remote sensing

MDPI

Article

HATF: Multi-Modal Feature Learning for Infrared and Visible Image Fusion via Hybrid Attention Transformer

Xiangzeng Liu [1], Ziyao Wang [1], Haojie Gao [1], Xiang Li [2], Lei Wang [2] and Qiguang Miao [1,*]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; xzliu@xidian.edu.cn (X.L.); zywang_6@stu.xidian.edu.cn (Z.W.); gaohj@stu.xidian.edu.cn (H.G.)
[2] NavInfo Co., Ltd., Beijing 100094, China; lixiang10081@navinfo.com (X.L.); wanglei4233@navinfo.com (L.W.)
* Correspondence: qgmiao@xidian.edu.cn

Abstract: Current CNN-based methods for infrared and visible image fusion are limited by the low discrimination of extracted structural features, the adoption of uniform loss functions, and the lack of inter-modal feature interaction, which make it difficult to obtain optimal fusion results. To alleviate the above problems, a framework for multimodal feature learning fusion using a cross-attention Transformer is proposed. To extract rich structural features at different scales, residual U-Nets with mixed receptive fields are adopted to capture salient object information at various granularities. Then, a hybrid attention fusion strategy is employed to integrate the complementing information from the input images. Finally, adaptive loss functions are designed to achieve optimal fusion results for different modal features. The fusion framework proposed in this study is thoroughly evaluated using the TNO, FLIR, and LLVIP datasets, encompassing diverse scenes and varying illumination conditions. In the comparative experiments, HATF achieved competitive results on three datasets, with EN, SD, MI, and SSIM metrics reaching the best performance on the TNO dataset, surpassing the second-best method by 2.3%, 18.8%, 4.2%, and 2.2%, respectively. These results validate the effectiveness of the proposed method in terms of both robustness and image fusion quality compared to several popular methods.

Keywords: transformer; U-Net; image fusion; inter-domain interaction

Citation: Liu, X.; Wang, Z.; Gao, H.; Li, X.; Wang, L.; Miao, Q. HATF: Multi-Modal Feature Learning for Infrared and Visible Image Fusion via Hybrid Attention Transformer. *Remote Sens.* **2024**, *16*, 803. https://doi.org/10.3390/rs16050803

Academic Editors: Qian Du, Jun Zhou, Danfeng Hong and Chenhong Sui

Received: 29 December 2023
Revised: 22 February 2024
Accepted: 23 February 2024
Published: 25 February 2024

1. Introduction

The progression of information technology has led to the gradual transformation of scene detection from single-sensor to multi-sensor systems. This transition is necessitated by the images acquired by single sensors being inadequate to effectively address the requirements of practical tasks. Image fusion is a crucial technology for achieving multi-sensor information fusion [1]. In comparison to a single image, a fused image incorporates abundant and complementary information from different sensors. This enhanced amalgamation is particularly beneficial for environmental perception and facilitates the processing of high-level visual tasks. Typically, common image fusion techniques include multi-focus fusion and multi-sensor fusion. As an important branch of multi-sensor fusion, infrared and visible image fusion is widely applied in the computer vision field, such as in target detection [2], video security surveillance [3], remote sensing image processing [4–7], and military reconnaissance [8].

Due to the different imaging mechanisms used, the difference between images captured by various imaging systems is huge in terms of contrast, color, and texture. Visible images have a high spatial resolution, rich color, and detailed textures; however, their imaging quality is dependent on illumination conditions. Therefore, in low light or adverse weather conditions, the use of visible imaging is not ideal. In comparison to visible imaging, infrared imaging captures the thermal radiation information of the target and provides good discrimination between warmer targets and backgrounds. Consequently, infrared

imaging is well-suited for use in challenging conditions, such as weak light environments, strong winds, dense fog, rain, and snow. Moreover, it remains relatively unaffected by light and weather interference, making it a reliable choice for such situations. However, infrared imagery is not well suited to human visual perception due to its low spatial resolution, limited textural detail, and poor clarity. Therefore, fusion of the images from these two types of sensors may help to achieve a more adaptive scene perception of the environment.

Over the past decade, extensive research in image fusion has yielded numerous methods, broadly categorized into traditional and deep learning-based approaches. Traditional methods, like multi-scale transform (MST) [9–11], sparse representation (SR) [12,13], low-rank representation [14–16], and saliency-based approaches [17,18], employ various techniques for fusion. However, they suffer from drawbacks such as operator dependency and computational intensity. In recent years, deep learning (DL) has emerged as a superior alternative, offering high adaptability and robustness. Depending on the training methods and network architecture, deep learning-based fusion methods can be broadly classified into two primary categories: non-end-to-end image fusion methods and end-to-end image fusion methods. In the case of non-end-to-end methods, pre-trained neural networks are commonly utilized to extract image features. Subsequently, these extracted features undergo fusion using predefined rules and are reconstructed to yield the ultimate fused image.

Unlike the non-end-to-end methods, end-to-end image fusion techniques simplify the fusion process and improve performance. Xu et al. [19] introduced FusionDN, a versatile network trained using a resilient weight-connection algorithm. Ma et al. [20] pioneered the use of generative adversarial networks (GANs) in image fusion. However, since these methods do not fully consider the illumination factor and cannot adapt to different light intensity distributions, they may sometimes produce distorted fusion results [21]. Meanwhile, although existing deep learning-based methods achieve high efficiency and good fusion quality, most of them highlight local features and lack consideration of global features during image fusion.

Despite achieving competitive performance, deep learning-based methods still exhibit certain drawbacks, including the following disadvantages:

i. Shallow features tend to have only local information and lack global information and cannot include contextual information from features at all scales;
ii. In the feature fusion stage, convolutional layers are initially employed to integrate the features, followed by their fusion. This process involves local and global interactions solely within the domain, with no cross-domain contextual interactions being executed.

To address the above drawbacks, we propose a hybrid attention Transformer fusion model (HATF). The main contributions of the proposed method can be summarized as follows:

i. A residual U-Net block (RUB) is utilized in each encoding block. The integration of a U-shaped structure nested within the RUB allows the network to capture richer local and global information simultaneously across all scales.
ii. Hybrid attention mechanisms are constructed within and between domains. Intra-domain self-attention is first adopted to extract global information from single-mode images. Subsequently, inter-domain cross-attention is applied to obtain interaction information of the dual-mode images. With the hybrid attention mechanism, the complementary information of the infrared and visible features is seamlessly integrated to yield a more informative fused image.
iii. We designed an adaptive fusion loss function based on different modal features, combined with a saliency loss, to achieve high-quality fusion of infrared and visible light images. Extensive experiments conclusively demonstrate the effectiveness of the proposed approach in significantly improving the quality of the fused images. Compared to state-of-the-art fusion methods, the proposed approach exhibits superior fusion performance.

The subsequent sections of this paper are organized as follows: Section 2 offers an extensive review of related work on visible and infrared image fusion, emphasizing the development and advantages of the Transformer. Section 3 provides a detailed description of HATF. Comparative experiments are conducted and their results analyzed in Section 4. The paper concludes with a summary in Section 5.

2. Related Works

This section presents the related work on deep learning-based image fusion methods and details the two networks used in the proposed model: the U-Net as the feature extraction backbone and the Transformer as the key module for feature filtering and fusion.

2.1. Deep Learning-Based Image Fusion Methods

The widespread adoption of deep learning (DL) in image fusion has demonstrated enhanced performance compared to traditional methods. This can be attributed to its inherent advantages, including its high adaptability, error tolerance, and noise resistance. Therefore, an enormous number of works have focused on using DL methods to solve image fusion tasks. In the case of non-end-to-end methods, Liu et al. [22] employed pre-trained Siamese neural networks to extract features and calculate fusion weights, which, in combination with image pyramids, were able to achieve both activity level measurement and weight calculation. In 2018, Li et al. [23] proposed DenseFuse, consisting of an encoder, a fusion strategy, and a decoder; with the help of dense convolutional blocks, the network was able to better extract and fuse deep features. To enhance DenseFuse, Li et al. [24] introduced Nestfuse by substituting the encoder with a multi-scale network and opting for a nested connection network as the decoder, and this modification was proven to be more effective in fusing both background details and salient regions within the image. To maximize the utilization of extracted information, they also introduced RFN-nest [25], which incorporates a residual fusion network in order to learn strategies that exhibit enhanced robustness and generalization capabilities.

Different from the above methods, the end-to-end image fusion method generates fusion images without the need for complex and time-consuming operations, thus improving fusion performance. Xu et al. [19] developed a general and efficient image fusion network, FusionDN, which is trained using an elastic weight joining algorithm and is highly versatile at handling multiple fusion tasks. In 2019, Ma et al. [20] pioneered the incorporation of generative adversarial networks (GANs) into image fusion systems. Tang et al. [21] introduced a progressive image fusion network called PIAFusion that considers the illumination factor during the feature extraction stage. As Transformer has made its mark in computer vision tasks, a number of works have deployed Transformer in image fusion methods. Chen et al. [26] proposed an end-to-end framework integrating transformer and a hybrid feature extractor to compensate for CNN's limitations in capturing long-range dependencies, improving fusion performance. The DATFuse model developed by Tang et al. [27] contains a dual-attention residual module for feature extraction and a Transformer module for capturing the global context, effectively preserving long-range dependencies. CMTFusion [28] captures global interactions and preserves complementary information by utilizing a cross-modal transformer (CMT) and demonstrates its utility in object detection and monocular depth estimation.

2.2. U-Net for Feature Extraction

U-Net [29] was proposed in 2015 and has gained significant popularity in various semantic segmentation tasks, including industrial fault recognition [30], satellite image segmentation [31,32], and medical image segmentation [33–35]. By modifying and extending the architecture of the full convolutional network (FCN), U-Net can generate more accurate segmentation results, especially when dealing with a limited number of training images. An important improvement to the U-Net system is that the transfer of contextual information from low to high levels is accomplished via cascade upsampling. U-Net ex-

hibits a U-shaped structure, as depicted in Figure 1. This structure is characterized by the symmetrical arrangement of the expansion path and the systolic path. The left systolic path is responsible for feature extraction, while the right extension path focuses on feature reconstruction. Each block of the systolic path contains two convolutional layers. To avoid gradient explosion, a modified linear unit (ReLU) is connected after each convolutional layer. Subsequently, a maximum pooling layer is performed for downsampling.

In recent years, numerous researchers have made modifications to the network structure and connectivity of the U-Net to further enhance its feature extraction capabilities. U-Net++ [36] greatly improves the efficiency of information transfer and feature reconstruction through nest connection. In the U-Net architecture, shallow convolutions primarily emphasize local texture features, while deep convolutions focus on semantic features. In contrast, U2-Net [37] incorporates both local and global intra-stage features without compromising the resolution of the feature map, which is achieved through the nesting of U-Net. With the advantage of the residual U-shaped network blocks in U2-Net, we are able to obtain more global information from the shallow high-resolution feature maps. Therefore, in our proposed method, we utilize U2-Net as the backbone network for extracting multimodal features.

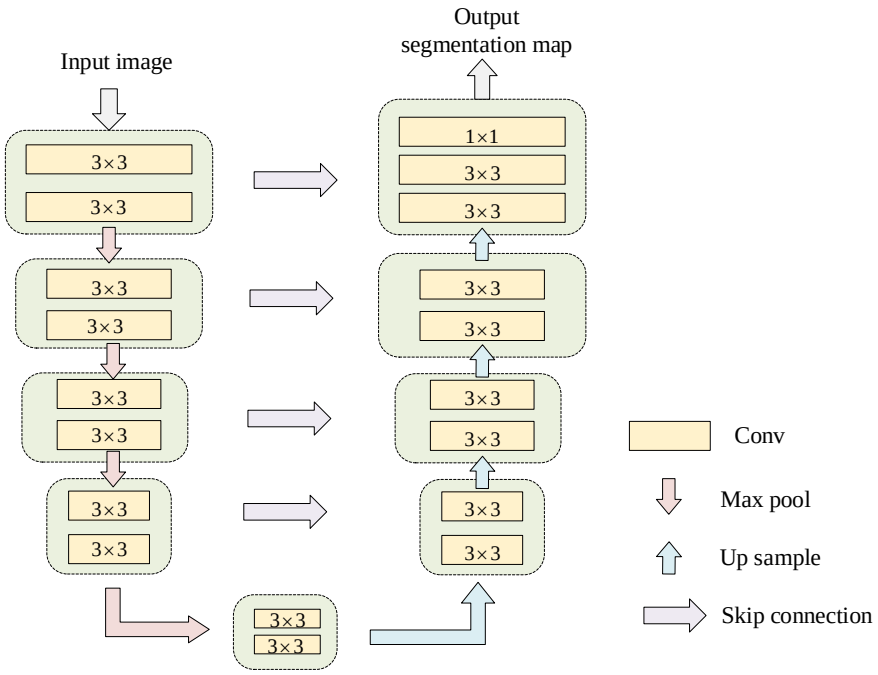

Figure 1. The architecture of U-Net.

2.3. Transformer for Fusion

The Transformer model initially achieved significant advancements in the field of natural language processing (NLP). The introduction of the ViT [38] demonstrated the immense potential of Transformer in computer vision tasks. Unlike CNNs, which focus on local features, the attention mechanism employed by the Transformer enables the modeling of long-range dependencies, facilitating the integration of global information across different scales. In recent years, the Transformer architecture has seen increased application in computer vision tasks, including in target detection, segmentation, and multi-object tracking. This highlights its effectiveness in capturing long-range dependencies and modeling global contexts. Liu et al. [39] introduced VST, a transformer-based dense prediction model, which incorporates task-related tokens and a patch task attention mechanism, presenting a novel paradigm for transformer-based models in dense prediction tasks.

Although Transformer has good representational capabilities, it is computationally expensive for high-resolution images. A number of studies have begun to explore more

suitable transformer architectures for computer vision tasks. Liu et al. [40] introduced the hierarchical Swin-Transformer, which implements feature characterization through shifted windows. It enhances efficiency by confining the computation of self-attention to non-overlapping local windows and enabling cross-window connections. The hierarchical structure is adaptive to modeling at different scales and exhibits linear computational complexity in relation to image size. To further improve the efficiency of model information transfer, we construct a hybrid attention mechanism to fuse complementary information between domains, resulting in an information-rich fused image.

3. Methodology

The fusion network incorporating the hybrid attention mechanism is described in detail in this section. Section 3.1 provides an introduction to the architecture of the fusion framework. Then, the residual U-Net feature extraction block is built in Section 3.2, and the Transformer module with hybrid attention is constructed in Section 3.3. Finally, the loss function and the training strategy are presented in Section 3.4.

3.1. Overall Network Structure

The HATF model is composed of three pivotal components: an encoder, a fusion module, and a decoder. The overall model structure is depicted in Figure 2, and detailed descriptions of each component as given below:

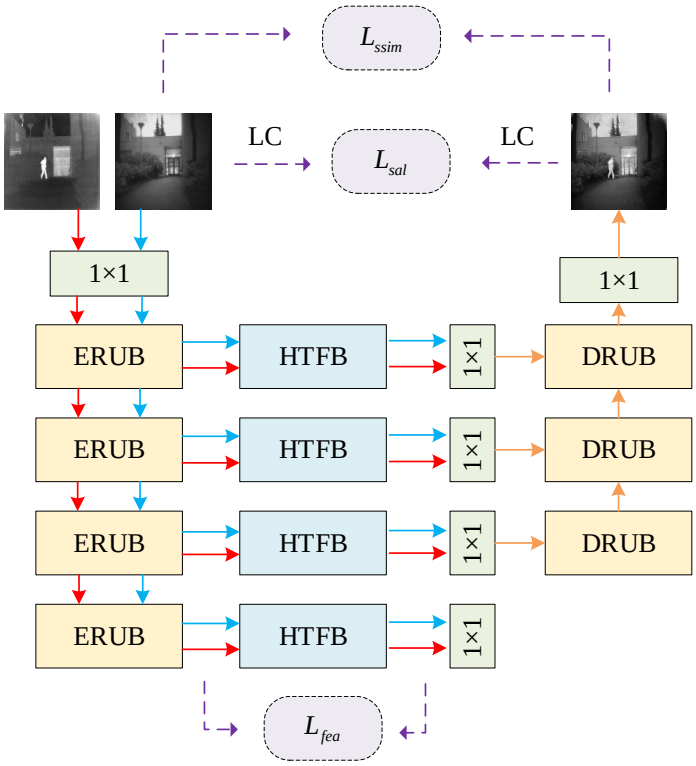

Figure 2. Network structure of HATF. The blue arrows and red arrows represent the branches of the visible and infrared images, respectively, and the yellow arrows represent the fused branches. The encoder is composed of 1 convolutional layer and 4 ERUBs. Subsequently, the features at different scales are processed through the HTFBs to obtain fusion features. Finally, the decoder is made up of 1 convolutional layer and 3 DRUBs.

(1) Encoder: The encoder adopts U2-Net as the backbone network for extracting multimodal features, which contain structural information at different granularities from shallow to deep. The multi-scale encoder is composed of 1 convolutional layer and 4 en-

coder residual U-Net blocks (ERUBs). To diminish the spatial resolution of the features, a max pooling layer is inserted between each ERUB.

(2) Hybrid Attention Transformer Fusion Block (HTFB): This block is constructed on Swin-Transformer with a hybrid attention mechanism. The hybrid attention mechanism is constructed from intra-domain self-attention units and inter-domain cross-attention units, which enables long-range dependency modeling and global interaction of features. The fusion features are obtained from the HTFBs, which are able to fully retain complementary infrared and visible information and better integrate multimodal features.

(3) Decoder: The decoder is made up of one convolutional layer and three decoder U-Net blocks (DRUBs). It reconstructs fusion features to generate the fused image. With the help of DRUBs, the reconstruction ability of the model for global information has been improved.

3.2. RUB-Based Feature Extraction

Feature extraction that encompasses both local and global information is crucial for image fusion. In the design of CNNs, the convolutional kernel size is typically 1×1 or 3×3, leading to a perceptual field that is too small to capture global information. On the other hand, the residual U-Net block (RUB) is adapted to extract global features across all scales. Therefore, we employ RUBs to extract more global information from the shallow, high-resolution feature maps. The residual U-Net block can be defined as $RUB - L(C_{in}, M, C_{out})$, where L is the number of U-Net layers; C_{in} and C_{out} represent the input and output channels, respectively; and M represents the number of internal channels. The composition of the RUB, as shown in Figure 3, can thus be described in three parts:

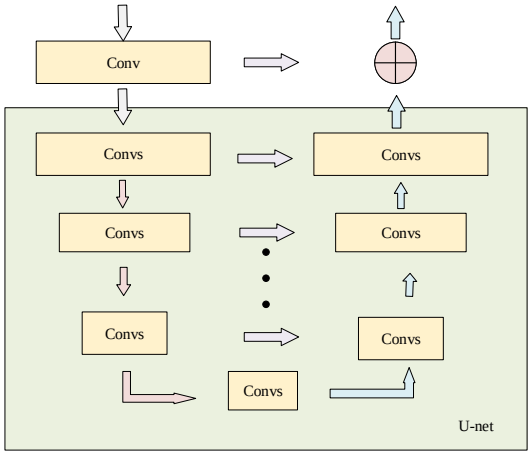

Figure 3. The structure of RUB. The input layer extracts and transforms the local features. After that, a U-Net encoder–decoder structure encodes multi-scale contextual information from input features and acquires decoder features. Additionally, the residual connection combines the local features with multi-scale depth features.

(1) Input layer: It conducts local feature extraction and transforms the input feature map $x(H \times W \times C_{in})$ into an intermediate feature $F_1(x)$ with C_{out} channels. The convolution kernel size is 3×3, and the activation function used is ReLU.

(2) U-Net encoder–decoder: Similar to U-Net, this structure encodes multi-scale context information from $F_1(x)$ and acquires decoder features $U(F_1(x))$, where U represents the U-Net-like structure. A larger L implies a deeper residual U-Net module. The module initially extracts multi-scale features from multiple downsamples and subsequently encodes them into a high-resolution feature map through successive upsampling, concatenation, and convolution. This approach helps alleviate the loss of detail associated with direct upsampling.

(3) Residual connection: It combines local features with multi-scale depth features to achieve improved preservation of structural information:

$$G_1(x) = F_1(x) + U(F_1(x)), \tag{1}$$

where $G_1(x)$ is the output of RUB.

The comparison between normal residual blocks and RUBs is shown in Figure 4. The extraction operations in normal residual blocks usually consist of one or more convolutional layers, which can be represented as

$$G_2(x) = F_2(F_1(x)) + F_1(x), \tag{2}$$

where $G_2(x)$ is the expected output. The main difference between RUBs and normal residual blocks is the replacement of the single-flow convolution with a U-Net structure. This makes it possible for the network to extract global information directly with very low computational overhead. In view of this, we employ a RUB as an encoder for the proposed method to thoroughly extract essential features at each scale.

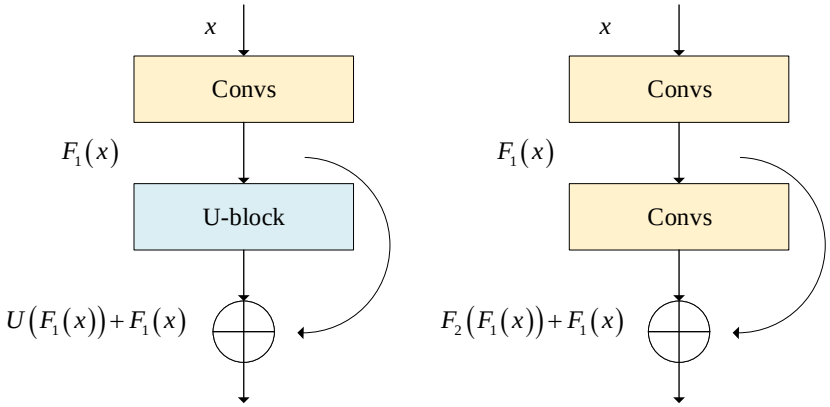

Figure 4. Comparison between RUBs (**left**) and ordinary residual blocks (**right**). In RUBs, a U-block is deployed to replace the single-flow convolution.

3.3. Hybrid Attention Transformer Fusion Block

After the extraction of multimodal features by the RUB, an HTFB is constructed to fuse these features using intra-domain self-attention and inter-domain cross-attention mechanisms. Firstly, multi-headed self-attentiveness (MSA) takes into account the global feature distribution, which helps the model capture information from multiple encoding subspaces. Then, to improve the feature tokens generated by MSA, a feedforward network (FFN) consisting of two multilayer perceptron (MLP) layers and a GELU activation layer is applied. Subsequently, layer normalization (LN) is implemented after the the MSA and FFN are run, and finally the residuals are deployed after being processed by these two modules. The process of intra-domain self-attention can be expressed as

$$\{Q, K, V\} = \{XW^Q, XW^K, XW^V\}, \tag{3}$$

$$O = Attention(Q, K, V). \tag{4}$$

Following intra-domain perception, inter-domain cross-attention interaction is implemented to further explore information common to different domains. As can be seen in Figure 5, the basic modules of inter-domain and intra-domain awareness are similar, the main difference being that inter-domain awareness adopts multi-headed cross-attention

(MCA) to achieve global content interaction. The whole process of inter-domain cross-attention interaction can be defined as below:

$$\{Q_1, K_1, V_1\} = \{X_1 W_1^Q, X_1 W_1^K, X_1 W_1^V\}, \tag{5}$$

$$\{Q_2, K_2, V_2\} = \{X_2 W_2^Q, X_2 W_2^K, X_2 W_2^V\}, \tag{6}$$

$$O_1 = Attention(Q_1, K_2, V_2), \tag{7}$$

$$O_2 = Attention(Q_2, K_1, V_1). \tag{8}$$

As shown in Formulas (5)–(8), $\{Q_1, K_2, V_2\}$ are querying features from domain2 that are similar to Q_1 in domain1, and similarly, $\{Q_2, K_1, V_1\}$ are querying features from domain1 that are similar to Q_2 in domain2, thus enabling the interaction between the two domains. Following the HTFB, a convolutional layer is adopted to integrate local information from different domains. Ultimately, the effective fusion of complementary multimodal features can be realized by cascading intra- and inter-domain hybrid attention.

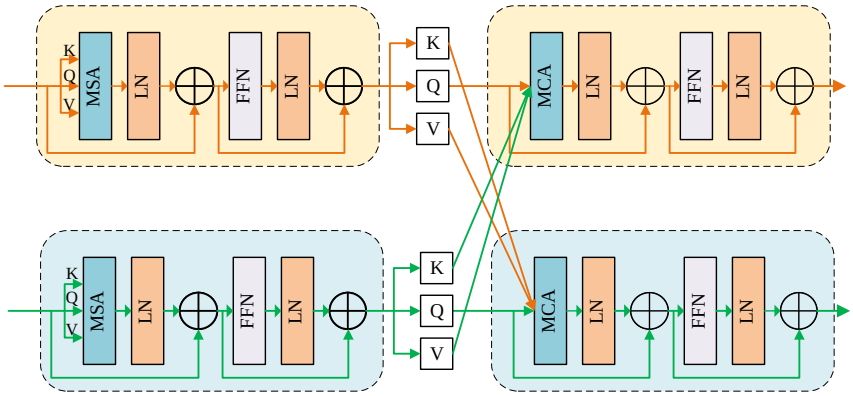

Figure 5. Structure of HTFB containing self-attention and cross-attention. HTFB first uses self-attention for intra-domain perception and then uses cross-attention for inter-domain interaction to further explore information associations across different domains.

3.4. Loss Function for HATF

In deep learning-based methods, the choice of the loss function is pivotal to fusion performance. To bolster feature extraction and fusion capabilities, we employ a two-stage training strategy and devise an adaptive loss function for the fusion of multimodal features. During the initial training stage, the encoder and decoder are directly connected to enhance feature extraction. In this stage, the loss function of the auto-encoder network (L_{econ}) comprises two terms: the content loss (L_{con}) and the structural similarity loss (L_{ssim}). L_{con} is calculated as follows:

$$L_{econ} = L_{con} + \lambda L_{ssim}, \tag{9}$$

where λ is a hyperparameter controlling the balance between the two terms. The content loss L_{con} facilitates the preservation of more details from the source image in the reconstructed image and can be expressed as follows:

$$L_{con} = ||O - I||_F^2, \tag{10}$$

where $||\cdot||$ represents the Frobenius norm, and O and I represent the output and input images, respectively. The structural similarity loss L_{ssim} is calculated to incorporate more structural features from the input images into the results and is defined as follows:

$$L_{ssim} = 1 - SSIM(I, O), \tag{11}$$

where $SSIM(\cdot)$ denotes structural similarity between two images, considering brightness, contrast, and structure. The inclusion of structural similarity loss aims to enhance the structural similarity of the results to the source images. In the second stage, HTFB undergoes training to augment the fusion feature's capability. For more effective fusion of multimodal features, an adaptive fusion loss function is devised for this stage. The total fusion loss L_{fusion} encompasses three components: structural similarity loss L_{ssim}, multimodal feature loss L_{fea}, and saliency loss L_{sal}. The comprehensive loss is calculated as follows:

$$L_{fusion} = \alpha L_{ssim} + \beta L_{fea} + L_{sal}, \tag{12}$$

where α and β are hyperparameters that balance the three terms. L_{ssim} is calculated as in Equation (11).

The multimodal deep features extracted by the encoder encompass diverse information, with shallow features being rich in detail, intermediate features representing structural information, and deep features primarily capturing regional features. To maximize the transfer of feature information to the fusion images, specific loss functions are designed for different modal features. The multimodal feature L_{fea} consists of three parts: detail features loss L_{df}, structural feature loss L_{sf}, and regional feature loss L_{rf}; they are formulated as follows:

$$L_{fea} = L_{df} + \mu L_{sf} + \rho L_{rf}, \tag{13}$$

$$L_{df} = ||\Phi_f^1 - (w_{ir}\Phi_{ir}^1 + w_{vis}\Phi_{vis}^1)||_F^2, \tag{14}$$

$$L_{sf} = 1 - \frac{cov(\Phi_f^{2,3}, (w_{ir}\Phi_{ir}^{2,3} + w_{vis}\Phi_{vis}^{2,3}))}{\sigma_{\Phi_f^{2,3}}\sigma_{w_{ir}\Phi_{ir}^{2,3}+w_{vis}\Phi_{vis}^{2,3}}}, \tag{15}$$

$$L_{rf} = ||\Phi_f^4 - (w_{ir}M_{ir}^4\Phi_{ir}^4 + w_{vis}M_{vis}^4\Phi_{vis}^4)||, \tag{16}$$

where Φ^i represents the features of each layer, μ and ρ are hyperparameters, M_{ir}^4 and M_{vis}^4 denote masks to remove noise from features, w_{ir} and w_{vis} are self-adapting weights, $cov(\cdot)$ denotes the covariance function, and σ denotes the standard deviation function. M_{ir}^4, M_{vis}^4, w_{ir}, and w_{vis} can be defined as follows:

$$M_{ir}^4 = \begin{cases} 1, \Phi_{ir}^4 \geq \theta \\ 0, \Phi_{ir}^4 < \theta \end{cases}, \tag{17}$$

$$M_{vis}^4 = \begin{cases} 1, \Phi_{vis}^4 \geq \theta \\ 0, \Phi_{vis}^4 < \theta \end{cases}, \tag{18}$$

$$w_{ir} = \frac{||\Phi_{ir}^i||_1}{||\Phi_{ir}^i||_1 + ||\Phi_{vis}^i||_1}, \tag{19}$$

$$w_{vis} = 1 - w_{ir}, \tag{20}$$

where θ is a constant set controlling the degree of noise removal.

To maintain the saliency of the thermal objects, the training process is supplemented with salient object detection information. During the fusion procedure, salient object regions are masked and an adaptive loss function is developed to drive feature extraction and reconstruction according to the masks. By selectively increasing the weights of salient objects and background textures, the fused images have high-quality texture details and clear saliency targets. Firstly, the LC [41] saliency extraction algorithm is applied to extract the saliency map from the infrared image. The saliency map is then normalized to obtain $\hat{M}_{sal}$. Finally, the saliency loss is calculated as follows:

$$L_{sal} = ||\hat{M}_{sal}F - \hat{M}_{sal}I_{ir})||_F^2 + ||(1 - \hat{M}_{sal})F - (1 - \hat{M}_{sal})I_{vis}||_F^2. \tag{21}$$

4. Experiments and Analysis

In this section, we perform comparative experiments to assess the performance of the proposed method against state-of-the-art methods. The experiments begin with an introduction to the evaluation metrics and datasets used. Subsequently, details about the training process are provided. Finally, the results of ablation experiments and comparison tests are presented to demonstrate the effectiveness of the proposed method.

Objective evaluation is employed to quantitatively assess the quality of the fusion results, aligning with the human visual system. In this section, five representative evaluation metrics are selected: entropy (EN), standard deviation (SD), mutual information (MI), structural similarity (SSIM) [42], and root mean square error (RMSE). EN and SD provide insights into the amount of information present in the fused image. MI quantifies the information conveyed from the input image to the output image. SSIM and RMSE assess the similarity of structures and the level of distortion in the fused image, respectively. The better the fusion method, the higher the values of EN, SD, MI, and SSIM, and the lower the value for RMSE.

Five datasets were employed for training and testing: [43] COCO datasets with rich scenes were selected to train the encoders. The TNO [44], FLIR, and LLVIP [45] datasets are infrared and visible image datasets containing various scenes, which are suitable for training and testing the proposed framework. In addition, supplementary experiments have been conducted on remote sense images to verify the generalizability of the proposed method. To assess the superiority of the proposed methods, we compared them with nine state-of-the-art fusion methods. These methods includes three traditional methods (DWT [46], DTCWT [47], and CVT [48]) and five deep learning methods (DenseFuse [23], FusionGAN [20], IFCNN [49], RFN-Nest [25], Swin-F [50], and MFST [51]). To be fair, the comparative experiments were implemented employing the code and parameters provided in the corresponding papers.

4.1. Training Details

The proposed method was trained using Python 3.8, and all experiments were conducted on a system equipped with an RTX 2080Ti GPU and an Intel i7-7700 CPU. The detailed settings of each module are shown in Tables 1 and 2. In the first stage, the encoder and decoder are trained by 80,000 images chosen from the MS-COCO dataset, which are first converted to greyscale and normalized to 256×256. In the second stage, CTFBs are mainly trained. One CTFB contains four Transformer modules, two for intra-domain self-attention and the other two for inter-domain cross-attention. In Transformer, both the partition window size and the number of heads are set to eight. In this stage, 12,000 pairs of images from the LLVIP dataset are selected for training, which are converted to greyscale and resized to 256×256. In addition, we give information on the computational and parametric quantities of HATF in Table 3.

Table 1. Configurations of encoder and decoder.

	Layer	Input Channel	Output Channel	Depth	Activation Function
Encoder	ERUB1	1	64	7	ReLU
	ERUB2	64	112	6	ReLU
	ERUB3	112	160	5	ReLU
	ERUB4	160	208	4	ReLU
Decoder	DRUB1	176	64	5	ReLU
	DRUB2	272	112	6	ReLU
	DRUB3	368	160	7	ReLU

Table 2. Configurations of HTFBs.

Layer	Number of Transformer	Channel	Window Size	Number of Heads
HTFB1	4	64	8	8
HTFB2	4	112	8	8
HTFB3	4	160	8	8
HTFB4	4	208	8	8

Table 3. Computational and parametric information of HTFBs.

	Input Size	FLOPs	Params
HATF	(320, 320)	44.79 G	5.19 M

4.2. Ablation Study

To validate the feasibility of the model's design, two ablation experiments are implemented in this section. First, the importance of RUBs in extracting multimodal features is discussed. Then, the effectiveness of the hybrid attention Transformer for feature fusion is verified.

4.2.1. Effect of RUBs on Extracting Multimodal Features

To confirm the effectiveness of the RUBs for feature extraction, the image reconstruction experiment was performed on the encoder. The performance of feature extraction and fusion on the TNO dataset was compared, with 2convs and RUBs as the encoders, respectively. As can be seen from Table 4, both models have advantages in the reconstruction of infrared and visible images, respectively, while the average error of the RUBs is smaller than that of 2convs, indicating that the RUBs can transfer more information during the fusion process, leading to an increase in the image quality.

Table 4. RMSE between the reconstructed images (2convs, RUBs) and source images. A smaller RMSE means a smaller deviation between the reconstructed image and the input image. Smaller values are marked in bold.

	Infrared Image	Visible Image	Avg
2convs	**5.1041**	8.2931	6.6986
RUB	5.5427	**7.2613**	**6.4020**

To provide a more intuitive understanding of the effectiveness of RUBs, the multimodal features extracted by the two networks are visualized in Figure 6. For the infrared image, the shallow features extracted by the RUBs have more edge features, and the human target is more distinct when compared to the background. The RUBs effectively integrate local and global features in the shallow feature maps of visible images, showcasing their ability to capture clear structural information.

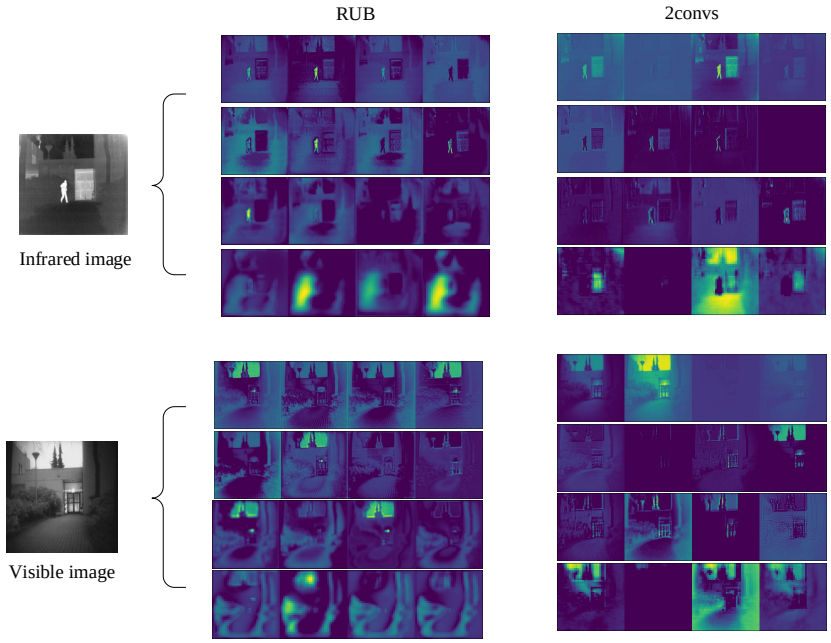

Figure 6. Visual comparison of extracted features between 2convs (**right**) and RUBs (**left**). In RUBs, the edges of the shallow features are more distinct and human targets are more visible.

Finally, to indicate the role of RUBs in the overall fusion process, the 2convs and the RUBs encoders are separately connected to a hybrid attention Transformer to construct the fusion network. The same loss function and datasets are adopted to train the two networks, and the results of their comparison are depicted in Figure 7. It is clear that the sky in Figure 7c is unnatural and artifacts exist, whereas in Figure 7d, the overall scene is blended more naturally. Figure 7g shows a higher overall brightness but insufficient resolution. In Figure 7h, the RUBs effectively combine shallow local features with deep features, improve the transmission efficiency of important information, and suppress noise to avoid artifacts. Table 5 presents the objective metric comparison results for the two encoders. The RUBs achieve the best performance, demonstrating that they are superior to 2convs in terms of maintaining the richness of an image's content and improving the visual quality of fused images.

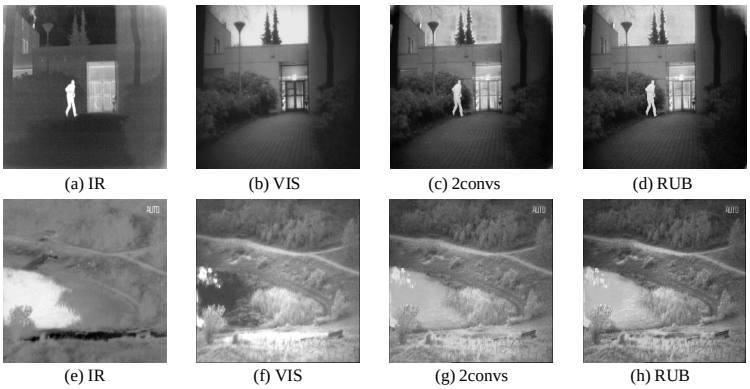

Figure 7. Comparison of fusion results obtained by 2convs and RUBs.

Table 5. Objective evaluation comparison of fusion results between 2convs and RUBs. The better values are marked in bold.

	EN	SD	MI	SSIM	RMSE
2convs	7.0107	43.7761	2.6785	0.8154	10.3867
RUB	**7.1107**	**46.7761**	**3.3608**	**0.8208**	**9.8033**

4.2.2. The Impact of HTFBs on Feature Fusion

To check the efficiency of HTFBs, the fusion results of the HTFBs and the ordinary Swin-Transformer (swin-T) are compared, which can be seen in Figure 8. For the first group of images, both fusion modules are able to produce satisfactory fusion results. In the case of the second group of images, the image fused by swin-T is relatively blurry, with glowing branches in the air, while the image obtained by the HTFBs has clearer door frames and renders people without artifacts. These results indicate that the hybrid attention mechanism can effectively integrate multimodal features, enhancing the complementarity of information. Table 6 presents the comparative results of objective evaluation on the TNO dataset. As we can see, the HTFBs are superior to swin-T in five metrics, which is attributed to the fact that the hybrid attention Transformer not only improves the fusion of intra-domain local features but also the fusion of inter-domain complementary features.

Table 6. Objective evaluation comparison of the fusion results between Swin-T and HTFB. The better values are marked in bold.

	EN	SD	MI	SSIM	RMSE
Swin-T	**7.1352**	44.1523	2.7125	0.8054	10.0157
HTFB	7.1107	**46.7761**	**3.3608**	**0.8208**	**9.8033**

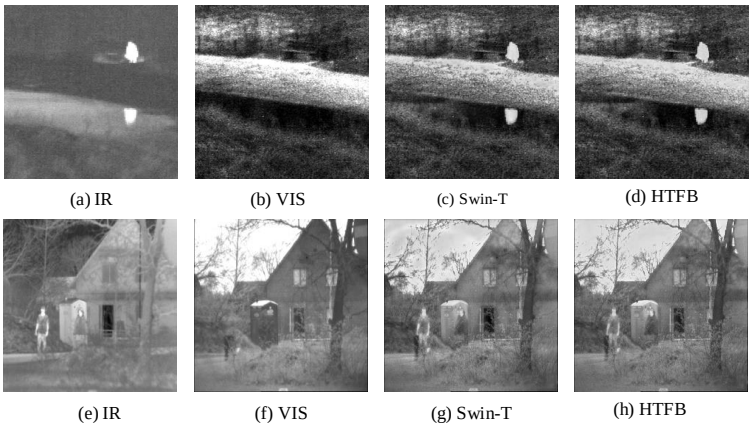

Figure 8. Comparison of fusion results between Swin-T and HTFBs.

4.3. Comparative Experiments and Analysis

To offer a comprehensive evaluation of the proposed method, nine different methods are compared based on five evaluation metrics using the TNO, FLIR, and LLVIP datasets. To further validate the generalization of the method, we tested it on remote sensing images and provided visualization results.

4.3.1. Results on TNO Dataset

To assess the performance of various methods in fusing infrared and visible images, we initially focused on the TNO dataset, which is a well-established benchmark used for evaluations. Figure 9 displays the fusion results of six image pairs using the proposed method as well as nine other methods. Among them, DWT, CVT, and DTCWT can effectively preserve the content of both the infrared and visible images. However, their fusion results suffer from blurred textures because the weighted averaging strategy leads to the loss of detailed information. Although FusionGan mainly preserves the pixel distribution in infrared images, it ignores the texture details in visible images. RFN-Nest does a good job of preserving the texture information and the greyscale distribution in the visible images, while also making thermal objects appear blurred. The results of DenseFuse are relatively low in brightness, while those of IFCNN are relatively high in brightness. The fused images of swin-F and MFST are visually good, but their contrast is low. Compared to the above methods, HATF achieved higher contrast and better visual results and is able to balance important object and background texture information well.

Table 7 showcases the quantitative comparison results among the relevant methods. The best values, second-best values, and third-best values are highlighted in bold, red, and blue, respectively. Overall, deep learning-based methods outperform traditional methods in all metrics, mainly because neural networks transmit information significantly more efficiently than traditional methods. Among them, Swin-F and MFST exhibit significant advantages, which can be attributed to the adoption of Transformer as the feature fusion module. This allows for enhanced fusion efficiency and better utilization of complementary information, leading to improved performance. The proposed method outperforms other methods in four metrics (EN, SD, SSIM, MI), and achieves sub-optimal performance on RMSE. The comparative results reveal that the proposed method generates fused images with enhanced information and textural detail and improved visual effects.

Table 7. Quantitative evaluation between HATF and related methods on TNO datasets. The first, second, and third best values are marked in bold, red, and blue, respectively.

	EN	**SD**	**MI**	**SSIM**	**RMSE**
DWT	6.5964	29.6984	0.6745	2.0510	10.2507
CVT	6.5371	28.1056	0.7149	1.8108	10.2445
DTCWT	6.4773	27.4436	0.7237	1.9163	10.2514
DenseFuse	6.7378	34.7623	0.7001	2.4726	10.2377
FusionGan	6.4919	27.9282	0.5140	2.3137	10.2673
IFCNN	6.6265	31.869	0.7155	2.5111	9.9390
RFN-Nest	6.9271	37.7383	0.7151	2.3238	10.2609
Swin-F	6.6041	37.8287	0.7847	3.2881	**8.9262**
MFST	6.9519	39.3726	0.7466	2.7028	10.1066
HATF	**7.1107**	**46.7761**	**0.8208**	**3.3608**	9.8033

Figure 9. Subjective comparison between HATF and related methods on TNO dataset.

4.3.2. Results on FLIR Dataset

To confirm the fusion capability of HATF for complicated scenes, further experiments were performed on the FLIR dataset. Intuitive comparison results of night and day scenes are presented in Figures 10 and 11. As can be seen in Figure 10, the results obtained by DWT, CVT, DTCWT, and DenseFuse preserve primary pixel distribution well; nevertheless, they lack the detailed information contained in visible images. FusionGan and RFN-Nest do not perform well on night scenes and produce results with poor visual effects. It is observed that IFCNN preserves infrared details well; however, there are significant contrast differences compared to the original image. Swin-F and MFST demonstrate strong performances on the FLIR dataset, producing fused images with enhanced backgrounds and salient object edges. In contrast to those methods, the proposed method does not only maintains the rich background texture and clear edge contours but also integrates salient thermal objects to produce images suitable for human visual perception. For the daytime scene (Figure 11), our fusion framework also produces a high-contrast image by achieving a balance between infrared and visible information.

The quantitative evaluation of the fusion capability of the ten methods on the FLIR dataset is presented in Table 8. From the table, it can be observed that the proposed method achieved the highest MI, the second-best results in terms of SSIM and SD, and the third-best results in terms of EN. RFN-Nest mainly focuses on constraining the pixel values between the fusion image and the input images, considering mainly local information during the fusion, which results in improved fusion performance in EN and SD. Our method is comparable to SWIN-F on SSIM and has significant advantages on EN, SD, and MI. Compared to MFST, the proposed method has advantages in all metrics except for a slightly lower EN and RMSE. In general, the proposed method excels in fusing images comprehensively, considering both structural features and saliency features. It performs particularly well in processing images with distinct structures and clear edges, showcasing its advantages in such scenarios.

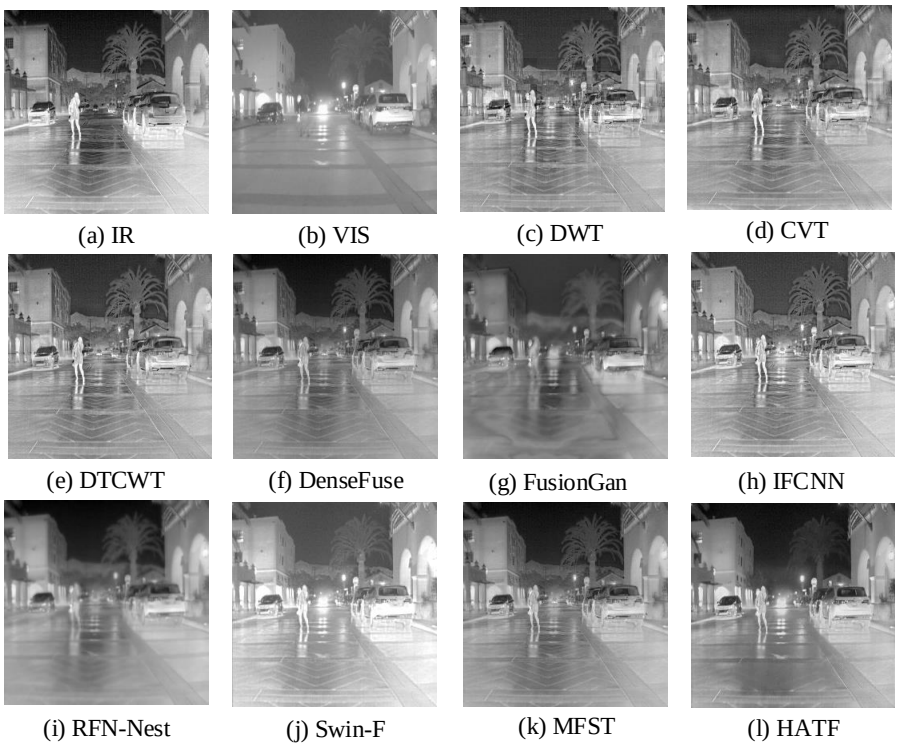

Figure 10. Subjective comparison between HATF and related methods on night scene.

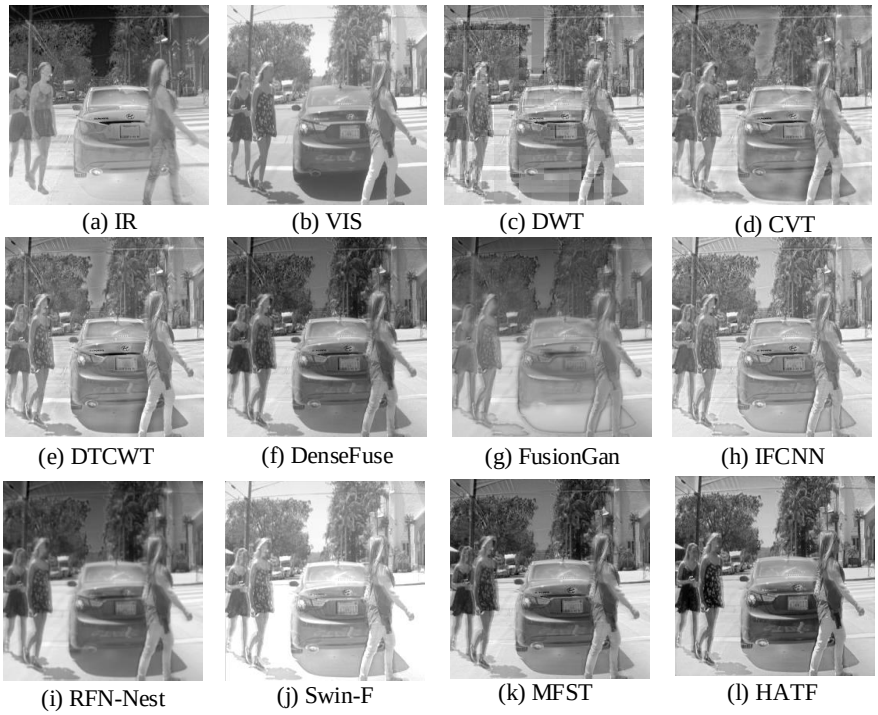

Figure 11. Subjective comparison between HATF and related methods on daytime scene.

Table 8. Quantitative evaluation comparing HATF with other methods on the FLIR dataset. The best, second-best, and third-best values are highlighted in bold, red, and blue, respectively.

	EN	SD	MI	SSIM	RMSE
DWT	7.2733	43.5689	0.5272	3.1864	10.1828
CVT	7.3018	43.8735	0.5383	2.8208	10.1606
DTCWT	7.2096	41.9806	0.5622	3.0607	10.1848
DenseFuse	7.4005	52.7755	0.6014	3.7431	10.0585
FusionGan	7.3209	47.5473	0.5947	3.3429	10.2062
IFCNN	7.1850	40.4748	0.6112	3.3764	9.6982
RFN-Nest	**7.5439**	**58.9438**	0.7006	3.5457	10.1004
Swin-F	6.0475	44.3140	**0.7715**	3.7731	**9.3039**
MFST	7.4811	52.4626	0.6435	3.7815	10.1030
HATF	7.469	52.8655	0.7651	**3.9437**	10.2133

4.3.3. Results on LLVIP Dataset

In addition, further comparison experiments were carried out on the LLVIP dataset, which was captured by surveillance cameras and has high-resolution and rich texture information. Figure 12 shows the comparison fusion results of the related methods for a typical surveillance scene with clear objects. Results obtained by traditional methods have more artifacts and noise, resulting in poor image quality. FusionGan and RFN-nest are not good enough at keeping the edges of thermal objects. Although IFCNN and Swin-F perform the fusion task better, they do not employ a multi-scale network to extract features, which results in the loss of structural information like contours. Meanwhile, the results obtained by MFST have lower contrast and less visible texture details. In comparison, the proposed method generates images with clear objects, sharp edges, and the best visual effects, demonstrating that it also provides better fusion performance on the LLVIP dataset.

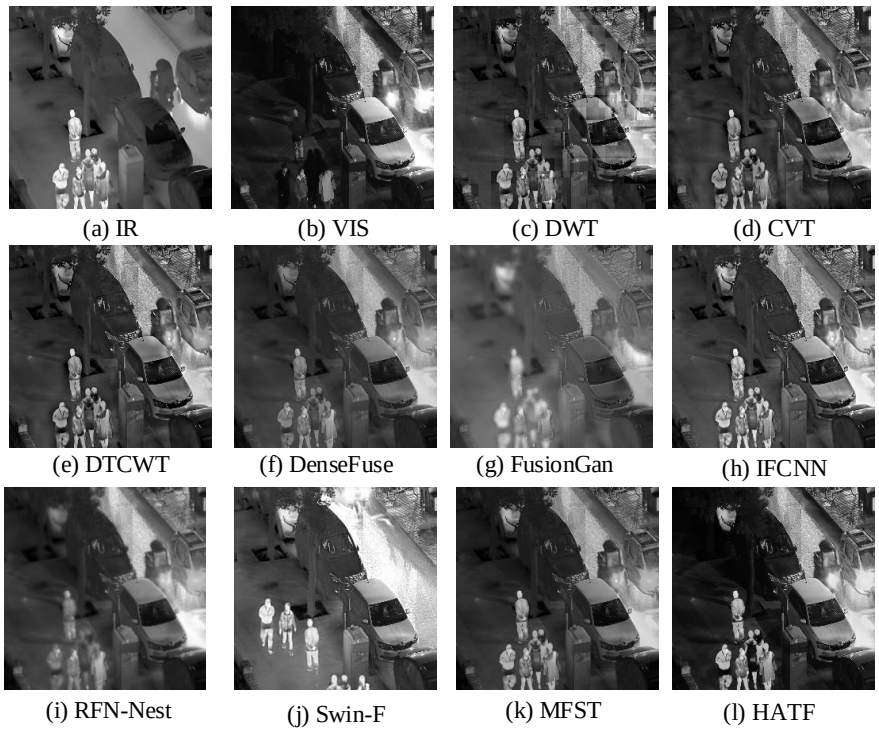

Figure 12. Subjective comparison between HATF and related methods on LLVIP dataset.

The quantitative comparison of the fusion performance for the ten methods on the LLVIP dataset is presented in Table 9. As can be seen from the table, the overall performance of IFCNN and RFN-nest are similar. Swin-F has a good performance on EN and SD and a poor performance on the other three metrics. MFST outperforms traditional methods in terms of SSIM and MI due to the more efficient information transfer in Transformer. HATF achieved optimal scores in three metrics: SSIM, MI, and RMSE. This shows that our method produces fused images of high quality on the LLVIP dataset, which can be attributed to the effective extraction and fusion of complementary features by the RUBs and HTFBs.

Table 9. Quantitative evaluation comparing HATF with other methods on the LLVIP dataset. The best, second-best, and third-best values are highlighted in bold, red, and blue, respectively.

	EN	SD	MI	SSIM	RMSE
DWT	7.1622	46.1461	0.5756	2.834	10.0145
CVT	7.1547	44.6028	0.5984	2.5841	10.0227
DTCWT	7.1415	44.4009	0.6177	2.701	10.0087
DenseFuse	7.1227	41.7651	0.6027	3.2511	9.9555
FusionGan	6.7159	30.4424	0.4887	2.7055	10.1182
IFCNN	7.3332	48.2112	0.5971	3.1225	9.7326
RFN-Nest	7.2456	45.1502	0.5533	2.8533	9.9285
Swin-F	**7.3665**	**54.6115**	0.6162	3.5557	9.3229
MFST	7.1779	46.9661	0.6875	3.2941	9.9281
HATF	7.0679	51.8918	**0.8574**	**3.8238**	**8.581**

4.4. Generalization Experiments on Remote Sensing Images

To validate the generalization of the proposed method, extensive experiments on remote sensing images are conducted in this section. Near-infrared (NIR) and visible image datasets to be fused are chosen from multimodal remote sensing images. Figure 13 showcases the remote sensing fusion images obtained by the proposed method. The images from the first to the third column represent the NIR, visible, and fused images, respectively.

The fusion results achieved by the proposed method demonstrate a notable enhancement of salient information from the NIR image while preserving color texture. Consequently, the proposed method significantly improves the information richness and visual impact of the fused image, affirming its robust generalization capability on remote sensing images.

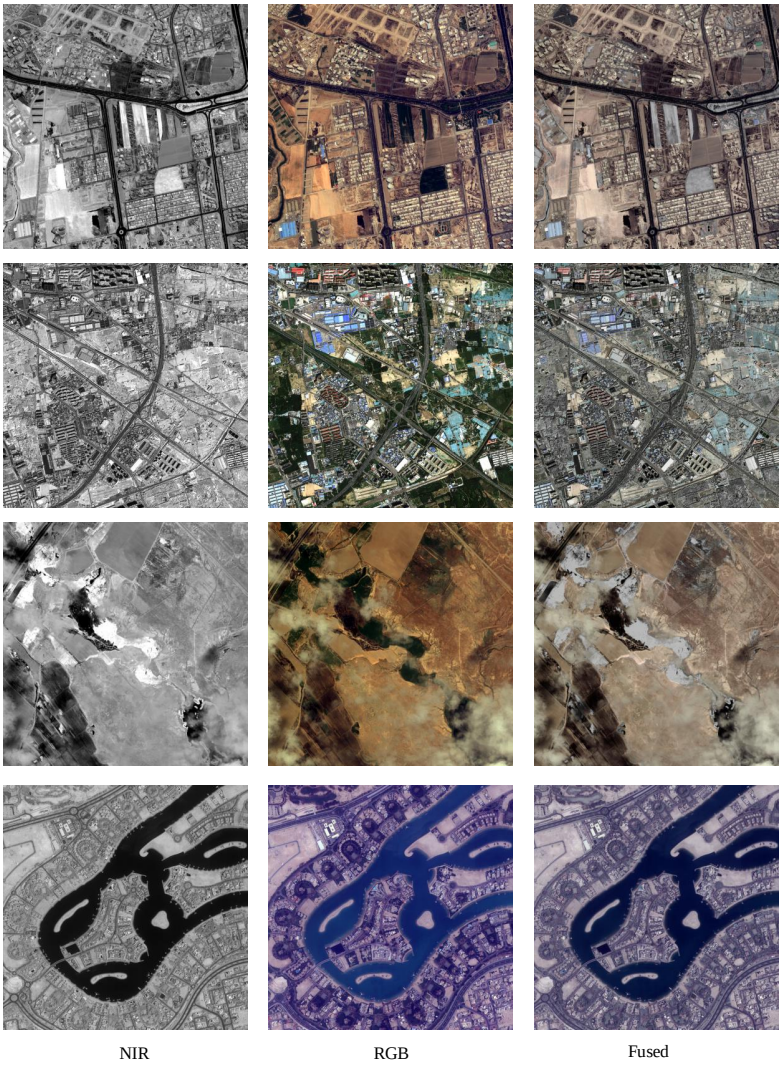

Figure 13. Generalization experimental results on remote sensing images.

5. Conclusions

In this paper, we present a hybrid attention Transformer fusion model (HATF) for infrared and visible image fusion task. The proposed method effectively improves the extraction efficiency of multimodal features and the fusion quality of images.

The innovations of our model are presented in three parts. Firstly, a residual U-Net block (RUB) is adopted to obtain more local and global information from shallow and deep layers. Secondly, the hybrid attention Transformer is constructed to fully retain complementary information and better integrate multimodal features. Finally, the adaptive loss function of multimodal features is designed to realize the high-quality fusion of infrared and visible images. Extensive comparative experiments on three datasets were conducted, and HATF achieved competitive results on several performance metrics. The experimental results demonstrate that the proposed method is effective at fusing various scene images and that it outperforms related popular methods, thus verifying the superiority of HATF.

Our experiments on satellite images validate that HATF can be applied to remote sensing image fusion scenarios. In our future work, we will further verify the performance

of HATF in several potential application scenarios, such as medical image fusion, multi-exposure image fusion, and multi-focus image fusion. In addition, we will also focus on the use of fused image techniques to enhance the performance of other visual tasks, such as target detection and image segmentation.

Author Contributions: Conceptualization, X.L. (Xiangzeng Liu); Methodology, X.L. (Xiangzeng Liu); Software, Z.W. and H.G.; Formal analysis, X.L. (Xiang Li) and L.W.; Investigation, Z.W.; Resources, X.L. (Xiang Li); Data curation, H.G., X.L. (Xiang Li) and L.W.; Writing—original draft, Z.W. and H.G.; Visualization, L.W.; Supervision, Q.M.; Project administration, X.L. (Xiangzeng Liu) and Q.M.; Funding acquisition, X.L. (Xiangzeng Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Basic Research Program of Shaanxi (2024JC-YBMS-467) and the Aeronautical Science Foundation of China (D023030002).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We acknowledge the authors of the TNO, FLIR and LLVIP datasets used in this study.

Conflicts of Interest: Authors Xiang Li and Lei Wang were employed by the company NavInfo Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
2. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
3. Arroyo, S.I.; Bussi, U.; Safar, F.; Oliva, D. A monocular wide-field vision system for geolocation with uncertainties in urban scenes. *Eng. Res. Express* **2020**, *2*, 025041. [CrossRef]
4. Rajah, P.; Odindi, J.; Mutanga, O. Feature level image fusion of optical imagery and Synthetic Aperture Radar (SAR) for invasive alien plant species detection and mapping. *Remote Sens. Appl. Soc. Environ.* **2018**, *10*, 198–208. [CrossRef]
5. Ma, J.; Yu, W.; Chen, C.; Liang, P.; Guo, X.; Jiang, J. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Inf. Fusion* **2020**, *62*, 110–120. [CrossRef]
6. Liu, W.; Yang, J.; Zhao, J.; Guo, F. A Dual-Domain Super-Resolution Image Fusion Method with SIRV and GALCA Model for PolSAR and Panchromatic Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
7. Ying, J.; Shen, H.L.; Cao, S.Y. Unaligned hyperspectral image fusion via registration and interpolation modeling. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
8. Kumar, K.S.; Kavitha, G.; Subramanian, R.; Ramesh, G. Visual and thermal image fusion for UAV based target tracking. In *MATLAB-A Ubiquitous Tool for the Practical Engineer*; IntechOpen: London, UK , 2011; p. 307.
9. Meng, F.; Song, M.; Guo, B.; Shi, R.; Shan, D. Image fusion based on object region detection and non-subsampled contourlet transform. *Comput. Electr. Eng.* **2017**, *62*, 375–383. [CrossRef]
10. Li, H.; Qiu, H.; Yu, Z.; Zhang, Y. Infrared and visible image fusion scheme based on NSCT and low-level visual features. *Infrared Phys. Technol.* **2016**, *76*, 174–184. [CrossRef]
11. Toet, A.; Hogervorst, M.A. Multiscale image fusion through guided filtering. In Proceedings of the Target and Background Signatures II. SPIE, Edinburgh, UK, 26–29 September 2016; Volume 9997, pp. 170–182.
12. Aishwarya, N.; Bennila Thangammal, C. An image fusion framework using novel dictionary based sparse representation. *Multimed. Tools Appl.* **2017**, *76*, 21869–21888. [CrossRef]
13. Zhu, Z.; Yin, H.; Chai, Y.; Li, Y.; Qi, G. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inf. Sci.* **2018**, *432*, 516–529. [CrossRef]
14. Li, H.; Wu, X.J. Infrared and visible image fusion using latent low-rank representation. *arXiv* **2022**, arXiv:1804.08992.
15. Li, H.; Wu, X.J.; Kittler, J. MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef] [PubMed]
16. Gao, C.; Song, C.; Zhang, Y.; Qi, D.; Yu, Y. Improving the performance of infrared and visible image fusion based on latent low-rank representation nested with rolling guided image filtering. *IEEE Access* **2021**, *9*, 91462–91475. [CrossRef]
17. Jian, L.; Rayhana, R.; Ma, L.; Wu, S.; Liu, Z.; Jiang, H. Infrared and visible image fusion based on deep decomposition network and saliency analysis. *IEEE Trans. Multimed.* **2021**, *24*, 3314–3326. [CrossRef]

18. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17. [CrossRef]
19. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. Fusiondn: A unified densely connected network for image fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12484–12491.
20. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
21. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83*, 79–92. [CrossRef]
22. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolut. Inf. Process.* **2018**, *16*, 1850018. [CrossRef]
23. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
24. Li, H.; Wu, X.J.; Durrani, T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [CrossRef]
25. 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]
26. Chen, J.; Ding, J.; Yu, Y.; Gong, W. THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor. *Neurocomputing* **2023**, *527*, 71–82. [CrossRef]
27. Tang, W.; He, F.; Liu, Y.; Duan, Y.; Si, T. DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3159–3172. [CrossRef]
28. Park, S.; Vien, A.G.; Lee, C. Cross-Modal Transformers for Infrared and Visible Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 770–785. [CrossRef]
29. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference—Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany, 5–9 October 2015; Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
30. Yu, T.; Wang, X.; Chen, T.J.; Ding, C.W. Fault Recognition Method Based on Attention Mechanism and the 3D-UNet. *Comput. Intell. Neurosci.* **2022**, *2022*, 9856669. [CrossRef] [PubMed]
31. Soni, A.; Koner, R.; Villuri, V.G.K. M-unet: Modified u-net segmentation framework with satellite imagery. In Proceedings of the Global AI Congress 2019, Kolkata, India, 12–14 September 2019; Springer: Berlin/Heidelberg, Germany, 2020; pp. 47–59.
32. Alsabhan, W.; Alotaiby, T. Automatic building extraction on satellite images using Unet and ResNet50. *Comput. Intell. Neurosci.* **2022**, *2022*, 5008854. [CrossRef]
33. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like Pure Transformer for Medical Image Segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 205–218.
34. Lou, A.; Guan, S.; Loew, M. DC-UNet: Rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation. In Proceedings of the Medical Imaging 2021: Image Processing SPIE, Online, 15–19 February 2021; Volume 11596, pp. 758–768.
35. Tran, S.T.; Cheng, C.H.; Nguyen, T.T.; Le, M.H.; Liu, D.G. Tmd-unet: Triple-unet with multi-scale input features and dense skip connection for medical image segmentation. *Healthcare* **2021**, *9*, 54. [CrossRef] [PubMed]
36. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef]
37. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
38. 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* **2020**, arXiv:2010.11929.
39. Liu, N.; Zhang, N.; Wan, K.; Shao, L.; Han, J. Visual saliency transformer. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4722–4732.
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
41. Zhai, Y.; Shah, M. Visual attention detection in video sequences using spatiotemporal cues. In Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 815–824.
42. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
43. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the 13th European Conference—Computer Vision–ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
44. Toet, A. The TNO multiband image data collection. *Data Brief* **2017**, *15*, 249–251. [CrossRef] [PubMed]

45. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3496–3504.
46. Zhan, L.; Zhuang, Y.; Huang, L. Infrared and visible images fusion method based on discrete wavelet transform. *J. Comput.* **2017**, *28*, 57–71. [CrossRef]
47. Sruthy, S.; Parameswaran, L.; Sasi, A.P. Image fusion technique using DT-CWT. In Proceedings of the 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), Kottayam, India, 22–23 March 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 160–164.
48. Nencini, F.; Garzelli, A.; Baronti, S.; Alparone, L. Remote sensing image fusion using the curvelet transform. *Inf. Fusion* **2007**, *8*, 143–156. [CrossRef]
49. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [CrossRef]
50. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
51. Liu, X.; Gao, H.; Miao, Q.; Xi, Y.; Ai, Y.; Gao, D. MFST: Multi-Modal Feature Self-Adaptive Transformer for Infrared and Visible Image Fusion. *Remote Sens.* **2022**, *14*, 3233. [CrossRef]