	6.6654	42.6293	**0.7254**	13.5709	0.3448	0.4327	0.7019

From Table 2, it can be observed that when μ is set to 100 and λ is set to 1, the image fusion network exhibits better performance.

5.3.3. The Loss Function during the Second Stage of Training

We kept the numerical values obtained in the first stage unchanged and conducted an ablation study on the balancing parameters of the loss function during the second stage of training. Referring to the conclusions drawn in reference[23], we first set β to 700 and conducted ablation experiments for both ω_{vi} and ω_{ir}. The experimental results are presented in Table 3.

From Table 3, it can be observed that the combination of ω_{ir} = 5 and ω_{vi} = 3 performs the best across almost all key performance indicators. This combination not only maintains the richness of image information and contrast but also effectively preserves the structural similarity, feature information, and visual fidelity of the images. Therefore, this combination is considered the optimal parameter setting, providing the best image fusion results.

To find the optimal value of β, which controls the balance between L_{detail} and L_{feature}, we conducted an ablation study by setting ω_{ir} to 5 and ω_{vi} to 3. Due to the larger difference in magnitudes between L_{detail} and L_{feature}, we experimented with β values of 100, 300, 500, 700, and 1000. The experimental results are presented in Table 4.

Table 3. The average values of objective metrics obtained were calculated by setting β to 700 during the second stage of training and by varying the values of ω_{ir} and ω_{vi}. The bold indicates the optimal value.

ω_{ir}	ω_{vi}	EN	SD	SSIM	MI	FMI$_{dct}$	FMI$_{w}$	VIF
2	2	6.8374	43.3473	0.7117	13.8329	0.3211	0.4121	0.6826
3	2	6.8479	43.5035	0.7322	13.8838	0.3415	0.4334	0.7038
	3	6.8558	43.6115	0.7380	13.9025	0.3448	0.4366	0.7086
4	2	6.8397	43.3764	0.7172	13.8530	0.3264	0.4178	0.6886
	3	6.8544	43.5587	0.7377	13.8955	0.3475	0.4389	0.7092
	4	6.8624	43.6606	0.7412	13.9225	0.3505	0.4419	0.7131
5	2	6.8655	43.7051	0.7230	13.8641	0.3324	0.4240	0.6945
	3	**6.8911**	43.9084	**0.7561**	**14.2369**	**0.3607**	**0.4505**	0.7240
	4	6.8833	44.1212	0.7473	13.9342	0.3557	0.4475	0.7174
	5	6.8694	43.7629	0.7278	13.8762	0.3382	0.4298	0.7003
6	2	6.8803	43.9310	0.7396	13.9350	0.3489	0.4408	0.7117
	3	6.8905	**44.1225**	0.7515	13.9753	0.3543	0.4461	0.6980
	4	6.8730	43.8484	0.7326	13.8878	0.3362	0.4274	**0.7272**
	5	6.8612	43.6517	0.7197	13.8518	0.3240	0.4152	0.6860
	6	6.8494	43.4549	0.7066	13.8164	0.3116	0.4030	0.6741

Table 4. The average values of objective metrics obtained were calculated by setting ω_{ir} to 5 and ω_{vi} to 3 during the second stage of training while varying the values of β. The bold indicates the optimal value.

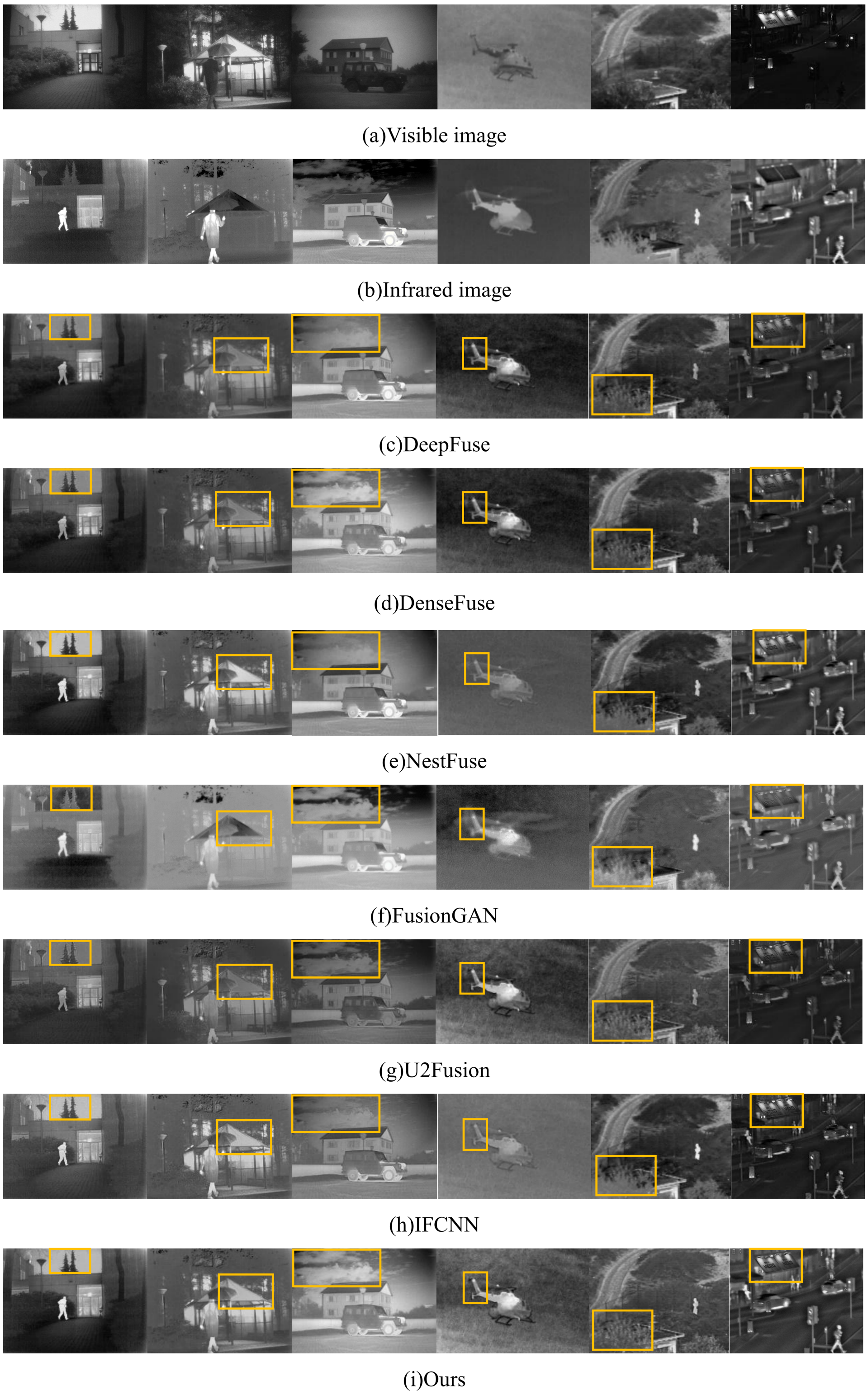
β	EN	SD	SSIM	MI	FMI$_{dct}$	FMI$_{w}$	VIF
100	6.6453	43.7460	0.7438	13.9182	0.3405	0.4230	0.7141
300	6.7841	44.6525	0.7532	14.0203	0.3544	0.4366	0.7324
500	**7.0876**	**45.9123**	**0.7564**	**14.2398**	**0.3975**	0.4371	**0.7360**
700	6.8911	43.9084	0.7561	14.2369	0.3607	**0.4505**	0.7240
1000	6.7140	43.5906	0.7505	14.1427	0.3460	0.4312	0.7244

From Table 4, it can be observed that when β is set to 500, all metrics except for FMI$_{w}$ are at their optimal values, with the value of FMI$_{w}$ being only slightly below the optimum. Considering all key performance indicators comprehensively, β = 500 offers the best overall performance. Therefore, we set β to 500 in our experiments.

5.4. Results Analysis and Comparison

5.4.1. Subjective Evaluation

To validate the effectiveness of our proposed method, we conducted a subjective comparative experiment on a subset of images from the TNO dataset, comparing them with various image fusion algorithms. The comparative results of each algorithm are shown in Figure 10. The comparison results show that the FusionGAN method fails to effectively preserve detailed texture information from the visible-light images. NestFuse and IFCNN methods demonstrate a good representation of the target contours but do not effectively retain the thermal radiation information from the infrared images. The DeepFuse, DenseFuse, and U2Fusion methods exhibit clear contour information and target features, but the introduction of excessive noise leads to poor fused image quality. Particularly in the yellow-boxed areas in Figure 10, none of these algorithms achieve the desired fusion results. In contrast, our algorithm maintains a better balance of information between the infrared and visible-light images in most fusion scenarios, resulting in superior fusion results.



(a)Visible image

(b)Infrared image

(c)DeepFuse

(d)DenseFuse

(e)NestFuse

(f)FusionGAN

(g)U2Fusion

(h)IFCNN

(i)Ours

Figure 10. Infrared and visible-light fusion results of each algorithm.

5.4.2. Objective Evaluation

To further validate the effectiveness of our proposed algorithm, we selected eight objective evaluation metrics for comparative analysis, and the comparative results are shown in Table 5.

Table 5. The average quantitative value of each evaluation index. The bold indicates the optimal value.

	EN	SD	SSIM	MI	$\mathbf{FMI}_{dct}$	$\mathbf{FMI}_{w}$	VIF	Params
DeepFuse	6.7581	41.8988	0.7329	13.5161	0.3777	0.4146	0.6582	2.9653M
DenseFuse	6.7634	41.9003	0.7335	13.5268	0.3810	0.4192	0.6573	3.1317M
NestFuse	6.9355	43.5519	0.6942	13.8709	0.3321	**0.4378**	0.7249	10.9310M
FusionGAN	6.5440	38.3732	0.6933	13.0673	0.2831	0.3534	0.5082	7.9873M
U2Fusion	6.4722	30.0504	0.7526	12.9444	0.3077	0.3540	0.4998	5.9896M
IFCNN	6.9521	44.5987	0.7054	13.9041	0.3574	0.4275	0.7279	8.4360M
Ours	**7.0876**	**45.9123**	**0.7564**	**14.2398**	**0.3975**	0.4371	**0.7360**	**2.0432M**

From Table 5, it is evident that our proposed method achieves the optimal performance across all seven metrics except FMIw. Additionally, it demonstrates significant advantages in terms of model complexity and computational efficiency. This demonstrates that the algorithm can effectively preserve detailed information from visible-light images while highlighting significant features from infrared images. Moreover, it maintains a relatively small parameter count. The reduced number of parameters implies a lighter-weight model and lower training and deployment costs, which are particularly suitable for resource-constrained environments such as mobile devices and real-time systems. This highlights the high practical value of our method, not only theoretically and experimentally excellent but also highly applicable in real-world scenarios.

6. Conclusions

In this article, we propose a novel method for lightweight infrared and visible image fusion based on nested connections and Res2Net. This method combines frequency-domain decomposition, depthwise separable convolution, nested connection networks, and multi-scale residual networks, achieving multi-scale feature extraction while maintaining a small model parameter count. Prior to inputting the source images into the fusion network, a mutually guided filtering operation is applied to better extract hierarchical representations of the images' high- and low-frequency domains. By improving depthwise separable convolution, the model reduces computational complexity while maintaining high fusion quality. Multi-scale feature extraction is realized through the use of nested connection structures. Through the designed R2FN network, image details are effectively preserved, and significant features of the infrared images are highlighted. Experimental comparisons with several classical image fusion algorithms, in terms of subjective and objective evaluations, demonstrate the superiority of the proposed method across multiple key performance indicators. Notably, the method exhibits advantages in lightweighting, significantly reducing the computational burden, while also enhancing feature extraction and fusion capabilities through the nested connection architecture and the R2FN fusion module. Consequently, this study not only advances the theoretical and practical aspects of image fusion technology but also opens up new pathways for its application in high-dynamic and dynamic environments.

Author Contributions: Conceptualization, X.T.; methodology, X.T.; software, X.T.; validation, X.T., Y.P., and Q.Y.; formal analysis, Y.P.; investigation, X.T.; resources, X.T.; data curation, X.T.; writing—original draft preparation, X.T.; writing—review and editing, X.T.; visualization, X.T.; supervision, Y.P.; project administration, X.T.; funding acquisition, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Projects of Basic Research Program in Yunnan Province (202401AS070105), the National Natural Science Foundation of China (61761025), and the Development Fund of Key Laboratory of Computer Technology Application in Yunnan Province (2021102).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author due to confidentiality restrictions.

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

References

1. Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; Ma, J. Image fusion meets deep learning: A survey and perspective. *Inf. Fusion* **2021**, *76*, 323–336. [CrossRef]
2. Cao, Y.; Guan, D.; Huang, W.; Yang, J.; Cao, Y.; Qiao, Y. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *Inf. Fusion* **2019**, *46*, 206–217. [CrossRef]
3. Bhatnagar, G.; Wu, Q.J.; Liu, Z. Directive contrast based multimodal medical image fusion in NSCT domain. *IEEE Trans. Multimed.* **2013**, *15*, 1014–1024. [CrossRef]
4. Zhang, H.; Ma, J.; Chen, C.; Tian, X. NDVI-Net: A fusion network for generating high-resolution normalized difference vegetation index in remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 182–196. [CrossRef]
5. Sundararajan, D. *Discrete Wavelet Transform: A Signal Processing Approach*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
6. Xia, J.; Chen, Y.; Chen, A.; Chen, Y. Medical image fusion based on sparse representation and PCNN in NSCT domain. *Comput. Math. Methods Med.* **2018**, *2018*, 2806047. [CrossRef] [PubMed]
7. Peng, H.; Li, B.; Yang, Q.; Wang, J. Multi-focus image fusion approach based on CNP systems in NSCT domain. *Comput. Vis. Image Underst.* **2021**, *210*, 103228. [CrossRef]
8. Zong, J.J.; Qiu, T.S. Medical image fusion based on sparse representation of classified image patches. *Biomed. Signal Process. Control* **2017**, *34*, 195–205. [CrossRef]
9. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Discriminative learned dictionaries for local image analysis. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; IEEE: New York, NY, USA, 2008; pp. 1–8.
10. Zhang, Q.; Fu, Y.; Li, H.; Zou, J. Dictionary learning method for joint sparse representation-based image fusion. *Opt. Eng.* **2013**, *52*, 057006. [CrossRef]
11. Li, H.; He, X.; Tao, D.; Tang, Y.; Wang, R. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognit.* **2018**, *79*, 130–146. [CrossRef]
12. Li, H.; Wu, X.J.; Kittler, J. Infrared and visible image fusion using a deep learning framework. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: New York, NY, USA, 2018; pp. 2705–2710.
13. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef] [PubMed]
14. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. Fusiondn: A unified densely connected network for image fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7 February 2020; Volume 34, pp. 12484–12491.
15. Song, X.; Wu, X.J.; Li, H. MSDNet for medical image fusion. In Proceedings of the Image and Graphics: 10th International Conference, ICIG 2019, Beijing, China, 23–25 August 2019; Part II 10; Springer: Berlin/Heidelberg, Germany, 2019; pp. 278–288.
16. 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]
17. Ram Prabhakar, K.; Sai Srikar, V.; Venkatesh Babu, R. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4714–4722.
18. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
19. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
20. Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J.; Jiang, J. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion* **2020**, *54*, 85–98. [CrossRef]
21. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [CrossRef]
22. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef] [PubMed]
23. 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]
24. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Sifre, L.; Mallat, S. Rigid-motion scattering for texture classification. *arXiv* **2014**, arXiv:1403.1687.

27. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
28. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
30. Guo, X.; Li, Y.; Ma, J. Mutually guided image filtering. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1283–1290.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
33. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA , 7–12 June 2015; pp. 1037–1045.
34. Toet, A. Data Title, 2014. Available online: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029/2 (accessed on 14 November 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.