electronics

MDPI

Article

A Visible and Synthetic Aperture Radar Image Fusion Algorithm Based on a Transformer and a Convolutional Neural Network

Liushun Hu, Shaojing Su, Zhen Zuo, Junyu Wei, Siyang Huang *, Zongqing Zhao, Xiaozhong Tong and Shudong Yuan

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; huliushun18@nudt.edu.cn (L.H.); ssjing@nudt.edu.cn (S.S.); z.zuo@nudt.edu.cn (Z.Z.); yujy@nudt.edu.cn (J.W.); zhaozongqing17@nudt.edu.cn (Z.Z.); tongxiaozhong@nudt.edu.cn (X.T.); yuanshudong21@nudt.edu.cn (S.Y.)

* Correspondence: huangsy1102@nudt.edu.cn

Abstract: For visible and Synthetic Aperture Radar (SAR) image fusion, this paper proposes a visible and SAR image fusion algorithm based on a Transformer and a Convolutional Neural Network (CNN). Firstly, in this paper, the Restormer Block is used to extract cross-modal shallow features. Then, we introduce an improved Transformer–CNN Feature Extractor (TCFE) with a two-branch residual structure. This includes a Transformer branch that introduces the Lite Transformer (LT) and DropKey for extracting global features and a CNN branch that introduces the Convolutional Block Attention Module (CBAM) for extracting local features. Finally, the fused image is output based on global features extracted by the Transformer branch and local features extracted by the CNN branch. The experiments show that the algorithm proposed in this paper can effectively achieve the extraction and fusion of global and local features of visible and SAR images, so that high-quality visible and SAR fusion images can be obtained.

Keywords: dual branch; feature extraction; image fusion; SAR images; visible images

Citation: Hu, L.; Su, S.; Zuo, Z.; Wei, J.; Huang, S.; Zhao, Z.; Tong, X.; Yuan, S. A Visible and Synthetic Aperture Radar Image Fusion Algorithm Based on a Transformer and a Convolutional Neural Network. *Electronics* **2024**, *13*, 2365. https://doi.org/10.3390/electronics13122365

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 6 May 2024
Revised: 9 June 2024
Accepted: 13 June 2024
Published: 17 June 2024

1. Introduction

In recent years, with the continuous development of remote sensing technology, SAR imaging, as an important imaging technology, has become a popular research field. SAR is a microwave remote sensing system with the characteristics of all-weather, all-day, certain penetration, etc., and can provide SAR images with rich structural information. However, SAR images are seriously contaminated by noise, resulting in low signal-to-noise ratios, and SAR image interpretation is more difficult. In comparison, traditional visible-light sensors receive rich, high-resolution, multi-spectral information from ground objects, but visible-light sensor imaging is easily affected by weather and other factors. On the other hand, although visible-light imaging technology covers a wide range of wavelengths from 400 to 700 nm, its principle of operation relies on the passive reception of spectral information reflected from features and, therefore, has limitations in filtering or highlighting spectral information in specific wavelength bands. In contrast, SAR is capable of actively emitting electromagnetic waves of a specific wavelength band and receiving their reflected signals, thereby accurately capturing and extracting information reflected from electromagnetic waves of a specific wavelength band, which endows SAR with greater flexibility and accuracy in specific tasks. Therefore, the organic fusion of visible-light and SAR images with complementary advantages can significantly enrich the useful information of images, which is of great significance in military reconnaissance, agricultural planning, target extraction, and other image processing work.

Currently, mainstream visible and SAR image fusion methods can be broadly categorized into two types, including traditional image fusion methods and deep learning-based image fusion methods. Traditional image fusion algorithms mainly include Laplace

Pyramid (LP) [1–3], Shear Wave (SW) [4–6], Discrete Wavelet Transform (DWT) [7–9], Non-Subsampled Shearlet Transform (NSST) [10–12], Sparse Representation (SR) [13–16], and other methods. However, traditional methods use complex transformations and manual rules, thus limiting the real-time performance of the algorithms and the integration of semantic information, which restricts their application in advanced visual tasks.

Deep learning-based image fusion methods include frameworks such as the auto-encoder (AE) [17,18], Convolutional Neural Networks [19,20], and Generative Adversarial Networks (GANs) [21–23]. These frameworks can automatically and efficiently learn feature information from visible and SAR images, resulting in highly accurate fusion results.

In image fusion methods based on deep learning, the auto-encoder is a commonly used fusion model. Its structure mainly consists of three parts: encoder, fusion decision, and decoder. The encoder is primarily used to encode the source images into low-dimensional representations in the latent space, capturing the key features of the images. The decoder reconstructs the original images by receiving the latent representations generated by the encoder. During the training process, an appropriate loss function is designed so that the decoder can reconstruct the input images as accurately as possible. After training, the encoder can encode data from different modalities into low-dimensional representations in the latent space, which are then fused according to the designed fusion method. The fused encoding is inputted into the decoder for reconstruction. Image fusion methods based on auto-encoders do not require manual design for feature extraction. They can effectively learn key information from the image and achieve fusion in an end-to-end framework, greatly simplifying the fusion process.

Among the many AE fusion frameworks, the auto-encoder approach based on CNN feature extraction and reconstruction has been proven to be one of the most effective methods. The three algorithmic processes shown in Figure 1 are currently the most commonly used methods for this approach. The processes shown in Figure 1a,b are based on a shared encoder algorithm process, while the one in Figure 1c is based on a private encoder method. However, these methods currently have some problems and shortcomings. Firstly, CNNs are convolution-based neural networks with inductive biases and translation invariance characteristics, which, while improving the efficiency of graphic feature computations, cause a loss of the receptive field, leading to weak global feature mining capabilities and difficulty in extracting global information to obtain high-quality fused images [24]; secondly, forward propagation in the fusion network may lead to the loss of some important feature information; and lastly, the method based on shared encoders in the figure cannot differentiate features from different modalities, while the method based on private encoders overlooks shared features.

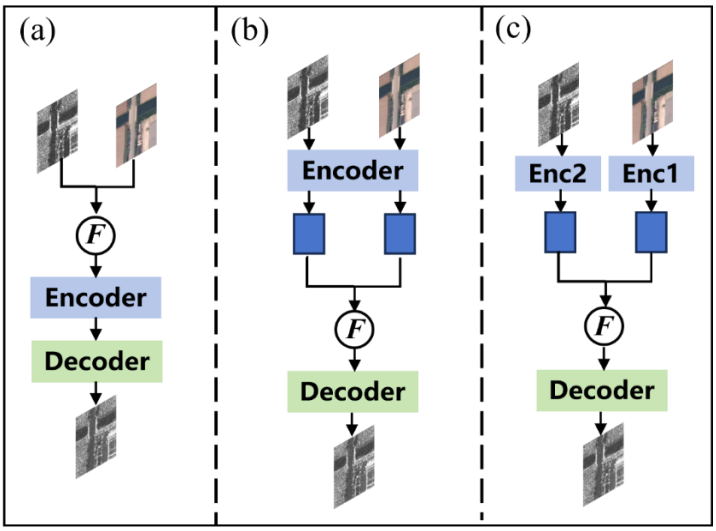

Figure 1. Existing AE fusion algorithm frameworks. (**a**,**b**) represent the process based on a shared encoder, while (**c**) represents the process based on private encoders.

Unlike CNNs, the Vision Transformer (ViT) [25] model architecture, which has recently become popular in the field of computer vision, utilizes mechanisms such as self-attention, multi-head attention, and positional encoding. This enables the model to effectively capture global dependencies within the input sequence, thereby providing outstanding global feature extraction capabilities. However, network models based on ViT are relatively complex and require substantial computation to achieve better performance.

To address these issues, this paper proposes a more rational fusion network architecture to solve the shortcomings and challenges in feature extraction and fusion. The fusion algorithm framework designed in this paper is illustrated in Figure 2.

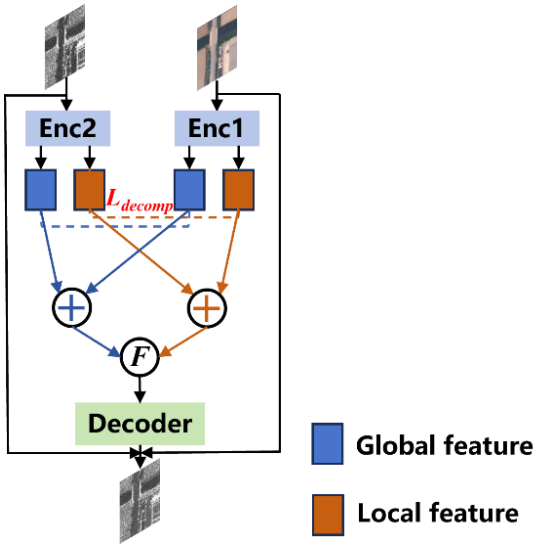

Figure 2. The dual-branch AE fusion algorithm framework designed in this paper.

First, addressing the lack of global feature extraction capability in CNNs, this paper introduces a dual-branch feature extraction network based on a Transformer and a CNN to separately extract and fuse global high-frequency features and local low-frequency features from visible and SAR images.

Second, addressing the potential loss of important feature information during the fusion process, this paper makes relevant improvements to the Transformer and CNN feature extraction models, enhancing the network's ability to capture important feature information. On one hand, based on the Transformer network structure, we introduce the LT [26] block to balance fusion image quality and reduce computational costs and the DropKey [27] mechanism in the network's attention layer to adaptively adjust attention weights, making the model focus on more useful information. On the other hand, based on the CNN network model, we have added the CBAM module, which enhances the network model's focus on important areas by introducing channel attention and spatial attention mechanisms, thereby reducing the loss of important information.

Third, regarding visible and SAR images, we believe that the large-scale environmental features such as background and contour of different modal data have high similarity, showing high correlation in global features, whereas for different modal textures and details, they show some differences and independence, demonstrating low correlation in local features. Therefore, we promote the feature extraction capability and effectiveness of different modal data by increasing the correlation of global features and reducing the correlation of local features in visible and SAR images.

In summary, the main contributions of this paper are as follows:

(1) We propose a dual-branch Transformer–CNN framework to extract and fuse global and local features of visible and SAR images, addressing the issue of insufficient feature extraction in traditional auto-encoder-based image fusion methods.

(2) On a macroscopic level, on the one hand, we have made innovative improvements to the dual-branch structure. Instead of traditionally concatenating global and local features of different modal data and then sending them to the decoder for reconstruction, we first concatenate the global features and then send the fused global features along with the local features of visible and SAR images to the decoder for reconstruction of the original images. On the other hand, we have introduced a residual structure to the model to enhance network performance and expressive capability, strengthening the extraction of complex features.
(3) On a microscopic level, we have made some improvements to the dual-branch feature extraction network model, specifically including the introduction of the LT and DropKey mechanisms in the Transformer feature extraction network and the addition of the CBAM module in the CNN feature extraction network to reduce the potential loss of important feature information during the forward propagation of the fusion network and enhance the robustness of the model.
(4) For the two-stage training process, we designed specific loss functions to suit different training tasks, achieving good results.

The specific chapters and arrangements of this paper are as follows: Section 2 introduces the related work on visible and SAR image fusion methods; Section 3 describes, in detail, the visible and SAR image fusion method and the related structure used in this paper; Section 4 introduces the related experimental work and presents the experimental results and analysis; finally, this paper concludes in Section 5.

2. Related Work

In this section, we mainly introduce some related work on image fusion methods.

2.1. CNN

Image fusion methods based on CNN mainly leverage the powerful feature extraction capabilities of CNN networks, retaining rich detail information in the fused images. In 2017, Liu et al. [28] introduced CNN networks into the field of image fusion. They trained the network using blurred background and foreground images to obtain binarized weight maps. During the testing phase, the source images were combined with the weight maps to produce fused multifocus images. Subsequently, many researchers introduced CNN network models into traditional methods, infusing rich semantic information into the fused images. For example, Li et al. [29] used the VGG19 network to further process the detail parts obtained through multi-scale decomposition, thus preserving rich texture information in the fused images. Liu et al. [30] used a downsampling sequence of convolutional weight maps as the fusion ratio map of two-branch downsampling sequences, avoiding manually designed fusion strategies. These methods all share a common issue: they do not fully consider the different information among different modal images.

2.2. Attention Mechanism

The attention mechanism is a commonly used module in image processing that is used to focus on the important features of the image and inhibit unnecessary regional responses. In 2014, the Google Mind team used the attention mechanism in the RNN model for image classification, which resulted in its research and use by many scholars. In general, the attention mechanism can be divided into soft attention, hard attention, and the self-attention that is used in the field of Natural Language Processing (NLP). Among them, the soft attention mechanism can currently be subdivided into channel attention, spatial attention, and its combination module [31]. Woo et al. [32] proposed CBAM through the channel dimension and spatial dimension in a combinatorial analysis study and confirmed that the performance of the network is enhanced by the accurate attention mechanism and the suppression of noisy information. CBAM is a combination of the channel attention module (CAM) and the spatial attention module (SAM) used for enhancing the performance of the feedforward convolutional attention module to enhance

the performance of CNNs. It can be integrated into any network model of CNN architecture with negligible computational cost and is a neural network that enables end-to-end training. Currently, CBAM has been applied to a variety of common Convolutional Neural Networks for enhancing network performance, such as ResNet [33], VGG [34], DenseNet [35], etc.

The CAM mainly models the importance of features, and its structure is shown in Figure 3. Its main process consists in using both the maximum pooling and mean pooling algorithms, then going through several MLP layers to obtain the transformed results, and finally applying them to the two channels separately to obtain the channel's attention results using the sigmoid function.

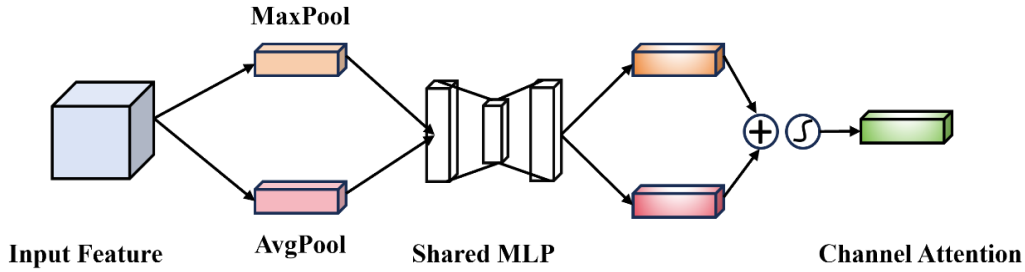

Figure 3. Channel attention module.

The SAM models the importance of spatial locations, and its structure is shown in Figure 4. Its main process consists in first downscaling the channel itself to obtain the maximum pooling and mean pooling results, respectively, and then stitching them into a feature map, which is then learned using a convolutional layer.

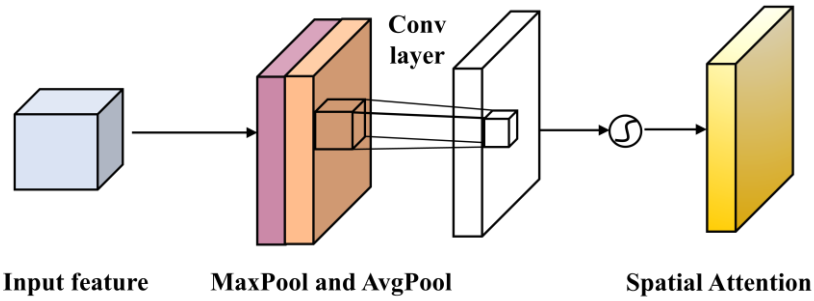

Figure 4. Spatial attention module.

2.3. Transformer and Its Variants

Transformer is a classic NLP model proposed by Vaswani et al. [36] in 2017 that relies entirely on self-attention to compute its inputs and outputs. The Vision Transformer (ViT) was introduced by Dosovitskiy [25] for computer vision applications. Compared to CNN, ViT and its variants have achieved many advanced results in image processing. For example, Wang et al. [37] proposed PVT, which integrates Transformer into CNN and trains on dense partitions of images to produce high-resolution outputs, overcoming the drawbacks of Transformer for dense prediction tasks. Wu et al. [26] proposed an efficient mobile NLP architecture, LT, which features long- and short-range attention to significantly reduce computational costs. Zamir et al. introduced the Restormer [38] structure, incorporating Multi-head Dconv Transfer Attention (MDTA) modules and gated-Dconv feed-forward network (GDFN) for multi-scale local/global representation learning in high-resolution images.

LT is a novel lightweight Transformer network with two enhanced self-attention mechanisms to improve the performance of edge deployment. For low-level features, Convolutional Self-Attention (CSA) is introduced. Unlike previous approaches that fused convolution and self-attention, CSA introduces local self-attention into the convolution within a kernel of size 3×3 to enrich the low-level features in the first stage of LT. For high-level features, Recursive Atrous Self-Attention (RASA) is proposed to compute similarity

mappings using multi-scale contexts, and a recursive mechanism is employed to increase the representational power of additional marginal parameter costs.

In the image recovery task, although the existing Transformer model can overcome the problems of the limited sensory field of CNN and its non-adaptability to the input content, its computational complexity grows quadratically with the spatial resolution, and thus it cannot be applied in the recovery task of high-resolution images. In contrast, Restormer, as an efficient Transformer network for image restoration, is applicable to the task of restoring and reconstructing large images by introducing an MDTA module and a new GDFN that models global connectivity.

The main improvements to ViT mainly focus on two aspects: on the one hand, enhancing or replacing the original network's ReLU structures due to insufficient non-linearity, as exemplified in LT; on the other hand, introducing the DropOut mechanism during the training process of Transformer networks to prevent overfitting, thereby helping the model extract more useful feature information. This paper introduces the DropKey mechanism based on the LT network to improve the network.

2.4. Regularization Method

In machine learning, when the model is continuously optimized, image blocks with a larger share of attention in the current iteration will tend to be assigned larger attention weights during the next iteration, thus predisposing them to overfitting problems. In order to solve such problems, many machine learning algorithms use related strategies to reduce the test error, which are collectively known as regularization. Currently, the main strategies used in deep learning are Parameter Norm Penalties, Early Stopping, DropOut, etc. In 2012, Alex proposed DropOut, which is based on the principle of improving the performance of neural networks by preventing the co-action of feature detectors to alleviate the neural network overfitting problem. And in that year's image recognition competition, Alex et al. used the DropOut algorithm in the AlexNet network to prevent the overfitting problem and eventually won the competition. As shown in Figure 5b, DropOut involves randomly discarding the attention weights after Softmax normalization, but this breaks the probability distribution of the attention weights and fails to penalize the weight peaks, resulting in the model still overfitting to locally specific information. In this paper, we use a novel regularization method, DropKey [27], shown in Figure 5c, which implicitly assigns an adaptive operator to each attention block to constrain the attention distribution by randomly dropping some of the key vectors (thus making it smoother) and also encouraging the model to pay more attention to the useful information of the other image blocks, which can help to capture globally robust features.

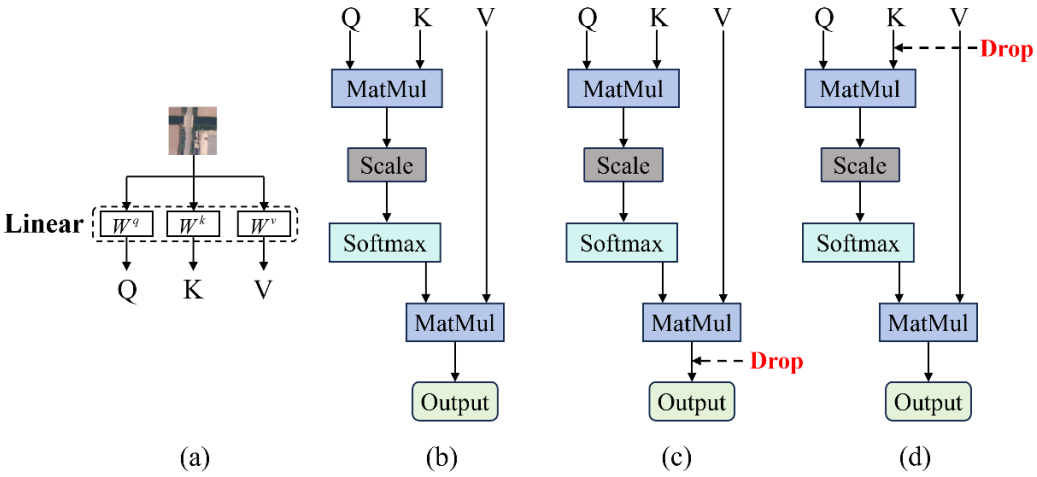

Figure 5. Comparison between DropOut and DropKey schematic diagrams. (**a**) Acquisition of Q, K, and V. (**b**) Transformer schematic diagram. (**c**) DropOut schematic diagram. (**d**) DropKey schematic diagram.

Q, K, and V, shown in Figure 5, are the three key components inside the self-attention mechanism in the Transformer network, denoted as query vectors, key vectors, and value vectors, respectively, which are all obtained from the input matrices by linear transformation, as shown in Figure 5a. In the self-attention mechanism, a weight distribution is obtained by calculating the similarity between the query vector and all the key vectors based on the query vector, which is used to weight and sum the associated value vectors. Firstly, the inner product (MatMul) of matrices Q and the vectors of each row of K is calculated, and in order to prevent the inner product from being too large, it is divided by the square root of dk (Scale), where dk is the dimension of the K matrix; secondly, the result of the above inner product is normalized using Softmax; finally, the Softmax matrix is obtained and then multiplied with the V matrix to obtain the final output.

3. Framework and Methodology

In this section, we introduce the method and framework we proposed; we have also designed the corresponding loss functions for this method. The algorithm framework of this paper is shown in Figure 6. Below, we will introduce it from four aspects: encoder, fusion strategy, decoder, and loss function.

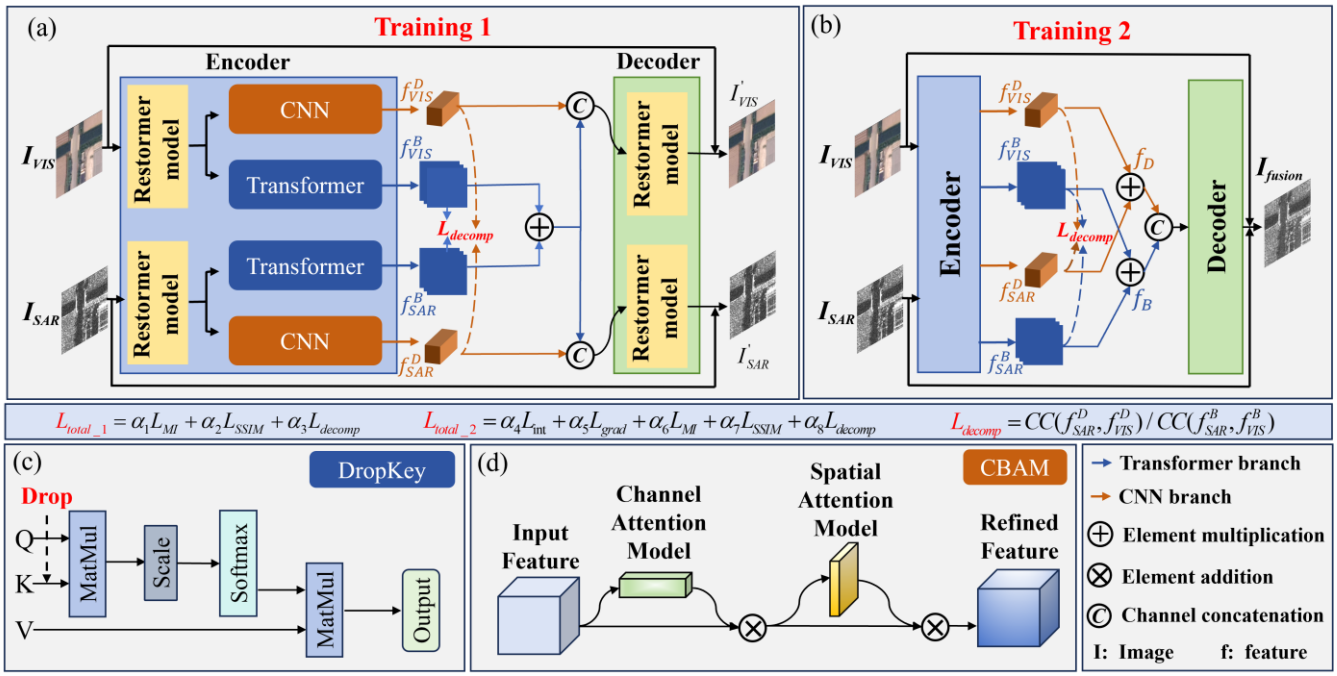

Figure 6. Fusion method structure of this experiment. (**a**) First stage of training process; (**b**) second stage of training process; (**c**) DropKey principle; (**d**) CBAM principle.

3.1. Encoder

The encoder part is mainly used for feature extraction of input images, which consists of three parts: shallow feature extraction, global feature extraction, and local feature extraction. The specific details are as follows:

Shallow Feature Extraction. Initially, the Restormer Block extracts shallow features from the input visible and SAR images and then continues to extract their global/local features based on the extracted shallow features. The Restormer Block has been proven to extract shallow features of images without increasing computational power, facilitating multi-scale global/local representation learning suitable for image reconstruction tasks.

Global Feature Extraction. Based on the shallow features extracted by the Restormer Block, we use the LT model to extract the global features of the input images. The LT model, by adopting long- and short-range attention, focuses more on the global information of images and reduces model parameters through a flattened feedforward network structure,

significantly reducing computational costs while maintaining the same performance. At the same time, we introduce the DropKey mechanism in the attention layer, randomly dropping some key values to reduce the model's over-reliance on certain neurons, helping to capture more robust global features.

Local Feature Extraction. Local feature extraction aims to extract detailed features such as texture information and corner features from image data. The CNN feature extraction network is currently one of the most effective methods for extracting image detail features. To capture more detailed feature information and reduce the loss of important information during the fusion process, we introduced the CBAM module at the front end of the CNN feature extraction network. This module adaptively adjusts the importance of different channel pieces of information and assesses the relevance of different spatial positions to enhance the network's focus on important areas.

3.2. Fusion Strategy

First, a fusion layer is constructed, whose main structure is similar to the feature extraction structure of the encoder. Therefore, we similarly adopt a Transformer network with the LT module and DropKey mechanism, as well as a CNN network with the CBAM module as the fusion strategy. For the first training stage, our approach is to first fuse and concatenate the global features extracted from the visible and SAR images, then send these concatenated features along with the local features of the visible and SAR to the decoder to reconstruct the original images. The purpose of this is to train an encoder that can extract global features of visible and SAR with higher relevance. For the second training stage, our approach is to input the visible and SAR images into the trained encoder, then fuse and concatenate the extracted global and local features, and send these concatenated features to the decoder for decoding to reconstruct the fused image.

3.3. Decoder

The decoders in the first and second training stages are structurally identical, both using Restormer Blocks as their basic unit, but they differ in function. The decoder in the first training stage mainly receives the global/local features from the visible and SAR images and ultimately reconstructs the original images, while the decoder in the second training stage receives the globally and locally concatenated features of the visible and SAR images and is capable of reconstructing the fused image.

3.4. Loss Function

Inspired by reference [39], this paper designs a two-stage training process. As introduced above, the tasks and functions realized in the first and second stages are not completely the same; therefore, we have designed specific loss functions for the training processes of both stages.

3.4.1. Training Stage 1

In training stage 1, the total loss function trained is calculated as follows:

$$L_{total\_1} = \alpha_1 L_{MI} + \alpha_2 L_{SSIM} + \alpha_3 L_{decomp} \tag{1}$$

where α_1, α_2, and α_3 refer to the adjustment coefficients, which are 3, 10, and 1, respectively. L_{MI}, L_{SSIM}, and L_{decomp} respectively refer to the mutual information loss, structural similarity loss, and feature decomposition loss of visible and SAR images, which are defined as follows:

- Mutual information loss

  The specific expression for the MI loss function is as follows:

$$L_{MI}(x, y) = H(x) + H(y) - H(x, y) \tag{2}$$

where x and y represent the original image and the reconstructed image, respectively. $H(x)$ and $H(y)$ represent the information entropy of the original image and the reconstructed image, respectively, and $H(x,y)$ represents the joint information entropy of the source image and the reconstructed image.

- Structural similarity loss

  The specific expression for the *SSIM* loss function is as follows:

$$L_{SSIM} = 1 - SSIM(I_f, I_{VIS}) + \beta(1 - SSIM(I_f, I_{SAR})) \tag{3}$$

where β represents the adjustment coefficient of 0.5. $SSIM(\ ,\ )$ is the structural similarity index, and its specific expression is as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

where x and y represent the original image and the reconstructed image, respectively; μ_x and μ_y represent the means of the original and reconstructed images; σ_x^2 and σ_y^2 represent the variances of the original and reconstructed images; σ_{xy} represents the covariance of the original and reconstructed images; $C_1 = (k_1 L)^2, C_2 = (k_2 L)^2$ is a constant used to maintain stability; and L is the dynamic range of image pixel values, with $k_1 = 0.01, k_2 = 0.03$.

- Feature decomposition loss

L_{decomp} is a loss function of our own design that aims to better distinguish between the extracted global feature information and the local feature information. It is defined as follows:

$$L_{decomp} = \frac{CC(f_{SAR}^D, f_{VIS}^D)}{CC(f_{SAR}^B, f_{VIS}^B)} \tag{5}$$

where $CC(\ ,\ )$ refers to the correlation coefficient operator; f_{SAR}^D and f_{VIS}^D respectively refer to the detailed local features extracted from SAR images and visible images; and f_{SAR}^B and f_{VIS}^B respectively refer to the global features extracted from SAR images and visible images. Equation (5) is designed based on the viewpoint we proposed earlier because, in our view, visible and SAR images should be highly correlated in terms of global feature information. In order to preserve the same global information for both types of images, the larger $CC(f_{SAR}^B, f_{VIS}^B)$, the better. In terms of local detail feature information, there are certain differences between the two types of images. In order to extract richer details, the smaller $CC(f_{SAR}^D, f_{VIS}^D)$, the better. Therefore, this article proposes the above loss function.

3.4.2. Training Stage 2

In training stage 2, the total loss function of the training is calculated as follows:

$$L_{total\_2} = \alpha_4 L_{\text{int}} + \alpha_5 L_{grad} + \alpha_6 L_{MI} + \alpha_7 L_{SSIM} + \alpha_8 L_{decomp} \tag{6}$$

where α_4, α_5, α_6, α_7, and α_8 are adjustment coefficients, which are 1, 1, 3, 10, and 1, respectively. On the basis of the first training stage loss function, two terms, L_{int} and L_{grad}, have been added, where L_{int} is the intensity loss of the image, which constrains the fused image to maintain a similar intensity distribution to the source image; and L_{grad} is the gradient loss of the image, forcing the fused image to contain rich texture details. The specific definition formula is as follows:

$$L_{\text{int}} = \frac{1}{H \times W} \| I_f - \max(I_{SAR}, I_{VIS}) \|_1 \tag{7}$$

$$L_{grad} = \frac{1}{H \times W} \| \nabla I_f - \max(\nabla I_{SAR}, \nabla I_{VIS}) \|_1 \tag{8}$$

where I and ∇I respectively refer to the operators of the image intensity and gradient magnitude, while H and W respectively refer to the height and width of the image.

4. Experimental Setup and Result Analysis

In this section, we first introduce the dataset used in this experiment, then detail some parameter configurations and the implementation process of the experiment, compare it with existing visible and SAR image fusion methods, and finally, conduct an ablation study to prove the advancement and reference value of our proposed image fusion method.

4.1. Dataset Introduction

The dataset used in this experiment is OGSOD-1.0 [40], a publicly available dataset downloaded from the Internet. The SAR images in OGSOD-1.0 are collected from the Chinese Gaofen-3 satellite in the C-band, Vertical–Vertical (VV), and Vertical–Horizontal (VH) polarization modes. These SAR images are provided by the 38th Research Institute of China Electronics Technology Group Corporation (CETGC), and their resolution is 3 m. The optical images are provided by Google Earth, and their resolution is 10 m. In addition, to increase the diversity of the training set, the original authors obtained permission from Michael Schmitt to extend the dataset by selecting an additional 3000 sample pairs from the SEN1-2 [41] dataset. Therefore, OGSOD-1.0 consists of a training set of 14,665 optical and SAR image pairs and a test set of 3666 SAR-only images, containing a total of more than 48,000 instance annotations. For this experiment, we selected 1048 pairs from the dataset as the training set and 100 pairs as the test set.

4.2. Evaluation Metrics

To verify the fusion performance of the algorithm proposed in this paper, the experiment quantitatively evaluates the fusion results from four aspects and a total of 12 common metrics: information-based, structure similarity-based, image feature-based, and human visual perception-based. The information-based image fusion metrics include entropy (EN), mutual information (MI), and peak signal-to-noise ratio (PSNR); the structure similarity-based metrics include Structural Similarity Index Measure (SSIM) and Mean Squared Error (MSE); the image feature-based metrics include Average Gradient (AG), Edge Intensity (EI), Standard Deviation (SD), Spatial Frequency (SF), and edge information-based index (Qabf); and the visual perception-based metrics include Sum of Correlated Differences (SCDs) and Visual Information Fidelity (VIF). They are categorized in Table 1. Except for the MSE, where a smaller value indicates higher image quality, higher values in all other metrics indicate better image quality after fusion.

Table 1. Classification of quantitative evaluation metrics used in the experiment [42].

Theory	Evaluation Metrics
Information Theory	EN, MI, PSNR
Structural Similarity	SSIM, MSE
Image Feature	AG, EI, SD, SF, Qabf
Visual Perception	SCD, VIF

4.3. Experimental Setup

All algorithm implementations were trained and tested on a high-performance workstation equipped with an Nvidia Tesla A100 GPU with 80 GB of memory and an AMD Ryzen Threadripper PRO 5995WX 64-Core CPU. The deep learning framework is PyTorch, using CUDA version 11.7. During the training phase, the input image size was set to 256×256, with a total of 140 training epochs, where the first and second phases were 40 and 100 epochs, respectively. The batch size was set to 16, with an initial learning rate of 10^{-4}, reduced by 50% every 20 epochs.

4.4. Comparison with SOTA Methods

To verify the effectiveness of the proposed image fusion method, the experiment compared the fusion results with advanced general image fusion methods, including DenseFuse [43], RFN-Nest [44], SeAFusion [45], SwinFusion [46], and YDTR [47].

4.4.1. Qualitative Comparison

To better evaluate the fusion performance of various algorithms, this experiment selected three pairs of visible-light and SAR images with rich texture details from the test set for comparative display. The original pairs of visible-light and SAR images are shown in Figure 7. From the figure, it can be seen that visible-light images have a better visual effect and clearly express local features such as buildings. However, their contour information is difficult to distinguish from the background. Conversely, SAR images express contour information more fully. Therefore, the fused image should include both local information, such as buildings, and global information, such as terrain contours.

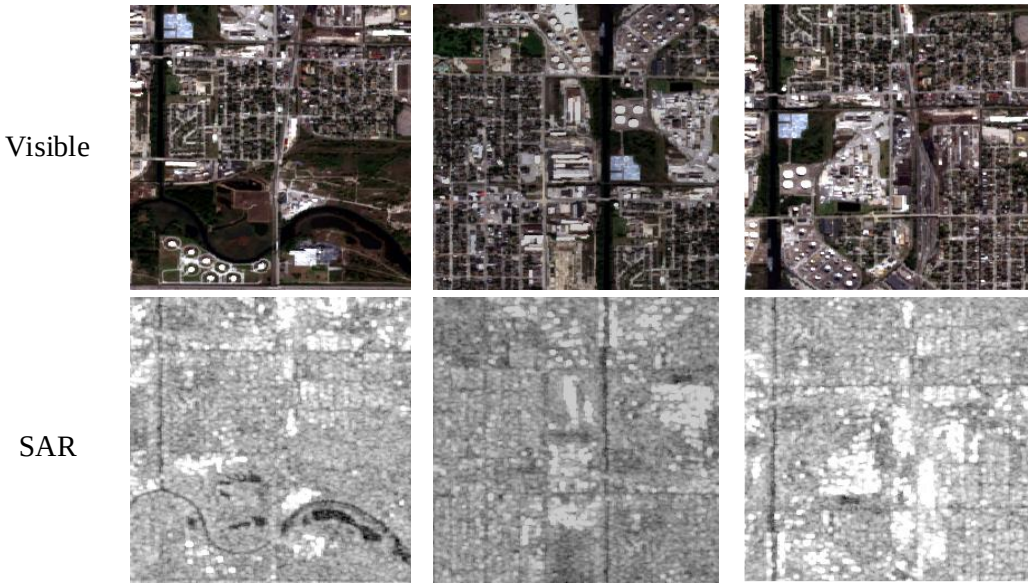

Figure 7. Three pairs of visible and SAR original images.

Figure 8 shows a visual comparison of the fused images obtained by our proposed visible and SAR image fusion method and the five methods mentioned above, with red boxes highlighting some detailed comparisons of the fused images from each method. From the comparison results, it can be seen that our proposed method captures more abundant texture details and clearer contour information in the fused images compared to the other five methods, and the fused images obtained by our method make the target objects more prominent and easier to distinguish from the background, helping us better understand various scenes.

Upon examination of the fused images, it becomes evident that they all exhibit a monochromatic appearance devoid of color. This is a notable departure from the conventional characteristics of visible and SAR fusion images, which will be elucidated below. In the execution of our algorithm, we first compress the RGB bands of the visible image into a single channel during the data processing stage, which results in the loss of color information. This is performed to ensure that the visible input and the SAR input have the same number of channels, which facilitates the overall execution of the algorithm.

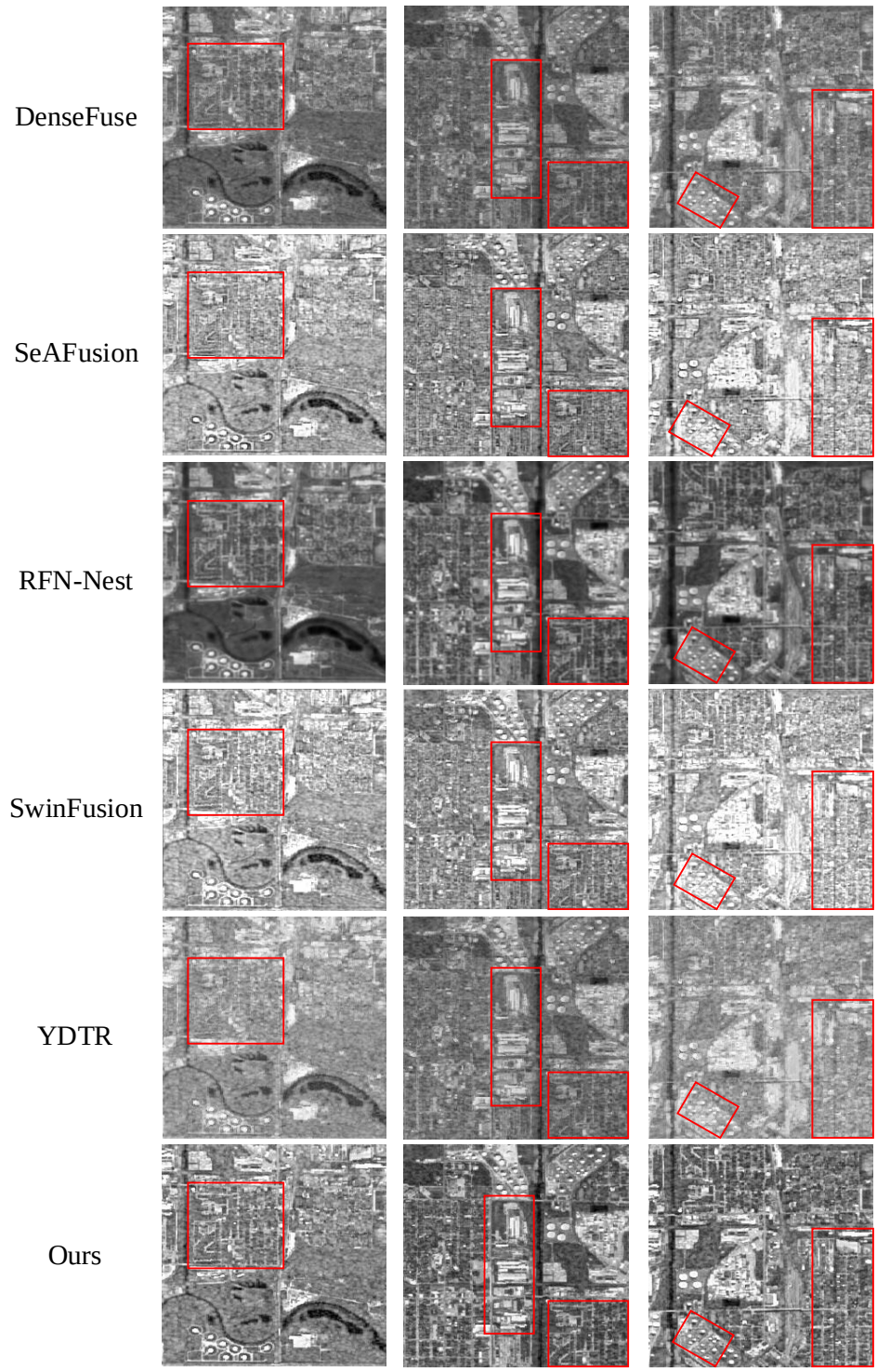

Figure 8. Comparison of fused images obtained by different methods. Red boxes highlight some detailed comparisons of the fused images from each method.

4.4.2. Quantitative Comparison

To verify the superiority of the proposed algorithm more objectively, Table 2 shows a quantitative index comparison between our method and the other five methods, where the bolded data are the best values for each index. From the experimental data in Table 2, it can be seen that the *SSIM* metrics obtained by some methods are greater than 1, which is not in line with common sense. This is because we have made a small change to the SSIM metrics when calculating them (the calculation expression is shown in Equation (9)). It can

be seen that the *SSIM* metrics in this paper are obtained by calculating the SSIM metrics from the fused images with the visible and SAR images, respectively, and then summing them, which is why the size of the metrics may be greater than 1. In addition, similar to this practice, MI, MSE, CC, PSNR, SCD, VIFF, Qabf, and other metrics are calculated.

$$M_{SSIM} = SSIM(I_f, I_{VIS}) + SSIM(I_f, I_{SAR}) \tag{9}$$

Table 2. Comparison of evaluation metrics for different fusion methods.

	DenseFuse	SeAFusion	RFN-Nest	SwinFusion	YDTR	Ours
EN	7.02	7.42	7.06	7.23	7.09	**7.54**
MI	1.71	1.64	1.51	1.75	1.78	**2.42**
PSNR	**15.47**	12.99	14.50	12.92	14.63	14.09
SSIM	1.04	1.01	0.80	1.06	1.05	**1.10**
MSE	**1974.79**	3614.99	2532.79	3656.79	2451.7	2820.47
AG	10.15	15.40	6.92	15.21	11.96	**17.15**
EI	41.50	48.25	37.81	48.37	42.74	**50.19**
SD	36.54	48.64	39.58	49.50	39.06	**52.88**
SF	24.28	33.43	14.31	34.88	30.17	**40.08**
Qabf	0.32	0.44	0.22	0.41	0.39	**0.56**
SCD	1.21	1.35	1.20	1.49	1.16	**1.57**
VIF	0.37	0.35	0.31	0.36	0.39	**0.58**

In order to show the comparison effect more intuitively, we normalized the data in Table 2 and plotted them as a radar chart, as shown in Figure 9. Since this is just a simple representation of the advantages and disadvantages of the metrics obtained by different methods, the normalization process we have adopted is to take the maximum value of the metrics in each category to be 1, while the metrics obtained by the other methods in this category are taken to be the ratio of their actual metrics to the maximum actual metrics in that category. However, the MSE metrics are special in that the smaller the metrics, the better the quality of the fused image generated. In order to facilitate the intuitive understanding of the human eye, we do the opposite of normalizing the metrics of the MSE. We set the minimum metric to 1, while the metrics obtained by other methods take the value of the ratio of the minimum actual metric to its actual metric.

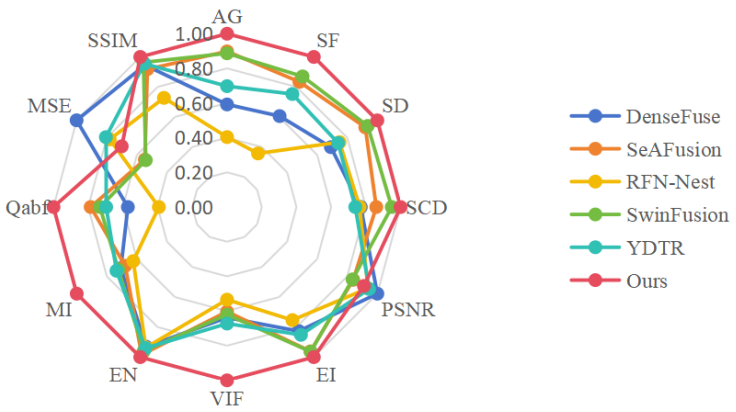

Figure 9. Radar chart of results of experimental metrics.

It can be seen that our proposed method performs well in all metrics except for 2 (PSNR and MSE) and is the best in the other 10 metrics, demonstrating that our method performs better in visible and SAR image fusion tasks. Specifically, our method performs best on the EN and MI metrics, indicating that it can fully mine and transfer the information from the source images to the fused images; it also performs best on the SSIM index, showing that it can retain the detailed information of the source images, being most similar to them; it

performs best on the AG, EI, SD, SF, and Qabf metrics, indicating that the fused images obtained by our method are of higher quality and clarity; and it performs best on the SCD and VIF metrics, demonstrating that the fused images obtained by our method have better visual effects.

4.5. Ablation Studies

In this section, we validate the rationality of different modules through a set of ablation experiments. Specifically, we conducted ablation studies on the dual-branch structure, residual structure, DropKey mechanism, CBAM module, and two-stage training used in our experiments. The details are as follows:

(1) Dual-branch structure: In this paper, we design a CNN-based and a Transformer-based dual-branch structure, and in order to prove the effectiveness of the dual-branch structure, we design ablation experiments as follows: (a) We use only the Transformer branch to complete the feature extraction, i.e., the CNN branch is replaced by the Transformer branch. (b) We use only the CNN branch to complete the feature extraction, i.e., the Transformer branch is replaced by the CNN branch.
(2) Residual structure: A comparative experiment is conducted by comparing scenarios with and without the introduction of the residual structure.
(3) DropKey: For the Transformer branch, a comparative experiment is conducted between using the DropKey mechanism and not using it.
(4) CBAM: For the CNN branch, an experiment is conducted comparing the use of the CBAM module against not using it.
(5) Two-stage training: This experiment introduced two-stage training to enhance fusion performance. In the ablation study, a one-stage training method directly trains the encoder, fusion layer, and decoder. The number of training rounds is consistent with the total number of rounds in the two-stage training, both at 140 rounds.

Based on the above ablation experimental setup, we obtained the experimental results and recorded them as shown in Table 3. From the comparative results, it is evident that in certain group comparisons, the methodology we employed exhibited a slight deterioration in several metrics. However, the overall enhancement across the majority of metrics substantiates the rational design of our proposed structure. Furthermore, it is noteworthy that the metrics derived from the dual-branch experiments significantly and consistently surpassed those obtained solely from the CNN branch. As for the results obtained by using only the Transformer branch, we can see that in the five indexes based on image features, the method using only the Transformer branch is even superior in four of them, which indicates that the Transformer branch we added has a strong feature extraction capability. However, in terms of overall performance, the dual-branch structure we use has a greater advantage in the other eight indicators, which indicates that the dual-branch structure adopted in this study is reasonable and effective. Additionally, in the ablation experiments involving DropKey and CBAM, our approach demonstrated notable improvements in the PSNR, MSE, and SSIM metrics. These results suggest that our method preserves more original image information and exhibits superior performance in representing details and textural features.

Additionally, Figure 10 intuitively displays the comparative results of the ablation experiments, highlighting certain aspects within red boxes to showcase detailed contrasts in the fusion images obtained by each group. An analysis of these comparisons reveals that the fusion images produced by our proposed method exhibit superior fusion quality. Specifically, the fusion images generated solely using the CNN branch contain more noise, which could hinder further image processing; certain texture details are lost in the fused image obtained without using residual structures; and the images resulting from only one-stage training also show some loss in textural detail, with weaker contrast between structures such as buildings and their backgrounds compared to our method. Furthermore, it is clearly visible that the fusion images obtained without employing DropKey and CBAM

lack detailed textures and appear more blurred, demonstrating the significant role of DropKey and CBAM utilized in our study.

Table 3. Results of ablation experiments.

	Information Theory			Structural Similarity			Image Features				Visual Perception	
	EN	**MI**	**PSNR**	**SSIM**	**MSE**	**AG**	**EI**	**SD**	**SF**	**Qabf**	**SCD**	**VIF**
(1) Dual-branch structure												
Transformer branch	7.52	1.73	13.47	1.01	3142.91	19.72	50.89	53.39	47.69	0.48	1.33	0.38
CNN branch	7.36	1.02	11.40	0.41	4898.22	7.45	34.36	42.53	23.62	0.15	0.07	0.15
Ours	7.54	2.42	14.09	1.10	2820.47	17.15	50.19	52.88	40.08	0.56	1.57	0.58
(2) Residual structure												
Nonresidual	7.52	2.41	13.14	0.98	3471.32	17.37	49.31	52.04	40.64	0.55	1.15	0.56
Ours	7.54	2.42	14.09	1.10	2820.47	17.15	50.19	52.88	40.08	0.56	1.57	0.58
(3) DropKey												
No Dropkey	7.51	1.67	13.64	0.94	3049.84	17.11	47.86	49.74	42.08	0.45	1.07	0.34
Ours	7.54	2.42	14.09	1.10	2820.47	17.15	50.19	52.88	40.08	0.56	1.57	0.58
(4) CBAM												
No CBAM	7.45	2.95	12.75	0.99	3886.71	16.23	49.47	51.02	37.53	0.55	0.97	0.75
Ours	7.54	2.42	14.09	1.10	2820.47	17.15	50.19	52.88	40.08	0.56	1.57	0.58
(5) Two-stage training												
One stage	7.53	2.22	13.23	1.01	3425.46	18.50	50.76	52.41	43.25	0.55	1.19	0.52
Ours	7.54	2.42	14.09	1.10	2820.47	17.15	50.19	52.88	40.08	0.56	1.57	0.58

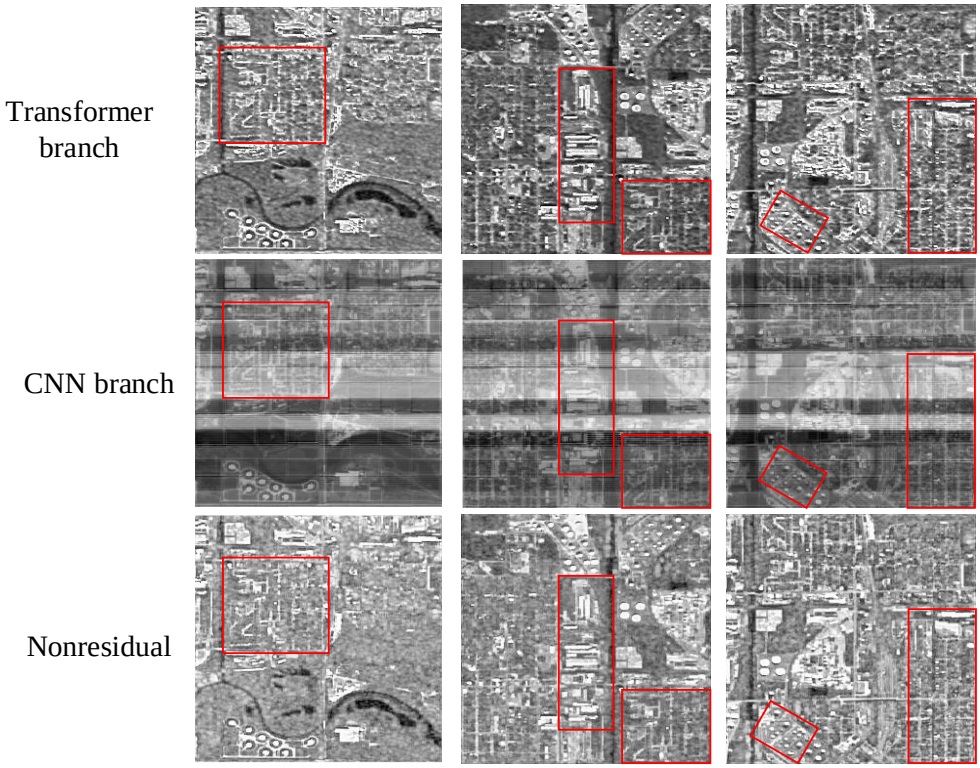

Figure 10. *Cont.*

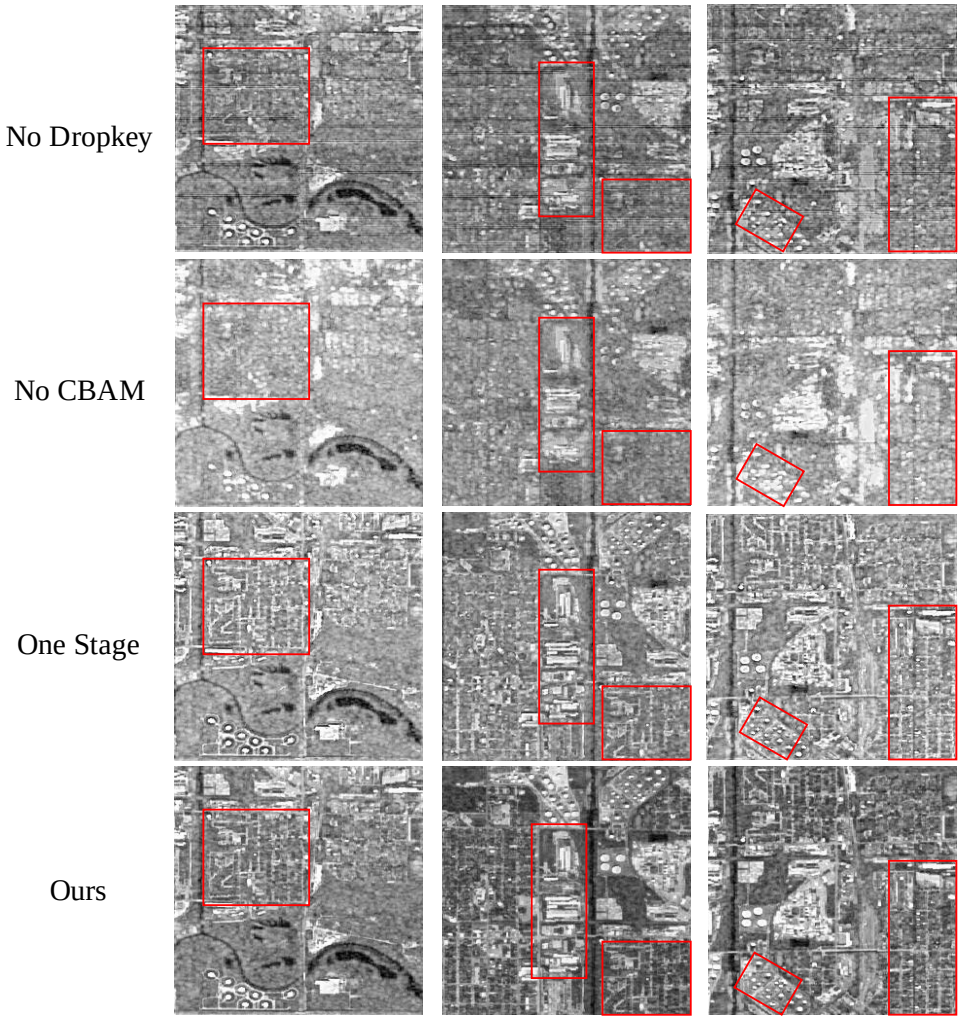

Figure 10. Display of ablation experiment results. Red boxes highlight some detailed comparisons of the fused images from each group.

In summary, the results of the ablation experiments show that our designed method is effective and rational.

5. Conclusions

This article proposes a visible and SAR image fusion method based on a dual-branch residual structure combining Transformer and CNN networks. It introduces the LT and DropKey mechanisms into the feature extraction network based on Transformer and incorporates the CBAM module into the feature extraction network based on CNN to better extract global and local features from both modalities. In addition, we have made certain improvements to the entire fusion network architecture by first fusing and concatenating the global features of the two modalities and then inputting the concatenated features separately with the local features of each modality into the decoder for reconstruction. To this end, we have also designed a specific loss function to adapt to this task. Finally, through comparative experiments with five other methods and ablation experiments, we have demonstrated the effectiveness and feasibility of our proposed method.

Author Contributions: L.H.: Writing the original draft, Methodology, Investigation, Software; S.S.: Supervision, Validation; Z.Z. (Zhen Zuo): Methodology, Writing—review and editing; J.W.: Methodology, Project administration; S.H.: Methodology, Visualization, Conceptualization, Software; Z.Z. (Zongqing Zhao): Software, Conceptualization; X.T.: Resources, Data curation; S.Y.: Software, Resources. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Youth Foundation of China grant number 62201598.

Data Availability Statement: The dataset OGSOD-1.0 [40] used in this study is openly available at the following links: https://github.com/mmic-lcl/Datasets-and-benchmark-code, (accessed on 5 May 2024).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Zhang, H.; Shen, H.F.; Yuan, Q.Q.; Guan, X.B. Multispectral and SAR Image Fusion Based on Laplacian Pyramid and Sparse Representation. *Remote Sens.* **2022**, *14*, 870. [CrossRef]
2. He, Y.Q.; Zhang, Y.T.; Chen, P.H.; Wang, J. Complex number domain SAR image fusion based on Laplacian pyramid. In Proceedings of the 2021 CIE International Conference on Radar (Radar), Haikou, China, 15–19 December 2021.
3. Zhang, T.W.; Zhang, X.L. Squeeze-and-Excitation Laplacian Pyramid Network With Dual-Polarization Feature Fusion for Ship Classification in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4019905. [CrossRef]
4. Dai, J.Y.; Lv, Q.; Li, Y.; Wang, W.; Tian, Y.; Guo, J.Z. Controllable Angle Shear Wavefront Reconstruction Based on Image Fusion Method for Shear Wave Elasticity Imaging. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.* **2022**, *69*, 187–198. [CrossRef] [PubMed]
5. Jia, H.Y. Research on Image Fusion Algorithm Based on Nonsubsampled Shear Wave Transform and Principal Component Analysis. *J. Phys. Conf. Ser.* **2022**, *2146*, 012025. [CrossRef]
6. Zhao, M.J.; Peng, Y.P. A Multi-module Medical Image Fusion Method Based on Non-subsampled Shear Wave Transformation and Convolutional Neural Network. *Sens. Imaging* **2021**, *22*, 9. [CrossRef]
7. Singh, S.; Singh, H.; Gehlot, A.; Kaur, J.; Gagandeep. IR and visible image fusion using DWT and bilateral filter. *Microsyst. Technol.* **2023**, *29*, 457–467. [CrossRef]
8. Amritkar, M.A.; Mahajan, K.J. Comparative Approach of DCT and DWT for SAR Image Fusion. *Int. J. Adv. Electron. Comput. Sci.* **2016**, *3*, 107–111.
9. Cheng, C.; Zhang, K.; Jiang, W.; Huang, Y. A SAR-optical image fusion method based on DT-CWT(Article). *J. Inf. Comput. Sci.* **2014**, *11*, 6067–6076. [CrossRef]
10. Zhang, K.; Huang, Y.D.; Zhao, C. Remote sensing image fusion via RPCA and adaptive PCNN in NSST domain. *Int. J. Wavelets Multiresolut. Inf. Process.* **2018**, *16*, 1850037. [CrossRef]
11. Liu, K.X.; Li, Y.F. SAR and multispectral image fusion algorithm based on sparse representation and NSST. In Proceedings of the 2nd International Conference on Green Energy and Sustainable Development (GESD 2019), Shanghai, China, 18–20 October 2019.
12. Shen, F.Y.; Wang, Y.F.; Liu, C. Change Detection in SAR Images Based on Improved Non-subsampled Shearlet Transform and Multi-scale Feature Fusion CNN. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1. [CrossRef]
13. An, F.P.; Ma, X.M.; Bai, L. Image fusion algorithm based on unsupervised deep learning-optimized sparse representation. *Biomed. Signal Process. Control.* **2022**, *71*, 103140. [CrossRef]
14. Ma, X.L.; Hu, S.H.; Yang, D.S. SAR Image De-noising Based on Residual Image Fusion and Sparse Representation. *KSII Trans. Internet Inf. Syst.* **2019**, *13*, 3620–3637.
15. Bai, L.; Yao, S.L.; Gao, K.; Huang, Y.J.; Tang, R.J.; Yan, H.; Max, Q.-H.M.; Ren, H.L. Joint Sparse Representations and Coupled Dictionary Learning in Multi-Source Heterogeneous Image Pseudo-color Fusion. *IEEE Sens. J.* **2023**, *23*, 1. [CrossRef]
16. Wang, J.W.; Qu, H.J.; Zhang, Z.H.; Xie, M. New insights into multi-focus image fusion: A fusion method based on multi-dictionary linear sparse representation and region fusion model. *Inf. Fusion* **2024**, *105*, 102230. [CrossRef]
17. Wang, H.Z.; Shu, C.; Li, X.F.; Fu, Y.; Fu, Z.Z.; Yin, X.F. Two-Stream Edge-Aware Network for Infrared and Visible Image Fusion With Multi-Level Wavelet Decomposition. *IEEE Access* **2024**, *12*, 22190–22204. [CrossRef]
18. Zhang, T.T.; Du, H.Q.; Xie, M. W-shaped network: A lightweight network for real-time infrared and visible image fusion. *J. Electron. Imaging* **2023**, *32*, 63005. [CrossRef]
19. Luo, J.H.; Zhou, F.; Yang, J.; Xing, M.D. DAFCNN: A Dual-Channel Feature Extraction and Attention Feature Fusion Convolution Neural Network for SAR Image and MS Image Fusion. *Remote Sens.* **2023**, *15*, 3091. [CrossRef]
20. Deng, B.; Lv, H. Research on Image Fusion Method of SAR and Visible Image Based on CNN. In Proceedings of the 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Dali, China, 12–14 October 2022.
21. Kong, Y.Y.; Hong, F.; Leung, H.; Peng, X.Y. A Fusion Method of Optical Image and SAR Image Based on Dense-UGAN and Gram–Schmidt Transformation. *Remote Sens.* **2021**, *13*, 4274. [CrossRef]
22. Li, D.H.; Liu, J.; Liu, F.; Zhang, W.H.; Zhang, A.D.; Gao, W.F.; Shi, J. A Dual-fusion Semantic Segmentation Framework with GAN For SAR Images. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022.
23. Ma, C.H.; Gao, H.C. A GAN based method for SAR and optical images fusion. In Proceedings of the Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing (APCOM and YSAOM 2021), Shanghai, China, 28–31 October 2022.

24. Liang, J.Y.; Cao, J.Z.; Sun, G.L.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
26. Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; Han, S. Lite Transformer with Long-Short Range Attention. *arXiv* **2020**, arXiv:2004.11886. [CrossRef]
27. Li, B.; Hu, Y.H.; Nie, X.C.; Han, C.Y.; Jiang, X.J.; Guo, T.D.; Liu, L.Q. DropKey for Vision Transformer. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–20 June 2023.
28. Liu, Y.; Chen, X.; Peng, H.; Wang, Z.F. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]
29. Li, H.; Wu, X.J.; Kittler, J. Infrared and Visible Image Fusion using a Deep Learning Framework. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
30. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z.F. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolut. Inf. Process.* **2018**, *16*, 1. [CrossRef]
31. Di, J.; Ren, L.; Liu, J.Z.; Guo, W.Q.; Zhang, H.K.; Liu, Q.D.; Lian, J. FDNet: An end-to-end fusion decomposition network for infrared and visible images. *PLoS ONE* **2023**, *18*, e0290231. [CrossRef]
32. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
33. Bai, Z.X.; Zhu, R.G.; He, D.Y.; Wang, S.C.; Huang, Z.T. Adulteration Detection of Pork in Mutton Using Smart Phone with the CBAM-Invert-ResNet and Multiple Parts Feature Fusion. *Foods* **2023**, *12*, 3594. [CrossRef] [PubMed]
34. Wang, S.H.; Fernandes, S.; Zhu, Z.Q.; Zhang, Y.D. AVNC: Attention-based VGG-style network for COVID-19 diagnosis by CBAM. *IEEE Sens. J.* **2021**, *22*, 1. [CrossRef] [PubMed]
35. Jia, J.H.; Qin, L.L.; Lei, R.F. Im5C-DSCGA: A Proposed Hybrid Framework Based on Improved DenseNet and Attention Mechanisms for Identifying 5-methylcytosine Sites in Human RNA. *Front. Biosci.* **2023**, *28*, 346. [CrossRef] [PubMed]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is all You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
37. Wang, W.H.; Xie, E.Z.; Li, X.; Fan, D.P.; Song, K.T.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
38. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
39. Zhao, Z.X.; Bai, H.W.; Zhang, J.S.; Zhang, Y.L.; Xu, S.; Lin, Z.D.; Timofte, R.; Van Gool, L. CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
40. Wang, C.; Ruan, R.; Zhao, Z.C.; Li, C.L.; Tang, J. Category-oriented Localization Distillation for SAR Object Detection and A Unified Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1. [CrossRef]
41. Schmitt, M.; Hughes, L.H.; Zhu, X.X. The SEN1–2 dataset for deep learning in SAR-optical data fusion. *arXiv* **2018**, arXiv:1807.01569. [CrossRef]
42. Zhang, X.; Ye, P.; Xiao, G. VIFB: A Visible and Infrared Image Fusion Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; Shanghai Jiao Tong University, School of Aeronautics and Astronautics: Shanghai, China, 2020.
43. Li, H.; Wu, X.J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef] [PubMed]
44. 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]
45. Tang, L.F.; Yuan, J.T.; Ma, J.Y. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
46. Wang, Z.S.; Chen, Y.L.; Shao, W.Y.; Li, H.; Zhang, L. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
47. Tang, W.; He, F.Z.; Liu, Y. YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer. *IEEE Trans. Multimed.* **2023**, *25*, 5413–5428. [CrossRef]