Article

RAN: Infrared and Visible Image Fusion Network Based on Residual Attention Decomposition

Jia Yu, Gehao Lu * and Jie Zhang

School of Information Science and Engineering, Yunnan University, Kunming 650500, China; yujia@itc.ynu.edu.cn (J.Y.); zjlyy@mail.ynu.edu.cn (J.Z.)
* Correspondence: glu@ynu.edu.cn

Abstract: Infrared image and visible image fusion (IVIF) is a research direction that is currently attracting much attention in the field of image processing. The main goal is to obtain a fused image by reasonably fusing infrared images and visible images, while retaining the advantageous features of each source image. The research in this field aims to improve image quality, enhance target recognition ability, and broaden the application areas of image processing. To advance research in this area, we propose a breakthrough image fusion method based on the Residual Attention Network (RAN). By applying this innovative network to the task of image fusion, the mechanism of the residual attention network can better capture critical background and detail information in the images, significantly improving the quality and effectiveness of image fusion. Experimental results on public domain datasets show that our method performs excellently on multiple key metrics. For example, compared to existing methods, our method improves the standard deviation (SD) by 35.26%, spatial frequency (SF) by 109.85%, average gradient (AG) by 96.93%, and structural similarity (SSIM) by 23.47%. These significant improvements validate the superiority of our proposed residual attention network in the task of image fusion and open up new possibilities for enhancing the performance and adaptability of fusion networks.

Keywords: image fusion; residual attention module; dense block; infrared image; visible image

1. Introduction

Image fusion is a key task in the field of digital image processing, aiming to combine image information from different sensors or modalities into a more informative and comprehensive image. With the continuous development of multi-modal imaging technologies, such as infrared and visible image fusion (IVIF) and medical image fusion, image fusion technology has been widely used in military [1], medical [2], remote sensing [3] and facial recognition applications [4]. Visible images (VIS) are typically captured by traditional camera devices, such as CCD or CMOS cameras. These devices are sensitive to the visible spectrum, providing high resolution and color reproduction performance. The visible spectrum, perceivable by the human eye, has excellent spatial resolution, making it suitable for capturing detailed information, such as the color, shape and texture of objects. Infrared images (IR) are usually captured by infrared sensors that are sensitive to infrared radiation. Infrared wavelengths are primarily used to capture the thermal radiation information of targets. IR images perform better in low-light environments and under special conditions, such as at night, or in adverse weather, like smoke and haze. Since visible and infrared images capture different types of information, fusing them can result in a richer image that provides more comprehensive information about the scene. This fusion is particularly useful for tasks such as target detection [5], tracking [6] and scene understanding [7]. With the improvement and enhancement of deep learning, image theory and computer resources, image fusion technology has made significant progress. Existing
fusion methods can generally be categorized into two groups: traditional algorithms and deep learning-based methods. Specifically, representative traditional methods include those based on multi-scale transform [8], sparse representation [9], subspace [10], saliency [11] and full-variate [12]. Synthesizing existing deep learning-based image fusion methods, the fusion work primarily addresses three key problems in image fusion: feature extraction, feature fusion and reconstruction of fusion results. Depending on the differences in the adopted architecture, these deep learning based methods [13] can be roughly summarized as Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), and Auto-Encoder (AE)-based methods [14]. Although current deep learning-based image fusion algorithms can produce better results, there are still some challenges in the field of image fusion:

- Some current image fusion methods may fail to effectively integrate infrared and visible information, resulting in loss of information or distortion in the fused image. These issues can arise from deficiencies in the feature extraction stage, ineffective handling of redundant information throughout the fusion process, or insufficient processing in complex scenes.
- Significant differences exist between infrared (IR) images and visible (VIS) images, such as variations in spectral range and resolution. Existing image fusion methods may not adequately consider and address these modal differences, thus affecting the fusion quality.
- Some image fusion methods rely on relatively shallow neural network structures, failing to fully leverage deep learning capabilities. Consequently, they are unable to comprehensively extract image features and semantic information, leading to suboptimal fusion results.

To address the aforementioned issues, our study introduces a novel network named IVIF, based on residual attention mechanism decomposition for infrared and visible image fusion. This model employs a dual-branch encoder to extract modality-specific features, shared features and target information. The encoding and decoding processes are enhanced through the use of attention modules and dense block modules. The framework is shown in Figure 1. Our research has achieved significant breakthroughs in the field of image fusion. Specifically, our main contributions include the following:

- Based on the information available to us, this is the first time a residual attention module has been introduced into an auto-encoding network, marking an innovative exploration of the structure of auto-encoding networks. It is also the first time that the attention module is applied to the IVIF task, injecting more powerful perceptual and attentional mechanisms into our network.
- By embedding attention modules and dense blocks into the network and applying them to the trunk and branch networks, respectively, we can better extract both local and global information, further increasing the depth and hierarchical features of the network. This enables the network to more effectively focus on target information and background information in the input images, significantly enhancing the performance and quality of image fusion.
- We have improved the attention module and the dense block. The test results indicate that our method achieves significant improvements across multiple key metrics, demonstrating the superior performance of our proposed residual attention network in the task of image fusion.

Furthermore, these technological advancements have brought practical application value to multiple industries, specifically including, but not limited to, improving the clarity and accuracy of night vision surveillance, enhancing the diagnostic capabilities of medical imaging, and boosting the information extraction ability of remote sensing images in complex environments. These applications significantly enhance the efficiency and outcomes in areas such as public safety, medical diagnostics and environmental monitoring. In summary, our research has not only propelled the scientific development of image fusion
technology but has also made a positive contribution to practical applications and societal welfare in related fields. In the future, we anticipate that these advancements will find broader applications in fields that rely on high-quality image analysis, such as intelligent transportation, disaster management, autonomous driving and more.

Figure 1. The overall framework of RAN. (a) Training network structure; (b) Testing network structure; (c) Attention module; (d) DenseBlock structure.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes the network mechanism proposed in this paper. Section 4 reports the experiments and describes the results, and conclusions are drawn in Section 5.

2. Related Work

This section briefly reviews representative work on deep learning-based IVIF fusion methods, followed by a description of the attention module and dense block used in the residual attention network (RAN). Additionally, it provides an overview of classical work represented by traditional two-scale decomposition methods.

2.1. Image Fusion Algorithms

2.1.1. CNN

CNN-based approaches aim to construct an end-to-end network to address three key challenges in IVIF tasks: feature extraction, feature fusion and image reconstruction. In this fusion framework, the design of the loss function and the network structure becomes the focus of research and has a crucial impact on the final fusion result [15]. Hou et al. [16] emphasized preserving the local features of the source image in the fusion result by improving the loss function, in particular, by combining the saliency features for enhancement. Long et al. [17] proposed an unsupervised IVIF network based on a dense network of aggregated residuals. Tang et al. [18,19] presented SEAFusion, an innovative semantic-aware realtime image fusion network, and PIAFusion, a progressive image fusion network designed to be light-aware. Zhou et al. [20] introduced adaptive visual enhancement and multi-scale structural block decomposition modules, significantly improving visual quality while reducing artifacts and noise. Wang et al. [21] proposed contrastive learning and self-supervised mechanisms, which enhance the model’s ability to perceive fine-grained feature compensation. Luo et al. [22] proposed the RSTSFusion algorithm, which integrates the advantages of multi-scale Transformer and Sobel operators. This approach excels in processing global information and extracting detailed features.
2.1.2. GAN

The GAN network consists of a generator and a discriminator, where the discriminator is used to detect false fusion images generated by the generator. By recognizing the fake images, the generator parameters can be adjusted until the discriminator is able to recognize the real image, which improves the accuracy of the discriminator and generates a higher-quality fusion image [23]. Therefore, GAN networks have been widely used in IVIF tasks. Ma et al. [24] first introduced an adversarial game between the generator and the discriminator. Fu et al. [25] designed a dense block and GAN-based approach to insert visible images directly into each of the entire network’s layer to fully preserve the detailed texture. Ma et al. [26] used the primary-secondary idea in processing infrared images; by considering both primary and secondary visual elements, they were able to extract key information from the image more efficiently and better capture useful features. Yang et al. [27] chose to use texture conditioning, which enhances or highlights specific aspects of the textures in an image by adjusting those textures. Zhou Yinan et al. [28] employed saliency algorithms, which utilize algorithms in computer vision regarding salient regions of an image in order to identify striking and salient portions of an image, to focus on them and to extract areas with important information. Li et al. [29] added a multi-scale attention mechanism to the generator and discriminator of the GAN so that the fusion network can focus more on the key regions of the source image to regenerate the fused image.

2.1.3. AE

Self-encoders usually consist of an encoder and a decoder. Self-coding image fusion network methods are usually divided into two steps. First, the self-encoder is pretrained on a large dataset to learn the feature representation of the source image. The encoder is responsible for encoding the source image into potential feature representations, and the decoder reduces these features to the original image. Then, in the fusion stage, we design a fusion network that utilizes a pretrained encoder to extract features from the source image. These features can be combined by either a manually designed fusion strategy or automatic network stitching to produce the final fused image. Wang et al. [30] innovatively introduced the Residual Density Block (RDB) as a key component of the network. On the other hand, Xu et al. [31] provided an effective approach for the IVIF task by employing a network consisting of two pairs of encoder-decoders to extract and decompose the feature maps, respectively. Wang et al. [32] proposed three different specifications of ZOF, LP-normalized based on dense connectivity attention models, which contribute a unique perspective to image fusion by highlighting and fusing depth features in spatial and channel dimensions. Li et al. [33] designed an IVIF architecture based on nest network and spatial/channel attention mechanism models, which provides a new way of thinking about the structure and performance of fusion networks. In addition, Ren et al. [34] proposed an image fusion network based on variational self-coding, which is divided into two parts: an image fusion network and an infrared feature compensation network. Li et al. [35] incorporated a Cross-Attention Mechanism (CAM) into the encoder network, significantly enhancing the ability to capture complementary information from multi-modal features, thereby improving the quality of the fused images.

2.2. Residual Attention Module and Dense Block

A residual attention network was initially proposed for image classification tasks by Wang et al. [36]. The structure of the attention module is shown in Figure 1c. The residual attention mechanism combines residual learning with attention modules, which is significantly different from traditional attention mechanisms. This combination allows for the optimization of feature maps at multiple levels, thereby enhancing both local and global feature extraction. Unlike traditional attention mechanisms that usually focus on a single type of attention, this mechanism captures hybrid attention, including spatial, channel and mixed attention types. The network employs an innovative approach by adding attention
modules at different levels of feature maps, thus significantly improving the network’s feature map extraction capability. Given that direct stacking of attention modules may lead to the problem of gradient vanishing in deeper layers of the network, the residual attention model employs residual connectivity to ensure that attention modules at different layers are fully learned. This strategy not only maintains accuracy but also significantly reduces the computational burden, to about 69% of that of large networks such as ResNet, making residual attention networks an excellent choice for performance in image processing tasks.

The dense block, proposed by Huang et al. [37], is illustrated in Figure 1d. It primarily addresses the issue of gradient vanishing in traditional convolutional networks. By establishing the connection relationship between different layers, the dense block maximizes the utilization of feature maps, better meeting the demands for feature extraction in image processing tasks. In architectures that combine self-encoding networks with dense blocks, the outputs of each layer are closely connected, effectively promoting the full retention of deep features. This design helps mitigate the gradient vanishing problem and ensures that all salient features are fully utilized in various processing strategies. Encoding with dense blocks maximizes the retention of deep features and ensures that all salient features are fully utilized in fusion strategies.

2.3. Two-Scale Decomposition

In the field of image fusion, two-scale decomposition is a technique that categorizes image information into high-frequency and low-frequency layers, which independently process the image’s detail and background information. This method facilitates the effective fusion of images from diverse sources, enhancing the overall quality of the fused images. This approach is extensively applied across various image fusion algorithms, significantly improving the results. Zhao et al. [38] introduced the DIDFuse method, which employs an encoder to split the image into two distinct feature maps: one for background containing low-frequency information and another for details containing high-frequency information. This approach uniquely integrates a mechanism that utilizes the $L_1$ norm to calculate attention for weight distribution, effectively enhancing the fusion process. Additionally, a decoder is used to reconstruct the original image, ensuring that the quality is preserved. Jiang et al. [39] proposed two-scale technology, integrating image morphology and fuzzy set theory, to partition the source image into a base layer and a detail layer, thereby optimizing the fusion process. Wang et al. [32] developed an end-to-end IVIF network that incorporates Residual Density Blocks (RDB), ensuring the effective extraction of deep features. This method separates the image into background and detail layers, which are subsequently combined through additive fusion, thereby enhancing information retention and feature transfer capabilities. Wu et al. [40] proposed a method using patch-based low-rank representation for two-scale decomposition, effectively distinguishing between the image’s base layer and detail layer. Additionally, Cheng et al. [41] utilized scale decomposition to separate the image data into differential features and common features. These approaches have shown excellent performance and results in the study of two-scale decomposition. These studies not only improve the efficiency and outcomes of image fusion technologies but also offer new perspectives and tools in the field of image processing, especially in the application potential for multi-source image data.

3. Method

In this section, we present the workflow of our proposed algorithm and network structure and the detailed structure of each module. The detailed workflow is shown in Figure 1.

3.1. Motivation

As described in Section 2.2, dual-scale decomposition involves partitioning the input image into two components: a background image that describes the overall structure and large-scale variations, and a detailed image that captures local details and small-scale
changes. Currently, many image decomposition algorithms rely on specific a priori knowledge in performing this process and employ manually designed decomposition methods, such as filters [42] or optimization-based methods [43]. These methods essentially act as feature extractors, converting the source image from the spatial domain to the feature domain. Despite the tremendous advantages of deep neural networks for data-driven feature extraction, however, as of now, there is no deep learning-based image decomposition algorithm that can effectively handle the work of fusing infrared and visible images. DenseFuse [44] introduced self-coding networks for image decomposition and reconstruction for the first time, but its approach simply uses an encoder to generate feature maps for IR images and VIS images without further in-depth applications. In contrast, DIDFuse employs a separate strategy for decomposing IR images and VIS images into background and detail feature maps. However, the network structure of this method is relatively simple and it is difficult to effectively utilize the feature information extracted by the convolutional neural network, which results in poor clarity of the generated fused images. Therefore, we propose a novel image decomposition network that achieves more complex and in-depth feature extraction by introducing an attention module and dense block into the encoder (as shown in Figure 1c,d). This network utilizes the encoder to perform two-branch decomposition and extracts different types of information and subsequently recovers the original image through a decoder. This innovative approach aims to bridge the gaps of previous methods to provide clearer, more informative fused images to advance the IVIF mission.

3.2. Proposed Network Structure

Our network architecture consists of two parts: an encoder and a decoder, as shown in Figure 1. First, we define some notations for clarity of presentation. The input paired IR and VIS images are notated as $I \in \mathbb{R}^{H \times W}, V \in \mathbb{R}^{H \times W}$. The output fused image is $F \in \mathbb{R}^{H \times W}$. The background and detail feature maps obtained by the encoder are, respectively, $I_1, I_2, V_1, V_2$. The source image can be decomposed into background and detail feature maps by the residual attention and dense block module in the encoder, respectively, $I_1, I_2, V_1, V_2$. Then, channel splice $I_1$ and $I_2, V_1$ and $V_2$. Finally, the spliced feature map is then passed through the residual attention modular network for inverse reconstruction decoding operation to output the fused image $F$.

3.2.1. Encoder Network

In Figure 1a, the structure of the encoder network consists of several key components, including a convolutional layer, three residual blocks, an attention module and a dense block. Each residual block consists of three convolutional layers, named Conv1, Conv2 and Conv3. In addition, each residual block contains three Relu layers, named Relu1, Relu2 and Relu3, as well as three normalization layers, named Bn1, Bn2 and Bn3, respectively. This hierarchical structure helps the network to understand the features of the image at a deeper level and improve the abstract expression of features. The attention module consists of residual blocks, pooling layers, up-sampling and down-sampling, which introduces local and global perception capabilities to the network. This allows the network to focus on key parts of the image in a targeted manner to better capture important information. This combination of local and global attention allows the network to decompose the background and detail features of the source image more effectively, providing strong support for subsequent image fusion tasks. The introduction of dense blocks, as shown in Figure 1c, further enhances the network’s ability to extract features. Tables 1 and 2 detail the configuration information of the networks, where the kernel size of the convolutional layer is generally $1 \times 1$, except for Conv2, which has a kernel size of $3 \times 3$. It is worth emphasizing that in view of the fact that IR and VIS images have different characteristics, although the same network structure is used for both feature extraction networks, their parameters are trained separately and independently. This design ensures that the networks are more flexible in learning and extracting features when dealing with different types of images.
Table 1. The network configuration of convolutional layers.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Kernel Size</th>
<th>Stride</th>
<th>Padding</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>False</td>
</tr>
<tr>
<td>Conv2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>False</td>
</tr>
<tr>
<td>Conv3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>False</td>
</tr>
</tbody>
</table>

Table 2. The network configuration of normalization layers.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Parameters</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bn1</td>
<td>128</td>
<td>Relu1</td>
</tr>
<tr>
<td>Bn2</td>
<td>64</td>
<td>Relu2</td>
</tr>
<tr>
<td>Bn3</td>
<td>64</td>
<td>Relu3</td>
</tr>
</tbody>
</table>

3.2.2. Decoder Network

The decoder network structure is shown in Figure 1a and consists of two convolutional layers, a residual block and an attention module. This combination is designed to increase the depth and complexity of the image features learned by the network. The stacking of the two convolutional layers allows the network to gradually extract higher-level abstract features, thus enhancing the representation of the image content. The introduction of the residual block helps the network to better capture and retain the information of the original image, improving image restoration. The attention module empowers the network with local and global perception, enabling it to focus on key parts of the image in a targeted manner to better capture important information. This combination of local and global attention properties allows the network to restore image details and structures more effectively, providing superior performance for image fusion tasks. In terms of dealing with long-range dependencies, by superimposing two convolutional layers, the network is able to learn a wider range of contextual information, which improves the understanding and restoration of global structures. This design not only helps to efficiently deal with features at different scales and levels in the image, but also provides strong support for reconstructing out the fused image in a retrograde manner. Therefore, the structure of the decoder network not only increases the depth and complexity of the network, but also provides multi-level, global and local information for the reduction process of image fusion, thus improving the overall performance.

3.2.3. Loss Function and Network Training

During the training process, our main goal is to train a self-encoder network that can adequately extract features, decompose the image and reconstruct the image completely. The entire training process is illustrated in Figure 1a. In the stage of image decomposition, the encoder generates a background feature map to extract common features from the source images, while also producing a detail feature map to capture distinctive features from the IR and VIS images. Therefore, our goal is to minimize the disparity between background feature maps to ensure the extraction of common features, while also minimizing the disparity between detail feature maps to preserve the unique characteristics of infrared and visible light images. To achieve this goal, we referenced the loss function defined by Zhao et al. [38] in DIDFuse and introduced it into the image decomposition process:

$$
\ell_1 = \Phi(\|V_1 - I_1\|_2^2) - \lambda_1 \Phi(\|V_2 - I_2\|_2^2),
$$

where $V_1$, $V_2$ are the background and detail feature map of the visible image and $I_1$, $I_2$ are the background and detail feature map of the infrared image. $\Phi(\cdot)$ is a hyperbolic tangent function used to restrict the gap to the interval $(-1, 1)$. In order to minimize the pixel-level difference between the reconstructed image and the original image to ensure
that the structural and detail information of the image is preserved as much as possible during the image reconstruction process, the loss function is designed as follows:

\[ \ell_2 = \lambda_2 f(I, I_F) + \lambda_3 f(V, V_F) + \lambda_4 \| \Delta V - \Delta V_F \|_1, \]  

(2)

where \( I \) and \( I_F \) and \( V \) and \( V_F \) denote the input and reconstructed images for the infrared and visible images, respectively. \( \Delta \) denotes the gradient operator, and

\[
f(O, \hat{O}) = \|O - \hat{O}\|_2^2 + \lambda \ell\text{SSIM}(O, \hat{O}),
\]  

(3)

where \( O \) and \( \hat{O} \) denote the above input image and reconstructed image, \( \lambda \) is a hyperparameter, and SSIM (Structural Similarity Index) is a metric used to measure the similarity between two images. Then, \( \ell_{\text{SSIM}} \) can be described as

\[
\ell_{\text{SSIM}}(O, \hat{O}) = \frac{1 - \text{SSIM}(O - \hat{O})}{2}.
\]  

(4)

Since feature extraction and image reconstruction are two crucial stages in the image fusion process, we need to consider their differences in terms of feature retention and reconstruction quality, especially in terms of texture consistency and structure preservation, and merging these two stages enables a comprehensive assessment of the fusion quality of an image. Therefore, Equations (1) and (2) are combined and summed into a total loss function:

\[
\ell = \ell_1 + \ell_2,
\]  

(5)

where \( \lambda, \lambda_1, \lambda_2, \lambda_3 \) and \( \lambda_4 \) are tuning parameters.

### 3.3. Fusion Strategy

In the previous phases, we have determined the structure of the encoder and decoder and defined the corresponding loss functions. Once the training is complete, we obtain a network structure containing only the encoder and decoder, as shown in Figure 1a. However, unlike the training phase, we introduce a fusion layer in the testing phase. The role of this layer is to fuse the background and detail feature maps, which is formulated as shown below:

\[
F_1 = \text{fusion}(I_1, V_1),
\]  

(6)

\[
F_2 = \text{fusion}(I_2, V_2),
\]  

(7)

\( F_1 \) and \( F_2 \) denote the fused background and detail feature maps, respectively. In this paper, we consider three convergence strategies:

- Summation method: \( F_1 = I_1 \oplus V_1, F_2 = I_2 \oplus V_2 \), where \( \oplus \) denotes element-by-element summation.

- Weighted average method: \( F_1 = \gamma_1 I_1 \oplus \gamma_2 V_1, F_2 = \gamma_3 I_2 \oplus \gamma_4 V_2 \), where \( \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1 \) and \( \gamma_i \) (\( i = 1,2,3,4 \)); the default values are all set to 0.5.

- \( L_1 \)-norm method: Referring to [44], we use \( L_1 \)-norm as the activity metric and combine it with the softmax operator. With \( \| I_i(\hat{x}, \hat{y}) \|_1 \) and \( \| D_i(\hat{x}, \hat{y}) \|_1 \) (\( i = 1,2 \)), we can obtain the background and detail feature maps, where \( I_1, I_2, V_1 \) and \( V_2 \) represent \( B_1, D_1, B_2, D_2 \), respectively, and \( (\hat{x}, \hat{y}) \) denote the feature maps and the fused feature maps, respectively, with corresponding coordinates. Then, the summed weights can be obtained by:

\[
\eta_1^{I}(\hat{x}, \hat{y}) = \frac{\psi(\| I_1(\hat{x}, \hat{y}) \|_1)}{\sum_{i=1}^{2} \psi(\| I_i(\hat{x}, \hat{y}) \|_1)},
\]  

(8)

\[
\eta_1^{V}(\hat{x}, \hat{y}) = \frac{\psi(\| V_1(\hat{x}, \hat{y}) \|_1)}{\sum_{i=1}^{2} \psi(\| V_i(\hat{x}, \hat{y}) \|_1)}.
\]  

(9)
where $\psi(\cdot)$ is a $3 \times 3$ box blurr, so we have:

$$F_1 = (\eta_1 I \otimes B_1) \oplus (\eta_2 I \otimes B_2), \tag{10}$$

$$F_2 = (\eta_1 V \otimes V_1) \oplus (\eta_2 V \otimes D_2), \tag{11}$$

where $\otimes$ denotes element-by-element multiplication.

4. Experimental

This section aims to validate the experimental effectiveness of our proposed fusion method. First, we detail the experimental setup for the training and testing phases. Then, through a series of ablation experiments, we delve into the impact of different elements in the fusion network. Finally, in order to objectively evaluate the performance of our proposed fusion framework, we perform a qualitative comparison with other state-of-the-art algorithms, including DIDFuse [38], PIAfusion [19], RFN-Nest [45], NestFuse [33] and DATFuse [46]. All experiments were conducted using a GeForce RTX 3060 GPU and an Intel(R) Core(TM) i7-12700H processor (2.30 GHz) from the 12th generation series, provided by the Intel Corporation (Santa Clara, CA, USA). The software environment comprised Pytorch version 1.8.1, paired with CUDA version 11, developed by the NVIDIA Corporation (Santa Clara, CA, USA).

**Fusion Performance Evaluation.** To evaluate the quality of image fusion, we employ eight key metrics: entropy (En), standard deviation (SD), spatial frequency (SF), average gradient (AG), visual information fidelity (VIF), correlation coefficient (CC), edge preservation measure (Qabf) and structural similarity index (SSIM). Entropy measures the richness of information in an image, while the standard deviation reflects the dispersion of gray-scale values, indicating contrast levels. Spatial frequency describes the frequency of gray-scale changes, highlighting the amount of detail in the image. Average gradient quantifies the intensity of gray-scale variations, serving as an indicator of image clarity. Visual information fidelity assesses how faithfully the fused image retains visual information from the original images. The correlation coefficient evaluates the statistical correlation between the fused and original images, indicating similarity. The edge preservation measure assesses how well the fused image maintains edge information, and the structural similarity index measures the likeness in structural information between the fused and source images. More details for these metrics can be seen in [14]. Higher values in these metrics typically indicate better fusion results as they represent successful retention of the original images’ characteristics and details.

**Datasets and Preprocessing.** Our experiments cover three datasets, namely, FLIR [47], TNO [48], NIR [49] and 21 pairs of infrared and visible images (select 21 pairs from TNO). These datasets are divided into training, validation and test sets in the experiments. For the FLIR dataset, we randomly selected 180 pairs of images from it to be used as training samples. Prior to training, all images were converted to gray-scale and centered and cropped to retain a region of $128 \times 128$ pixels. This series of datasets and preprocessing steps provides a rich and diverse sample of images for our experiments to evaluate the performance of the proposed method in different scenarios.

**Hyperparameter Settings.** The setting of hyperparameters is crucial for the training and performance of the network. In our experiments, we empirically set specific values for the tuning parameters in the loss function as $\lambda_1 = 0.05$, $\lambda_2 = 2$, $\lambda_3 = 2$, $\lambda_4 = 10$ and $\lambda = 5$. In the training phase, we chose the Adam optimizer for network training up to 120 epochs. To better process the data, we divided the training data into batches, each containing 24 samples. The initial value of the learning rate was set to $10^{-3}$, which was reduced by a factor of 10 for every 40 epochs as the training progressed, which helped to tune the model parameters more finely at the later stages of training to improve convergence. To monitor the progress of training and network performance, we plotted the loss curve versus epochs, as shown in Figure 2.
After 120 epochs of training, all the loss curves stabilize, indicating that the network has achieved better convergence under this training condition. These careful choices of hyperparameter settings and training strategies help ensure that the model performs well in the final testing phase.

### 4.1. Experiments on Fusion Strategy

As mentioned in Sections 3 and 4, fusion strategies play a crucial role in our model. In this experimental session, we choose three classical manual fusion strategies commonly used in current fusion networks, i.e., “add”, “average”, and “\(L_1\) norm”, and conduct detailed experiments. In order to clearly demonstrate the effects of different fusion strategies, we applied these three fusion strategies on the NIR dataset and present an example of the fusion results in Figure 3.

Subsequently, a detailed assessment of the fusion results for the NIR dataset was performed, including a comprehensive consideration of eight quality metrics. The specific evaluation results are organized and summarized in Table 3, providing a comprehensive comparison of each fusion strategy. In the table, we can clearly see that adopting the summation strategy obtains higher values in various indexes, especially in EN, SD, SF, AG, VIF and Qabf. In view of these results, we chose the summation strategy in the subsequent experiments to obtain better fusion performance.

**Table 3.** The results of different integration strategies on eight indicators for the NIR dataset. Boldface and under-line show the best and second-best values, respectively.

<table>
<thead>
<tr>
<th>Fusion Strategy</th>
<th>EN</th>
<th>SD</th>
<th>SF</th>
<th>AG</th>
<th>VIF</th>
<th>CC</th>
<th>Qabf</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>7.32</td>
<td>59.11</td>
<td>26.42</td>
<td>8.98</td>
<td>0.73</td>
<td>0.76</td>
<td>0.49</td>
<td>1.21</td>
</tr>
<tr>
<td>Average</td>
<td>6.85</td>
<td>37.88</td>
<td>12.74</td>
<td>4.66</td>
<td>0.71</td>
<td>0.77</td>
<td>0.47</td>
<td>1.25</td>
</tr>
<tr>
<td>L1-norm</td>
<td>6.89</td>
<td>38.76</td>
<td>13.07</td>
<td>4.80</td>
<td>0.72</td>
<td>0.77</td>
<td>0.48</td>
<td>1.26</td>
</tr>
</tbody>
</table>

![Figure 2. The loss curves over 120 epochs.](image)
4.2. Ablation Studies

In this section, our objective is to validate the impact of the Densenet modules and residual attention mechanisms on image fusion performance. For this purpose, two distinct network configurations were systematically compared.

**Densenet.** To explore the efficacy of the Densenet module, we constructed a benchmark network devoid of Densenet components, exclusively utilizing residual attention mechanisms instead. The architecture of this network, which includes both training and testing phases, is depicted in Figure 4.

![Figure 4. The AE without Densenet. (a) Training network structure; (b) Testing network structure.](image)

We conducted experiments using this benchmark network against a proposed network that integrates the Densenet module to assess performance enhancements. The fusion results of 21 pairs of infrared and visible images were evaluated using eight qualitative metrics and are summarized in Table 4.

Table 4. The results of network structure with and without Densenet on eight indicators for the 21 pairs of infrared and visible images. Boldface is used to indicate the best values.

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>SD</th>
<th>SF</th>
<th>AG</th>
<th>VIF</th>
<th>CC</th>
<th>Qabf</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-densenet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proposed</td>
<td>6.61</td>
<td>34.18</td>
<td>9.42</td>
<td>3.52</td>
<td>0.58</td>
<td>0.48</td>
<td>0.42</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>6.87</td>
<td>40.81</td>
<td>12.29</td>
<td>4.32</td>
<td>0.60</td>
<td>0.48</td>
<td>0.44</td>
<td>0.98</td>
</tr>
</tbody>
</table>

The outcomes demonstrate that the experimental network with the Densenet module significantly improved performance across these metrics, thereby confirming the effectiveness of the Densenet branch in enhancing image fusion capabilities.
Residual Attention. In another set of experiments, we developed a network that lacks the residual attention module, consisting entirely of conventional convolutional layers and a dual-branch structure with the Densenet module. This configuration is illustrated in Figure 5, encompassing both training and testing stages. We compared this network with a proposed network that incorporates the residual attention module.

![Figure 5](image_url) The AE without residual attention modules. (a) Training network structure. (b) Testing network structure.

The fusion results on the TNO dataset were assessed using eight qualitative metrics and are compiled in Table 5.

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>SD</th>
<th>SF</th>
<th>AG</th>
<th>VIF</th>
<th>CC</th>
<th>Qabf</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-Residual</td>
<td>6.79</td>
<td>42.09</td>
<td>10.71</td>
<td>3.94</td>
<td>0.60</td>
<td>0.49</td>
<td>0.38</td>
<td>0.87</td>
</tr>
<tr>
<td>Proposed</td>
<td>7.01</td>
<td>43.70</td>
<td>12.59</td>
<td>4.56</td>
<td>0.63</td>
<td>0.50</td>
<td>0.44</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Comparative analysis revealed that networks integrating the residual attention module also exhibited significant improvements in performance metrics, further substantiating the efficacy of residual attention mechanisms in boosting image fusion performance.

In summary, these comparative experiments not only clarify the individual contributions of the Densenet and residual attention modules but also showcase their potential in handling complex image fusion tasks. These findings provide crucial experimental evidence and theoretical support for further research and application of image fusion technology.

4.3. Comparison with Other Models

To rigorously evaluate the differences in image fusion performance between our proposed method and existing state-of-the-art algorithms, we selected five representative fusion methods for comparison. These include NestFuse, DIDFuse, RFN-Nest, PIAFusion and DATFuse.

Qualitative comparison. To comprehensively evaluate the visual effects of the image fusion results, we carefully selected typical infrared (IR) and visible light (VIS) images from the TNO and NIR datasets, which include rich target and background information. By processing these images through various fusion models, we can demonstrate the effectiveness and performance of each model in handling complex scenes. The specific fusion results are displayed in Figure 6.
Infrared image  
Visible image  
NestFuse  
RFN-Nest  
DIDFuse  
PIAFusion  
DATFuse

Figure 6. The experimental results of different methods on the TNO and NIR datasets.

By comparing and observing the images obtained from NestFuse, DIDFuse, RFN-Nest, PIAFusion, DATFuse and our proposed fusion method, it can be clearly seen that the new method better preserves and highlights the target details in the fused images and makes the target contours clearer, and at the same time, effectively fuses the information of IR and VIS images. Compared to other algorithms, our method presents more natural and realistic colors and has higher fidelity in terms of target boundaries and textures. This visual performance highlights the superiority of our proposed fusion method in terms of image quality and information retention, and provides a more competitive and practical solution for the fusion of infrared and visible images.

Quantitative comparison. We used two datasets, an NIR dataset and a TNO dataset. We performed a comprehensive comparison of five existing fusion methods and our proposed fusion framework using eight quality metrics. Tables 6 and 7 list the average values of these metrics across all fused images.

Table 6. The average values of eight quality metrics for NIR fused images. Boldface and under-line show the best and second-best values, respectively.

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>SD</th>
<th>SF</th>
<th>AG</th>
<th>VIF</th>
<th>CC</th>
<th>Qabf</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NestFuse</td>
<td>7.33</td>
<td>49.34</td>
<td>18.38</td>
<td>6.66</td>
<td>0.71</td>
<td>0.73</td>
<td>0.59</td>
<td>1.18</td>
</tr>
<tr>
<td>DIDFuse</td>
<td>7.00</td>
<td>46.88</td>
<td>11.28</td>
<td>4.27</td>
<td>0.62</td>
<td>0.49</td>
<td>0.40</td>
<td>0.86</td>
</tr>
<tr>
<td>RFN-Nest</td>
<td>7.31</td>
<td>50.47</td>
<td>9.73</td>
<td>4.17</td>
<td>0.65</td>
<td>0.75</td>
<td>0.40</td>
<td>1.17</td>
</tr>
<tr>
<td>PIAFusion</td>
<td>7.28</td>
<td>51.24</td>
<td>20.38</td>
<td>7.65</td>
<td>0.72</td>
<td>0.74</td>
<td>0.58</td>
<td>1.22</td>
</tr>
<tr>
<td>DATFuse</td>
<td>6.86</td>
<td>35.84</td>
<td>19.31</td>
<td>6.58</td>
<td>0.72</td>
<td>0.75</td>
<td>0.53</td>
<td>1.20</td>
</tr>
<tr>
<td>Proposed</td>
<td>7.32</td>
<td>59.11</td>
<td>26.42</td>
<td>8.98</td>
<td>0.73</td>
<td>0.76</td>
<td>0.49</td>
<td>1.21</td>
</tr>
</tbody>
</table>

By analyzing Tables 6 and 7, it can be seen that compared to other methods, the fusion framework proposed in this paper achieves five best values and three best values in the NIR datasets and TNO dataset, respectively. Additionally, out of a total of 16 metrics, it
obtained eight best values, ranking first overall. The study found that while other methods may perform better on certain metrics, the proposed fusion method maintains strong competitiveness across all metrics, demonstrating a more comprehensive performance.

Table 7. The average of eight quality metrics for TNO fused images. Boldface and under-line show the best and second-best values, respectively.

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>SD</th>
<th>SF</th>
<th>AG</th>
<th>VIF</th>
<th>CC</th>
<th>Qabf</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NestFuse</td>
<td>7.02</td>
<td>40.60</td>
<td>9.96</td>
<td>3.82</td>
<td>0.83</td>
<td>0.48</td>
<td>0.52</td>
<td>0.89</td>
</tr>
<tr>
<td>DIDFuse</td>
<td>7.00</td>
<td>46.88</td>
<td>11.28</td>
<td>4.27</td>
<td>0.62</td>
<td>0.49</td>
<td>0.40</td>
<td>0.86</td>
</tr>
<tr>
<td>RFN-Nest</td>
<td>6.99</td>
<td>37.24</td>
<td>5.87</td>
<td>2.67</td>
<td>0.55</td>
<td>0.51</td>
<td>0.33</td>
<td>0.81</td>
</tr>
<tr>
<td>PLAFusion</td>
<td>6.88</td>
<td>39.40</td>
<td>11.31</td>
<td>4.40</td>
<td>0.69</td>
<td>0.44</td>
<td>0.56</td>
<td>0.97</td>
</tr>
<tr>
<td>DATFuse</td>
<td>6.44</td>
<td>27.53</td>
<td>9.56</td>
<td>3.57</td>
<td>0.68</td>
<td>0.46</td>
<td>0.50</td>
<td>0.93</td>
</tr>
<tr>
<td>Proposed</td>
<td>7.01</td>
<td>43.70</td>
<td>12.59</td>
<td>4.56</td>
<td>0.63</td>
<td>0.50</td>
<td>0.44</td>
<td>0.98</td>
</tr>
</tbody>
</table>

The fusion model proposed in this paper performs best across all datasets, further validating its superiority. This indicates that our fusion network not only achieves excellent fusion performance but also significantly enhances the quality of the image content. The generated images are not only clearer but also exhibit more realistic visual effects, showcasing a higher level of visual information and image realism.

5. Conclusions

To address the shortcomings of IVIF and existing fusion methods in decomposing image features, this paper proposes a self-encoder fusion network based on a residual attention module. By incorporating the residual attention module and dense blocks into the encoder, the network efficiently extracts background and detail feature maps from the input images. Subsequently, these feature maps are inversely reconstructed into the final fusion image using attention modules. In tests on NIR and TNO public datasets, the proposed method produces fusion images with high clarity, prominent targets and distinct contours, aligning well with human visual perception. Compared to other representative fusion methods, our approach demonstrates improvements in fusion quality evaluation metrics, including energy (EN), standard deviation (SD), spatial frequency (SF), average gradient (AG), visual information fidelity (VIF), correlation coefficient (CC), Qabf and structural similarity index (SSIM), particularly excelling in the clarity of the fused images. Additionally, the proposed method exhibits excellent fusion performance in challenging scenarios such as blur, occlusion, backlighting and smoke, highlighting its high practicality.

Despite the simplicity and effectiveness demonstrated by the proposed residual attention network (RAN) in image fusion tasks, it still exhibits certain limitations in its fusion strategy. Currently, the RAN network employs a traditional summation strategy, which proves insufficient when addressing the complex demands of image fusion. Therefore, a potential research direction is to introduce advanced attention mechanisms or Transformer mechanisms. By adopting adaptive learning strategies, we can dynamically adjust the feature fusion weights, thereby enhancing the role of the fusion module within the image fusion network. Through these efforts, we aim to achieve more significant advancements in the field of image fusion and promote the widespread application of this technology.

Author Contributions: Conceptualization, J.Y.; methodology, J.Y.; investigation, J.Y.; data curation, J.Z.; Resources, G.L.; Writing—original draft, J.Y.; supervision, G.L.; project administration, G.L.; funding acquisition, G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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


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.