remote sensing

MDPI

Article

MFST: Multi-Modal Feature Self-Adaptive Transformer for Infrared and Visible Image Fusion

Xiangzeng Liu [1], Haojie Gao [2], Qiguang Miao [1,*], Yue Xi [2], Yunfeng Ai [3] and Dingguo Gao [4]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; xzliu@xidian.edu.cn
[2] Guangzhou Institute of Technology, Xidian University, Xi'an 510555, China; gaohj@stu.xidian.edu.cn (H.G.); xiyue@xidian.edu.cn (Y.X.)
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China; aiyunfeng@ucas.ac.cn
[4] School of Information of Science and Technology, Tibet University, Lhasa 850000, China; gdg@utibet.edu.cn
* Correspondence: qgmiao@xidian.edu.cn

Abstract: Infrared and visible image fusion is to combine the information of thermal radiation and detailed texture from the two images into one informative fused image. Recently, deep learning methods have been widely applied in this task; however, those methods usually fuse multiple extracted features with the same fusion strategy, which ignores the differences in the representation of these features, resulting in the loss of information in the fusion process. To address this issue, we propose a novel method named multi-modal feature self-adaptive transformer (MFST) to preserve more significant information about the source images. Firstly, multi-modal features are extracted from the input images by a convolutional neural network (CNN). Then, these features are fused by the focal transformer blocks that can be trained through an adaptive fusion strategy according to the characteristics of different features. Finally, the fused features and saliency information of the infrared image are considered to obtain the fused image. The proposed fusion framework is evaluated on TNO, LLVIP, and FLIR datasets with various scenes. Experimental results demonstrate that our method outperforms several state-of-the-art methods in terms of subjective and objective evaluation.

Keywords: infrared image; visible image; transformer; image fusion; multi-modal feature; focal self-attention

Citation: 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. https://doi.org/10.3390/rs14133233

Academic Editors: Riccardo Roncella and Mattia Previtali

Received: 27 May 2022
Accepted: 1 July 2022
Published: 5 July 2022

1. Introduction

Image fusion refers to combining the images obtained by different types of sensors to generate a robust or informative image for subsequent processing and decision-making [1,2]. The technique is important for the fields of target detection [3], image enhancement [4], video surveillance [5], remote sensing [6–9], defogging [10], and so on. Due to differences in the imaging mechanism of the sensors, the scene information captured by infrared and visible images is very different in contrast and texture. Visible images are mainly reflection imaging, which is strongly dependent on lighting conditions. They usually have the characteristics of high spatial resolution, rich color, and texture details, which can offer a good source of perception in favorable lighting conditions. However, they are vulnerable to insufficient light or bad weather conditions. The infrared images reflect the thermal radiation of an object and are almost unaffected by weather and light. However, they usually have low spatial resolution and lack detailed texture information. Therefore, the fusion of two images provides more comprehensive information than a single image, which is very useful for subsequent high-level applications [11,12].

Currently, infrared and visible image fusion techniques can be divided into two categories: traditional methods and deep learning-based methods. In the past decades, traditional methods have been proposed for the fusion of pixel-level or fixed features. Traditional image fusion methods mainly include multi-scale transform (MST) [13,14], sparse

representation (SR) [15,16], salience [17,18] and low rank representation (LRR) [19,20]. The MST methods design appropriate fusion strategies to fuse the sub-layers obtained by using some transform operators, and the result is achieved through the inverse transformation. As a representative of MST method, Vanmali et al. [21] employed the laplacian pyramid as the transform operator and generated the weight map that was used to fuse the corresponding layers by considering local entropy, contrast, and brightness; therefore, good results can be achieved under the conditions of bad light. Yan et al. [22] constructed an edge-preserving filter for image decomposition, which can not only preserve the edge but also attenuate the influence of infrared background, ensuring that the fused image contains rich background information and salient features. However, MST method has a strong dependence on the choice of transformation, and its inappropriate fusion rules can introduce artifacts to the results [23]. Compared with MST, the goal of SR is to learn an over-complete dictionary to sparsely represent the source image, and the fused image can be reconstructed from the fused sparse representation coefficients. Bin et al. [24] adopted a fixed over-complete discrete cosine transform dictionary to represent infrared and visible images. Veshki et al. [25] used a sparse representation with identical support and Pearson correlation constraints without causing strength decay or loss of important information. For target-oriented fusion methods, salience methods can maintain the integrity of the significant target area and improve the visual quality of the fused images. Ma et al. [26] employed the rolling guidance and Gaussian filter as a multi-scale decomposition operator and used a visual saliency map to make the fusion result contain more visual details. Liu et al. [27] proposed a method combining salient object extraction and low-light region enhancement to improve the overall brightness of the image and make the results more suitable for human perception. As an efficient representation method, LRR is to decompose the images with low-rank representation and then fuse the sub-layers with appropriate rules. Gao et al. [22] proposed the combination of latent low-rank representation (LatLRR) and rolling guidance image filter (RGIF) to extract sub-layers from the images, which improved the fusion quality in terms of image contrast, sharpness, and richness of detail information. Although traditional methods have achieved indicated good performance, they still have three drawbacks: (1) the quality of handcrafted features determines the effect of fusion; (2) some traditional methods such as SR are very time-consuming; (3) specific fusion strategies need to be designed for various image datasets.

Recently, due to the advantages of strong adaptability, fault tolerance, and anti-noise capabilities, deep learning (DL) has been widely used in image fusion and has achieved better performance than traditional ones. According to the difference in network structure and output, DL-based fusion methods can be divided into two categories: non-end-to-end learning and end-to-end learning. For the former, the neural networks only extract deep features or output weights as the consideration of fusion strategy. Liu et al. [28,29] obtained the activity level measurement of the images through the Siamese convolutional network and combined it with the Laplace pyramid to realize the efficient fusion of infrared and visible images. Jian et al. [30] proposed a fusion framework based on decomposition network and salience analysis (DDNSA). They combined saliency map and bidirectional edge intensity to fuse the structural and texture features, respectively, and the fusion result can retain more details from the source images. While the end-to-end methods directly produce the fusion results through the network without very sophisticated and time-consuming operations. Xu et al. [31] proposed the FusionDN by employing a densely connected network to extract features effectively, which can be applied to multiple fusion tasks with the same weights. Ma et al. [32] proposed a new end-to-end model, termed DDcGAN, which established an adversarial game between a generator and two discriminators for fusing infrared and visible images at different resolutions. In the past two years, many methods have begun to adopt the framework of feature extraction-fusion-image reconstruction. This framework can maximize the capabilities of feature extraction and feature fusion, respectively, and ultimately improve the quality of fusion. Yang et al. [33] proposed a method based on dual-channel information cross fusion block (DICFB) for cross extraction and preliminary

fusion of multi-scale features, and the final image is enhanced by saliency information. By considering the illumination factor in the feature extraction stage, Tang et al. [34] proposed a progressive image fusion network termed as PIAFusion, which can adaptively maintain the intensity distribution of significant targets and retain the texture information in the background.

Although the above-mentioned methods have achieved competitive performance, they still have as following disadvantages:

1. The design of the multi-feature fusion strategy is simple and does not make full use of feature information.
2. CNN-based methods only consider local features in the fusion process without modeling long-range dependencies, which will lose global context meaningful for the fusion results.
3. End-to-end methods lack obvious feature extraction steps, resulting in poor fusion results.

In order to alleviate the drawbacks mentioned above, this paper presents a novel fusion framework based on the focal Transformer fusion model and multi-modal feature self-adaptive fusion strategy. The main contributions of our paper can be summarized as follows:

1. To fully utilize both local information and global context, a new fusion model that introduces the optimized focal self-attention is constructed.
2. To effectively utilize the multi-modal feature, an adaptive fusion strategy is designed according to the representation of different layer features, which makes the fusion results retain more structural features from the source images.
3. Experiments show that our method outperforms the existing state-of-the-art fusion methods in both subjective appraisement and objective evaluation on multiple datasets.

The rest of this paper is organized as follows: The related works of fusion for visible and infrared images are reviewed, and the development and superiority of the transformer are described in Section 2. A detailed description of the multi-modal feature self-adaptive transformer is given in Section 3. Comparative experiments and analysis are performed in Section 4. Finally, conclusions are drawn in Section 5.

2. Related Works

In this section, we first review a special network structure: auto-encoder. It employs a two-stage training strategy, which can improve the performance of feature extraction and fusion. Then, we introduce the development of a transformer in the field of computer vision and its great potential in image fusion.

2.1. Auto-Encoder-Based Methods

In CNN-based fusion methods, the last layer is often used as output features or to produce fusion results, which will lose the meaningful information contained by the middle layers. In order to solve this problem, Li et al. [35] proposed DenseFuse for infrared and visible image fusion, which is composed of an encoder network, fusion strategy, and decoder network. In which the encoder network comprised of convolution layers and dense blocks are used to extract deep features, and the decoder network is applied to reconstruct the image. In their fusion phase, the addition strategy or $l_1 - norm$ strategy is adopted to fuse the deep features, which can preserve more details from the source images.

To improve DenseFuse, Li et al. [36] proposed Nestfuse, in which the encoder network is changed to a multi-scale network, and the nest connection architecture is selected as the decoding network. Due to their design of spatial/channel attention fusion strategies, the model can better fuse the background details and salient regions in the image. However, this handcrafted strategy cannot effectively utilize multi-modal features. Therefore, Li et al. [37] further proposed RFN-nest, adopting a residual fusion network to learn the fusion weight. Although these methods achieve good results to some extent, they adopt the same fusion strategy for multi-modal features, which ignores the differences between these features

at various modals. In order to improve the fusion quality, the focal transformer model is adopted, and a self-adaptive fusion strategy is designed for multi-modal features.

2.2. Transformer-Based Method

Transformer [38] was first applied to natural language processing and has achieved great success. Unlike CNN's focus on local features, the transformer's attention mechanism can help it establish long-range dependence so as to make better use of global information in both shallow and deep layers. The proposal of a vision transformer [39] shows that the transformer has great potential in computer vision (CV). In recent years, more and more researchers have introduced transformers into CV, such as object detection, segmentation, multiple object tracking, and so on. Liu et al. [40] proposed VST, which adopts T2T-ViT as the backbone, introducing a new multitask decoder and reverse T2T token upsampling method. Unlike some methods in which class tokens are directly used in image classification via using multilayer perceptron on the token embedding, VST recommends that patch-task-attention should be carried out between patch tokens and task tokens to predict saliency and boundary map.

Although the transformer has better representation ability, it needs enormous computational overhead when processing high-resolution images. To alleviate the challenge of adapting the transformer from language to vision, many researchers began to explore the transformer structure more suitable for CV. Liu et al. [41] proposed a Swin transformer, in which the key is the shift window scheme, which limits the self-attention calculation to non-overlapping local windows and allows cross window connection so as to improve the efficiency. Inspired by the Swin transformer, Li et al. [42] proposed a multi-path structure of transformer called LG-Transformer, which can carry out local-to-global reasoning on the multiple granularities of each stage and solve the problem of lack of global reasoning in the early stages of the previous models. These methods of applying coarse-grained global attention and fine-grained local attention improve the performance of the model but also weaken the modeling ability of the transformer's original self-attention mechanism. Therefore, Yang et al. [43] proposed a focal transformer, which combines fine-grained local interaction with coarse-grained global interaction. In the work of the focal transformer, a new mechanism called focal self-attention is introduced, in which each token attends to its nearest surrounding tokens in fine granularity and far away tokens in coarse granularity. This method can capture both short-term and long-term visual dependencies, and the computational efficiency is greatly improved.

In view of the advantages of the focal transformer, we introduce focal self-attention into the fusion task and propose a novel self-adaptive fusion strategy according to the characteristics of multi-modal features.

3. Methodology

In this section, the multi-modal feature self-adaptive transformer is presented. Section 3.1 introduces the architecture of the fusion model. Then, Section 3.2 presents the fusion strategy in detail. Finally, the loss function and training phase are given in Section 3.3.

3.1. The Architecture of the Fusion Model

The proposed network is an end-to-end network consisting of three main parts: encoder, transformer fusion strategy, and decoder. The architecture of our fusion model is illustrated in Figure 1, which can be described as follows:

(1) Encoder network: the multi-scale encoder network accepts one input image (infrared image Ir and visible image Vi) and generates multi-modal deep features (Φ_{ir}^{i} and Φ_{vi}^{i}), which contains one convolution layer and four encoder blocks. Each encoder block contains two convolution layers followed by the ReLU operation and one max-pooling layer.

(2) Transformer fusion block: the multi-modal features (Φ_{ir}^{i} and Φ_{vi}^{i}) extracted from different source images are fed to the transformer fusion blocks (TFB) to obtain the fused

features (Φ_f^i), which carry out fine-grained local fusion and coarse-grained global fusion at the same time, helping the model fuse both local features and global context.

(3) Decoder network: the fused features (Φ_f^i) are input into the decoder to generate the fusion result (fused image F). The decoder consists of 6 decoder blocks and one convolution layer. As shown in Figure 1, these blocks are combined through nest connection, which greatly improves the image reconstruction ability. The encoder and decoder in this paper are constructed according to the structure in NestFuse.

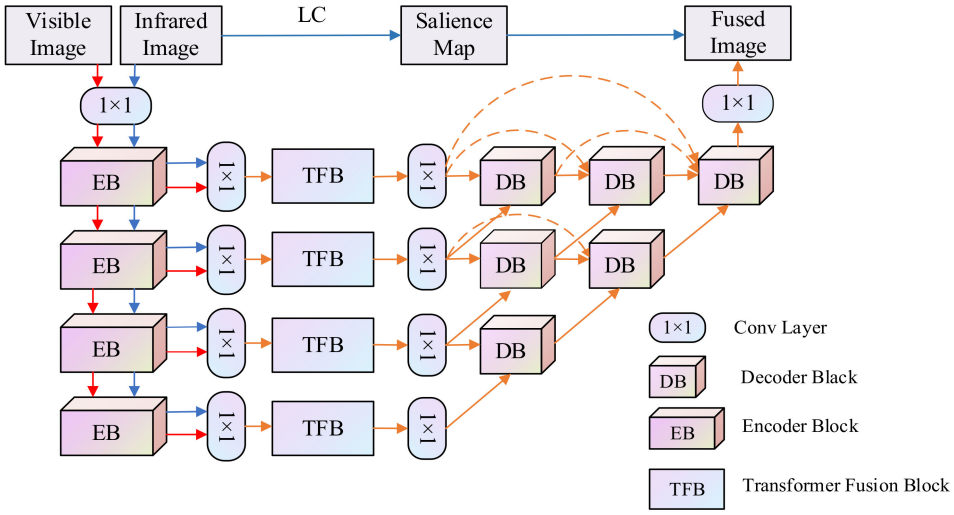

Figure 1. Overview of the proposed Multi-modal Feature Self-adaptive Transformer fusion model.

3.2. Transformer Fusion Strategy

Most of fusion methods generally use a simple strategy to fuse all extracted features, which causes the loss of feature information during the fusion process. Therefore, the fusion strategy for features should be carefully designed. In this paper, our fusion strategy consists of two parts, one is a self-adaptive fusion strategy designed according to the characteristics of multi-modal features, and the other is transformer fusion blocks used to fuse local features and global context. These two parts will be elaborated in the following two subsections.

3.2.1. Multi-Modal Features Self-Adaptive Fusion Strategy

The extracted multi-modal deep features contain a variety of information. Multi-modal features extracted by encoder are shown in Figure 2. It can be seen that the shallow layer features have abundant details, the middle layer features represent the structural information of the image and the deep features are mainly region features. Therefore, when fusing multi-modal features, we design specific loss functions to ensure that the feature information can be transferred to the fusion result to the greatest extent.

As mentioned above, the first layer feature contains more details, and detail feature (pixel) loss L_{df} is designed at the pixel level to retain more details and textures for fusion. The detail feature loss L_{df} is calculated as follows:

$$L_{df} = \left\| \Phi_f^1 - \left(w_{ir} \Phi_{ir}^1 + w_{vi} \Phi_{vi}^1 \right) \right\|_F^2, \quad (1)$$

where Φ_{ir}^1, Φ_{vi}^1 denotes the first layer features of infrared image and visible image, respectively, Φ_f^1 denotes the fused feature of the first layer features. w_{ir} and w_{vi} are self-adaption weights, which are defined as follows:

$$w_{ir} = \frac{\left\|\Phi_{ir}^1\right\|_1}{\left\|\Phi_{ir}^1\right\|_1 + \left\|\Phi_{vi}^1\right\|_1}, w_{vi} = 1 - w_{ir}. \tag{2}$$

while, the features of the middle layer contain more structural features such as contour and edge. The purpose of the structural feature (correlation) loss L_{sf} is to retain the integrity of structure and edges from extracted features. L_{sf} is defined as follows:

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

where $\mathrm{cov}(\cdot)$ denotes covariance function, and σ denotes the standard deviation function.

Furthermore, the deepest features have the lowest resolution, and their foreground and background are obviously distinguished; hence, the region feature loss function L_{sf} is defined as follows:

$$L_{rf} = \left\|\Phi_f^4 - \left(w_{ir}M_{ir}^4\Phi_{ir}^4 + w_{vi}M_{vi}^4\Phi_{vi}^4\right)\right\|_F^2, \tag{4}$$

where M_{ir}^4 and M_{vi}^4 denotes the masks to remove the noise from the features, and the calculation is as follows:

$$M_{ir}^4 = \begin{cases} 1, \Phi_{ir}^4 \geq \theta \\ 0, \Phi_{ir}^4 < \theta \end{cases}, \tag{5}$$

$$M_{vi}^4 = \begin{cases} 1, \Phi_{vi}^4 \geq \theta \\ 0, \Phi_{vi}^4 < \theta \end{cases}, \tag{6}$$

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

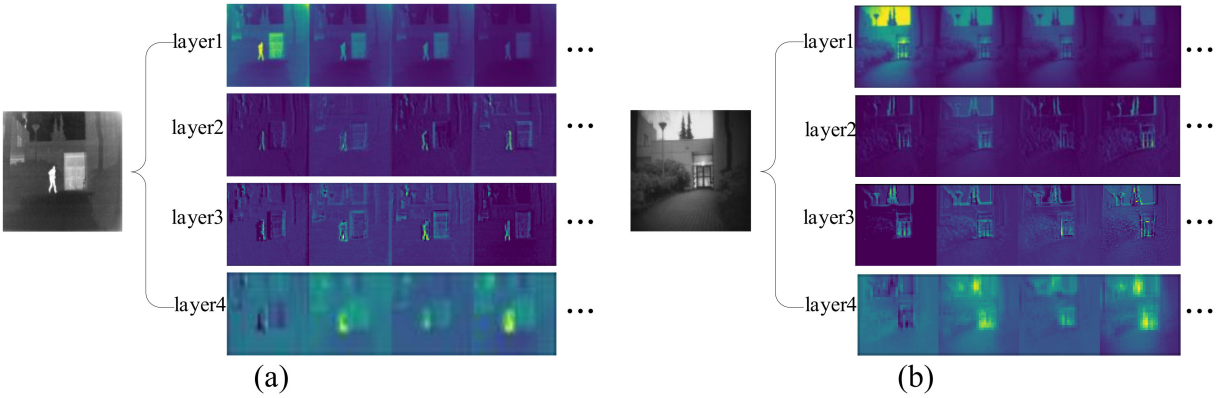

Figure 2. Multi-modal features extracted by encoder. (**a**) shows the features extracted from infrared image, (**b**) shows the features extracted from visible image.

3.2.2. Transformer Fusion Block

Focal self-attention incorporates both fine-grained local interaction and coarse-grained global interaction, which reduces quadratic computational overhead of processing high-resolution images and make it more suitable for fusion tasks than traditional transformer. Therefore, in this paper, focal transformer is employed as the fusion module. The feature fusion stage contains four Transformation fusion blocks (TFB), and each TFB consists of 2 focal transformer layers. The structure of focal transformer layer is shown in Figure 3. The focal transformer layer has two sub-layers. The first is the focal self-attention mechanism

to capture local and global features, and the second is a multi-layer perception network to improve the modeling ability of complex processes. Each sub-layer has layer normalization to stabilize the data distribution and facilitate training and residual connection to solve the problem of vanishing gradient.

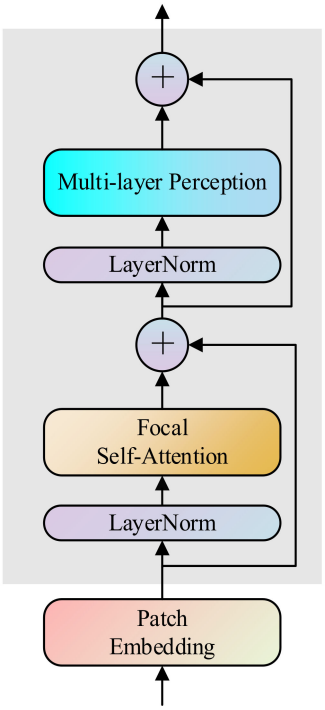

Figure 3. The architecture of the Focal Transformer fusion layer.

The key mechanism of focal transformer is focal attention. Unlike the standard self-attention, focal self-attention pays attention to fine-grained tokens locally and coarse-grained tokens globally. Therefore, it can cover the areas covered by the standard self-attention, while the cost is much lower. The detailed mechanism of focal self-attention is shown in Figure 4. The focal self-attention is performed at the window level. Computational steps of focal transformer can be described as follows:

(1) The input x of all layers L is split into sub-window with the size of $s_w^l \times s_w^l$. After sub-window pooling operation, we can obtain the feature map $\left\{x^l\right\}_1^L$ which provide rich information of both coarse-grained and fine-grained. Where focal level L is the number of granularity, focal window size s_w^l is the size of sub-window at level $l \in \{1, \ldots, L\}$, focal region size s_r^l is the number of sub-windows horizontally and vertically at level l.

(2) With the three linear projection layers, the query of the first layer and the key and value of all layers are calculated as follows:

$$f_q\left(x^i\right) = W^q x^i, \tag{7}$$

$$f_k\left(x^i\right) = W^k x^i, \tag{8}$$

$$f_v\left(x^i\right) = W^v x^i, \tag{9}$$

$$Q = f_q\left(x^1\right), K = \left\{K^l\right\}_1^L = f_k\left(\left\{x^1, \ldots, x^L\right\}\right), V = \left\{V^l\right\}_1^L = f_v\left(\left\{x^1, \ldots, x^L\right\}\right), \tag{10}$$

where W^q, W^k and W^v are three matrices obtained by learning. For the queries inside the i-th window Q_i, we extract the $s_r^l \times s_r^l$ keys and values from K^l and V^l around the window where the query lies in, and then gather the keys and values from all L to obtain $K_i = \{K_i^1, \ldots, K_i^L\} \in R^{s \times d}$ and $V_i = \{V_i^1, \ldots, V_i^L\} \in R^{s \times d}$, where s is the sum of focal region from all levels.

(3) Finally, with the learnable relative position bias $B = \left\{B^l\right\}_1^L$, the focal self-attention for Q_i can be calculated as follows:

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}} + B\right) V_i, \tag{11}$$

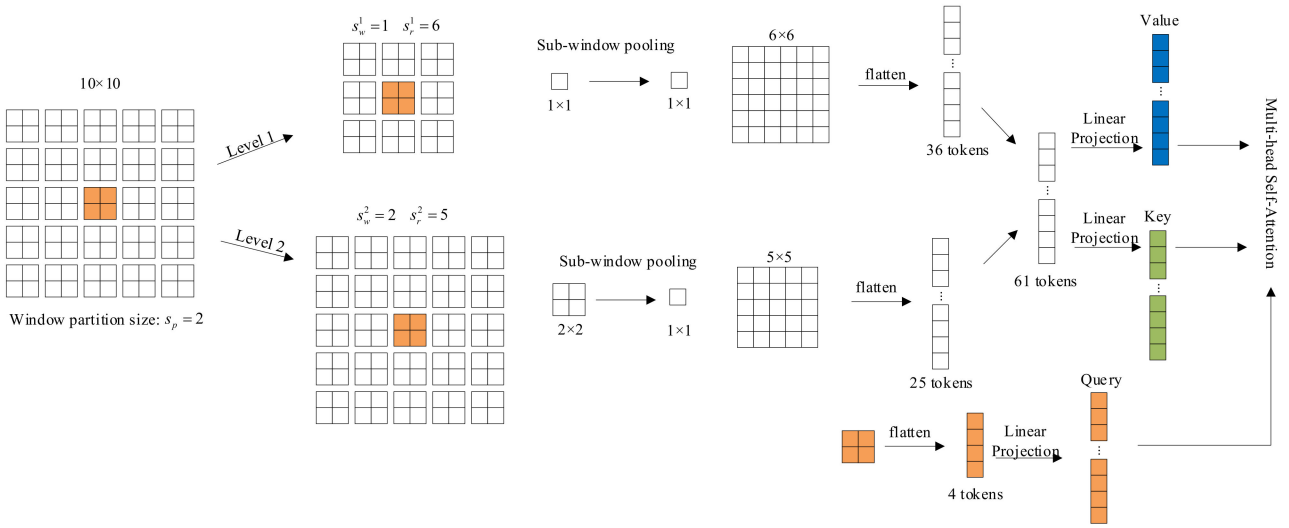

Figure 4. An illustration of focal self-attention mechanism. Each square in the figure is a visual token from the input feature map. Assuming the input feature map is of size 10×10, partition it into 5×5 windows of size 2×2. The 2×2 orange square in the middle is selected as a query, and its surroundings tokens at multiple granularities are extracted as its keys and values. For the first level, a 6×6 grid around the query is extracted as fine-grained tokens. For the second level, focus on the region of the entire feature map to obtain coarse-grained tokens, at which point 2×2 tokens are pooled into 1×1 tokens. Then the tokens of two levels are concatenated, and the keys and values corresponding to the query can be obtained through linear projection.

3.3. Loss Function

Inspired by RFN-Nest, our fusion network adopts two-stage training strategy as well. Loss L_{enco} and loss L_{fusion} are utilized to guide the optimization of the encoder and transformer fusion blocks, respectively.

3.3.1. Loss Function of Encoder

In the first stage, the encoder and decoder are trained together to improve the ability of feature extraction. Through the maximum pooling operation, the encoder receives one input and generates four scale features. Then, the extracted features are directly sent to the decoder to reconstruct the image. The decoder uses short cross-layer connections, which can greatly improve the reconstruction ability of features.

The loss function of encoder–decoder network L_{enco} consists of two parts: the content loss L_{con} and the structural similarity loss L_{ssim}, which can be formulated as follows:

$$L_{enco} = L_{con} + \lambda L_{ssim}, \tag{12}$$

where λ is a hyperparameter to balance these two terms. Content loss L_{con} enables the output to preserve more content details of the source image, which is calculated as follows:

$$L_{con} = \|O - I\|_F^2, \tag{13}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm, O is the output image and I is the input image.

SSIM loss helps the fused image contain more structural features from the source images. The calculation of SSIM loss L_{ssim} is defined as follows:

$$L_{ssim} = 1 - ssim(I, O), \tag{14}$$

where $ssim(\cdot)$ denotes the structural similarity [44], which is an index to measure the similarity of two images from three different aspects: brightness, contrast, and structure. SSIM loss makes the output structurally more similar to the input image. The calculation of SSIM can be formulated as follows:

$$SSIM(X, Y) = \sum_{x,f} \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \cdot \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{15}$$

where $SSIM_{x,y}$ denotes the structural similarity between source images X and Y; x and y denote the image patches of source images in a sliding window, respectively; μ_x and μ_y denote the mean values of source images, respectively; σ_{xy} denotes the covariance of source and fused images; σ_x and σ_y denote the standard deviation (SD). C_1, C_2, and C_3 are the parameters used to make the algorithm stable.

3.3.2. Loss Function of Fusion

In the second training phase, we connect the transformer fusion blocks between the encoder and decoder and freeze the weight of the encoder and decoder. The loss function L_{fusion} in fusion phase consists of three parts: the structural similarity loss L_{ssim}, the multi-modal feature loss L_{fea} and the salience loss L_{sal}. The fusion loss L_{fusion} is calculated as follows:

$$L_{fusion} = \alpha L_{ssim} + \beta L_{fea} + L_{sal}, \tag{16}$$

where α, β are hyperparameters to balance these three terms.

Visible image has significant and distinct structure, while the infrared image has lots of noise and fuzzy structure. Hence L_{ssim} here only restricts the similarity in structure and details between the visible image and the fused image to make the fused image more similar to the visible image. The calculation of SSIM loss L_{ssim} is the same as Formula (5), and the inputs are the fused image F and the visible image I_{vi}, respectively.

For multi-modal features, The feature loss L_{fea} contained three items: detail feature loss L_{df}, structure feature loss L_{sf} and region feature loss L_{rf}. The feature loss L_{fea} is calculated as follows:

$$L_{fea} = L_{df} + \mu L_{sf} + \rho L_{rf}, \tag{17}$$

where μ and ρ are the hyperparameters, which are used to balance the weight of these three items. The calculation of L_{df}, L_{sf} and L_{rf} are given in Section 3.2.2.

In order to preserve the salient objects in the infrared image, salience information is introduced to the fusion process. Firstly, the LC salience extraction algorithm is employed to detect the infrared source image to obtain the salience map M_{sal}. Then the salience map M_{sal} is normalized to obtain the $\hat{M}_{sal}$. Finally, the salience loss is designed as shown in the following formula:

$$L_{sal} = \left\|\hat{M}_{sal} * F - \hat{M}_{sal} * I_{ir}\right\|_F^2 + \left\|\left(1 - \hat{M}_{sal}\right) * F - \left(1 - \hat{M}_{sal}\right) * I_{vi}\right\|_F^2, \tag{18}$$

Through the carefully designed loss function, the multi-modal feature information is transferred to the fusion results to the greatest extent. The fused results of the pro-

posed method contain abundant texture details, distinct edge contour, and good visual salience. In Section 4, we will verify the effectiveness of the proposed method through comparative experiments.

4. Experimental Results and Analysis

In this section, we conduct an experimental analysis of the proposed method on TNO, FLIR, and LLVIP datasets with various scenes. The training details and parameter settings are introduced in Section 4.1. In order to demonstrate the effectiveness of the proposed MFST, three datasets and six quality metrics are employed in comparative and evaluative experiments, which are introduced in Section 4.2. Then, the comparative experimental results of the proposed method and state-of-the-art methods are analyzed in Section 4.3.

4.1. Datasets and Training Details

The four datasets adopted in the training and testing phases are as follows:

(1) MS-COCO dataset [45]: The COCO dataset has over 330,000 images and is a large, rich dataset for object detection, segmentation, and captioning. The images in COCO are mainly intercepted from daily scenes, with complex backgrounds, a large number of targets, and a small target size.
(2) TNO dataset [46]: The TNO Image Fusion Dataset contains multispectral (intensified visual, near-infrared, and longwave infrared or thermal) nighttime imagery of different military relevant scenarios registered with different multiband camera systems.
(3) FLIR dataset: This dataset provides annotated thermography datasets and corresponding unannotated RGB images. The dataset contains a total of 14,452 infrared images, of which 10,228 are from multiple short videos; 4224 are from a video with a duration of 144 s. All video scenes are streets and highways.
(4) LLVIP dataset [47]: This dataset contains 15,488 pairs of images, most of which were taken at very dark scenes, and all of the images are strictly aligned in time and space.

The proposed model is implemented in Pytorch. All the experiments are conducted on an NVIDIA GeForce RTX 2080Ti GPU and 3.6-GHz Intel Core i7-7700 CPU. The configurations for encoder, decoder, and transformer fusion blocks are shown in Tables 1 and 2. The encoder and decoder are trained in the first training stage. At this time, the features extracted by the encoder are fed to the decoder directly. The dataset for training contains 80,000 images selected in MS-COCO. These pictures are converted into gray-scale and reshaped into 256×256. λ in Equation (12) is set to 100. Batch size and epoch are set to four and two, respectively. The learning rate is 1×10^{-4}.

In the second training phase, the weights of the encoder and decoder are frozen, and four TFBs are connected between the encoder and decoder. Some parameters of TFB are set as follows: The window partition size is set to seven, and the focal self-attention layer is set to two to obtain both fine-grain local attention and coarse-grain global attention. The focal window size is set to one, and the focal region size is set to thirteen in the first focal level. For the coarse-grain global attention, the focal window size is set the same as the window partition size seven, but the focal region size is decreased to obtain {37, 19, 9, 5} for the four blocks. The LLVIP dataset is employed to train our TFB. We selected 12,000 pairs of infrared and visible images in the training set. These images are converted to gray-scale and reshaped into 256×256 as well. α and β in Equation (16) are set to 700 and 1, μ and ρ in Equation (17) are set to 1000 and 1, respectively. Batch size and epoch are both set to 2. The learning rate is 1×10^{-4}. Figure 5 displays loss curves versus iteration index. It is shown that all loss curves are very flat after 250 iterations. The training and validation loss curves show that the model is able to converge with this configuration.

Table 1. Model configurations for encoder and decoder network. EB1 denotes the encoder block in the first row, DB11 denotes the decoder in the first row and the first column.

	Layer	Input Channel	Output Channel	Size	Stride	Activation
Encoder	EB1	1	64	3	1	ReLU
	EB2	64	112	3	1	ReLU
	EB3	112	160	3	1	ReLU
	EB4	160	208	3	1	ReLU
Decoder	DB11	176	64	3	1	ReLU
	DB12	240	64	3	1	ReLU
	DB13	304	64	3	1	ReLU
	DB21	272	112	3	1	ReLU
	DB22	384	112	3	1	ReLU
	DB33	368	160	3	1	ReLU

Table 2. Model configurations for Transformer Fusion Block network. TFB1 denotes the Transformer Fusion Block in the first row.

	Number of Layers	Channel	Window Partition Size	Focal Window Size	Focal Region Size
TFB1	2	64	7	1, 7	13, 37
TFB2	2	112	7	1, 7	13, 19
TFB3	2	160	7	1, 7	13, 9
TFB4	2	208	7	1, 7	13, 5

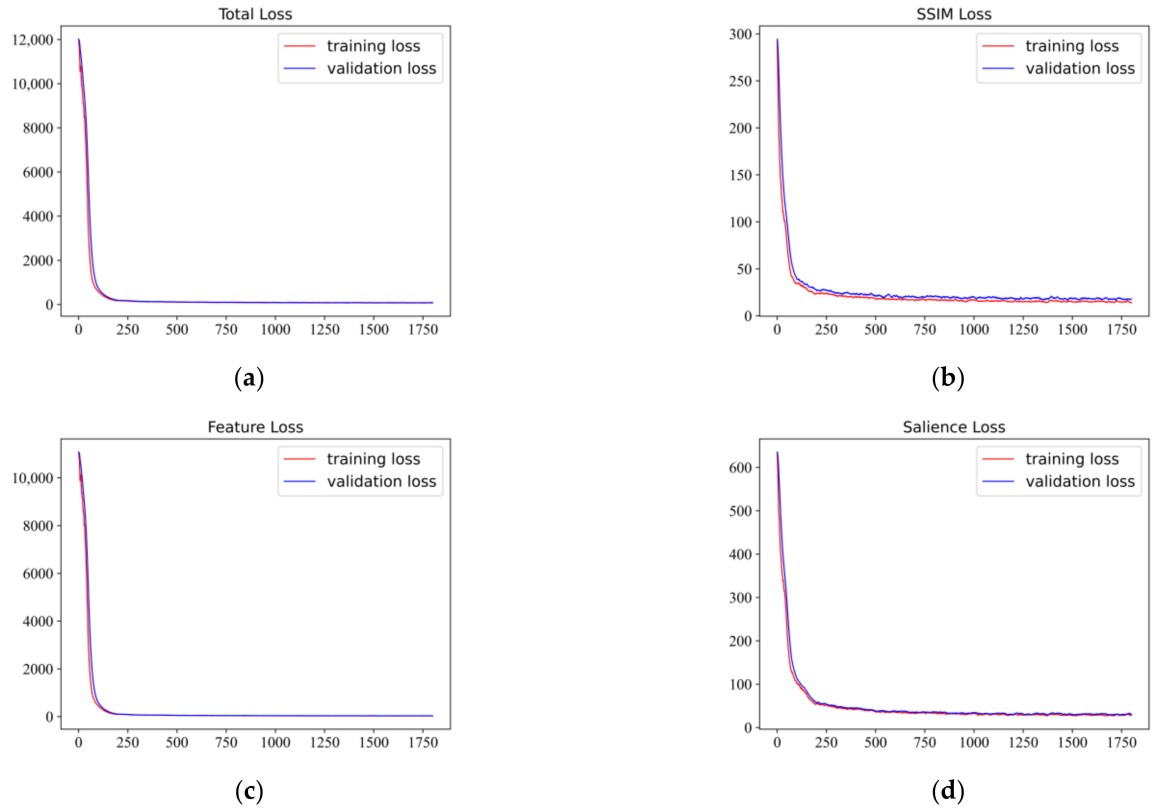

Figure 5. Fusion loss curves over 1800 iterations. (**a**) shows the training and the validation curves of total loss. (**b**–**d**) are training and the validation curves of SSIM loss, feature loss and salience loss respectively.

4.2. Evaluation Metrics

To quantitatively evaluate the fusion performance, we select six evaluation metrics, including entropy (EN), standard deviation (SD), the sum of the correlations of differences (SCD), structural similarity index measure (SSIM), mutual information (MI), and root mean square error (RMSE) [48]. These metrics measure the performance of the fusion method from different aspects, such as the amount of information, the information transmitted by the source images, visual quality, and so on. Their definitions are described as follows.

EN measures the amount of information contained in a fused image based on information theory and can be defined as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \ \log_2 \ p_l, \tag{19}$$

where L denotes the number of gray levels and p_l denotes the normalized histogram of the corresponding gray level in the fused image. The larger the value of EN, the more information contained in the fused image, and the better the performance of the fusion method.

The SD metric is based on the statistical concept that reflects the distribution and contrast of the fused image, which is defined as follows:

$$SD = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-\mu)^2}, \tag{20}$$

where F denotes the fused image, $F(i,j)$ is the pixel value at coordinate (i,j) in image F, and μ denotes the mean value of F. A larger SD represents a higher contrast of the region, which can attract more attention due to the sensitivity of the human visual system, which means the fusion method achieves a good visual quality.

The SCD does not directly use the correlation between the source image and the fused image to evaluate the quality of the fused image but considers the source images and their influence on the fused images to calculate the quality. It is defined as follows:

$$SCD = r(D_1, S_1) + r(D_2, S_2), \tag{21}$$

where difference images D_1 and D_2 can be obtained by $D_1 = F - S_1$ and $D_2 = F - S_2$. F, S_1 and S_2 denote the fused image, the first and second input image, respectively. The $r(\cdot)$ function calculates the correlation between S_1 and D_1, S_2 and D_2 as:

$$r(D_k, S_k) = \frac{\sum_i \sum_j \left(D_k(i,j) - \overline{D}_k\right)\left(S_k(i,j) - \overline{S}_k\right)}{\sqrt{\left(\sum_i \sum_j \left(D_k(i,j) - \overline{D}_k\right)\right)\left(\sum_i \sum_j \left(S_k(i,j) - \overline{S}_k\right)\right)}}, \tag{22}$$

where $k = 1, 2$, $\overline{S}_k$ and $\overline{D}_k$ are the average of the pixel values of S_k and D_k, respectively. A higher SCD index indicates that the fusion method achieves a good performance.

The human visual system is sensitive to structure loss and distortion. The calculation of the SSIM index is shown in Formula (11). Due to the visible images having more structural features than the infrared ones, we calculate the SSIM between the fused images and the visible source images. The larger the value of SSIM, the better the structure is maintained from visible source images.

The MI metric is a quality index that measures the amount of information transferred from source images to the fused image. MI is a fundamental concept in information theory

and measures the dependence of two random variables. The definition of the MI metric is given as follows:

$$MI = MI_{A,F} + MI_{B,F}, \tag{23}$$

where $MI_{A,F}$ and $MI_{B,F}$ denote the amount of information that is transferred from infrared and visible images to the fused image, respectively. The MI between two random variables can be calculated by the Kullback–Leibler measure, which is defined as follows:

$$MI_{X,F} = \sum_{x,f} P_{X,F}(x,f) \log \frac{P_{X,F}(x,f)}{P_X(x)P_F(f)}, \tag{24}$$

where $P_X(x)$ and $P_F(f)$ denote the marginal histograms of source image X and fused image F, respectively. $P_{X,F}(x,f)$ denotes the joint histogram of source image X and fused image F. A larger MI indicates that the more information transferred from the source image to the fusion result, which means the better performance.

The root mean squared error (RMSE) metric is similar to the MSE metric and is defined as follows:

$$RMSE_F = \frac{RMSE_{IF} + RMSE_{VF}}{2}, \tag{25}$$

$$RMSE_{XF} = \sqrt{\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(X(m,n) - F(m,n))^2}, \tag{26}$$

where X denotes infrared images I or visible images V, $RMSE_{IF}$ and $RMSE_{VF}$ denote the dissimilarity between the fused and infrared/visible images. A small RMSE metric indicates that the fused image has a small amount of error and distortion and hence the fusion method achieves a good performance.

4.3. Ablation Study

In this paper, focal self-attention is introduced to facilitate the model paying attention to both local features and global context, and a self-adaptive fusion strategy is designed to fuse and transmit multi-modal feature information more comprehensively. In order to verify the effectiveness of these two parts for improving the fusion effect, we set up four different configurations in the ablation experiment and used quantitative metrics to evaluate and analyze the results. The four configurations including with focal self-attention and self-adaptive fusion strategy (WS, WF), with focal self-attention and without self-adaptive fusion strategy (WS, OF), without focal self-attention and with self-adaptive fusion strategy (OS, WF), and without focal self-attention and self-adaptive fusion strategy (OS, OF). When focal attention is not used, the feature dimension is reduced by two convolutional layers with kernel size 1×1. When the self-adaptive fusion strategy is not used, only the SSIM loss and the saliency loss are included in the total loss function. Our ablation experiments are conducted on the TNO dataset.

The average quantitative results of the ablation experiments are shown in Table 3. The average performance of WS, OF, and WF, OS is better than OS, OF, which indicates that both focal self-attention and self-adaptive fusion strategy play a certain role in the fusion process. The performance of OS, WF is higher than that of WS, OF, which shows that the self-adaptive fusion strategy has a relatively greater effect on fusion results than focal self-attention. The models of WS, WF achieved the best average results in five metrics, demonstrating that these two parts are indispensable in the proposed method, and both have a great effect on the improvement of fusion performance. At the same time, we also noticed that on MI metric, OS, WF performs better than WS, WF, because, without the transformer fusion block, the loss of feature information is the least in the transmission process. However, the absence of a transformer fusion block means that the model can only focus on local areas when fusing features and lacks global context information, resulting in OSWF's performance not being as good as WSWF on other metrics.

Table 3. Objective evaluation of outputs of with focal self-attention and self-adaptive fusion strategy (WS, WF), with focal self-attention and without self-adaptive fusion strategy (WS, OF), without focal self-attention and with self-adaptive fusion strategy (OS, WF) and without focal self-attention and self-adaptive fusion strategy (OS, OF) on TNO dataset.

Method	EN	SD	SCD	SSIM	MI	RMSE
OS, OF	6.6784	35.7851	1.5017	0.7342	2.6247	10.2413
WS, OF	6.7561	38.1576	1.5578	0.7439	2.6782	10.1892
OS, WF	6.8242	39.0014	1.5876	0.7411	2.843	10.1678
WS, WF	6.9519	39.3726	1.6011	0.7466	2.7028	10.1066

4.4. Experimental Results Analysis

In order to verify the superiority of the proposed method, we tested it on three datasets: TNO, FLIR, and LLVIP. A total of 21 pairs of images are selected from TNO, and 50 pairs of images are selected from FLIR and LLVIP, respectively. Because the major scenes of FLIR and LLVIP are both roads, including pedestrians, vehicles, and other targets, we discuss the results of these two datasets together, and the results of TNO are discussed separately. Eight existing state-of-art fusion methods are selected for comparison, including four traditional methods (DWT [49], DTCWT [50], CVT [51] and NSCT [52]) and four deep learning methods (DenseFuse [35], FusionGan [53], IFCNN [54], and RFN-Nest [37]). The parameters of these methods are set in strict accordance with the relevant reference.

4.4.1. Fusion Results Analysis on TNO

Six pairs of typical image fusion results obtained by the proposed method and the other eight methods on the TNO dataset are shown in Figure 6, in which source images include people, cars, umbrellas, houses, ships, trenches, rivers, and other targets and scenes. It can be seen that most of the methods can achieve ordinary fusion results. However, there are still deficiencies in the fusion of salience targets and texture details. At the same time, these methods are easy to introduce artifacts that reduce the quality of fusion results. Results of DWT, NSCT, CVT, and DTCWT have an indistinct edge, dim significant target, and some artifacts. This is because they all use manual methods to decompose the source images and employ the "weighted-average" or "choose-max" strategy to fuse the decomposed components, which leads to the loss of texture details from the source images and the decline of the quality of fusion results. The FusionGan method is to preserve more infrared image content so that some details in the visible images are lost, and the edges are blurred. By contrast, for RFN-Nest, the texture information in the visible images can be completely retained, while the thermal target from the infrared source is dimmed. This poor performance can be seen from the first and sixth pair of image fusion results. DenseFuse and IFCNN can better complete the fusion task; however, they adopt a general strategy to fuse the extracted features, which leads to the loss of contour information and an increase in noise.

Compared with the above fusion methods, the fusion results of the proposed method have a richer texture and sharper edge, and the thermal targets from infrared images are more prominent and have higher visual quality, which is due to two points: (1) the specific loss functions could help the learning of self-adaptive fusion strategy, making edges features can be transferred to the results commendably. (2) focal self-attention facilitates the fusion of local and global multi-modal features. In turn, the visual quality of the fusion results is significantly improved.

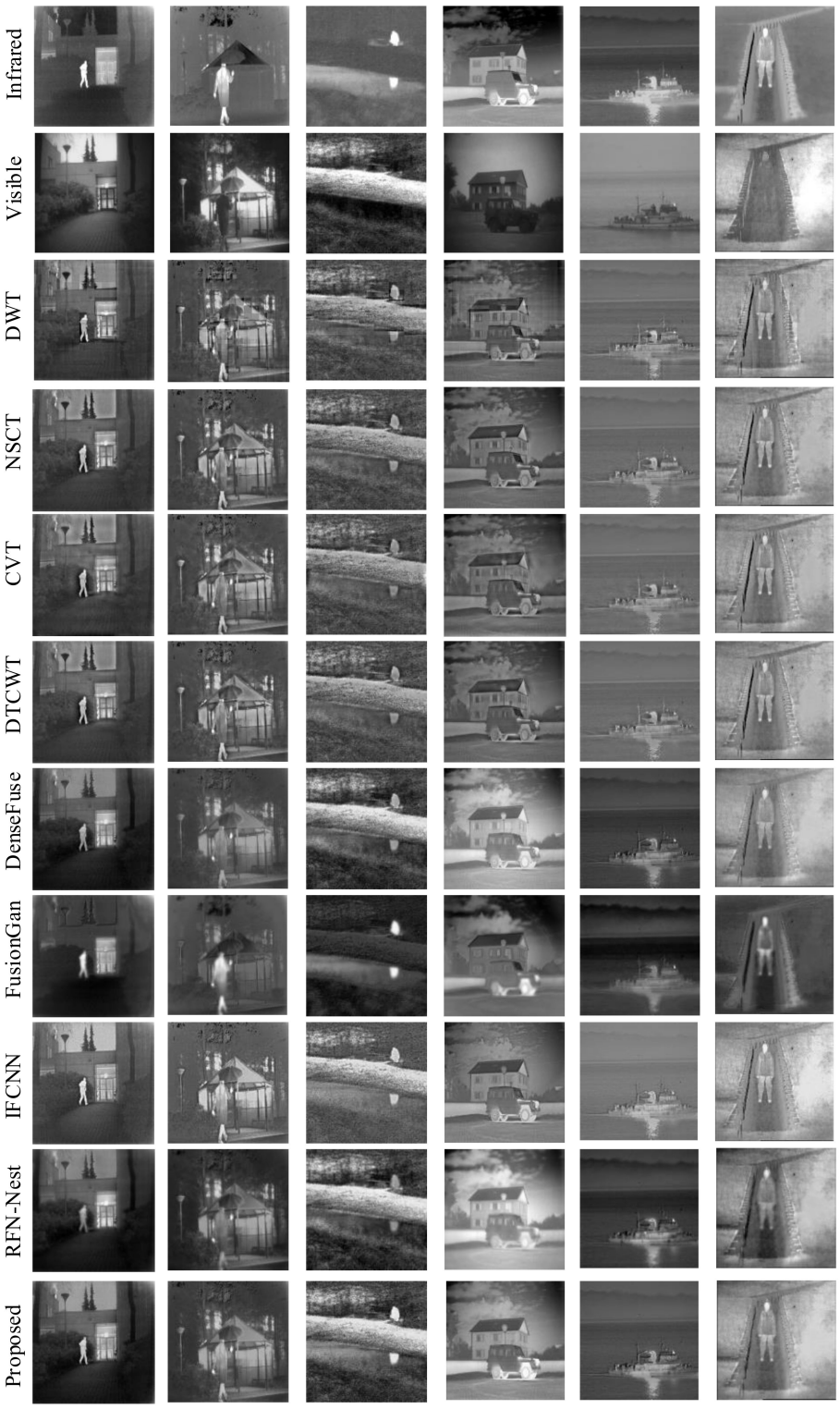

Figure 6. Subjective comparison between the proposed method and other state-of-art methods on 6 pairs of images from TNO dataset.

In order to evaluate the proposed method more comprehensively, nine methods are compared in terms of six metrics on 21 pairs of images from TNO datasets. The comparative results are given in Figure 7. In the view of informativeness (EN) and texture richness (SD) of fused results, we can see that the proposed method and RFN-nest are in the first echelon, while FusionGan and NSCT perform poorly because FusionGan only preserves details of

the infrared images and NSCT has information loss during the transformation process. IFCNN mainly focuses on the pixel error between the fused image and the original image, while the proposed method pays attention to more comprehensive information such as structural information and texture information. Therefore, the proposed method is inferior to IFCNN on RMSE. Since the SSIM in our experiments denotes the similarity measure between the fused image and the visible image, the performance of FusionGan is very poor, while other methods have good performance. On MI and SCD that measure the amount of information transferred from the source images to the fused images, FusionGan and traditional methods perform poorly; this is because FusionGan's generative adversarial networks are not mature enough, and traditional methods will lose a lot of information during transform. In contrast, the proposed method performs best on MI and SCD, which is due to the contribution of the multi-modal feature self-adaptive fusion strategy and transformer fusion block.

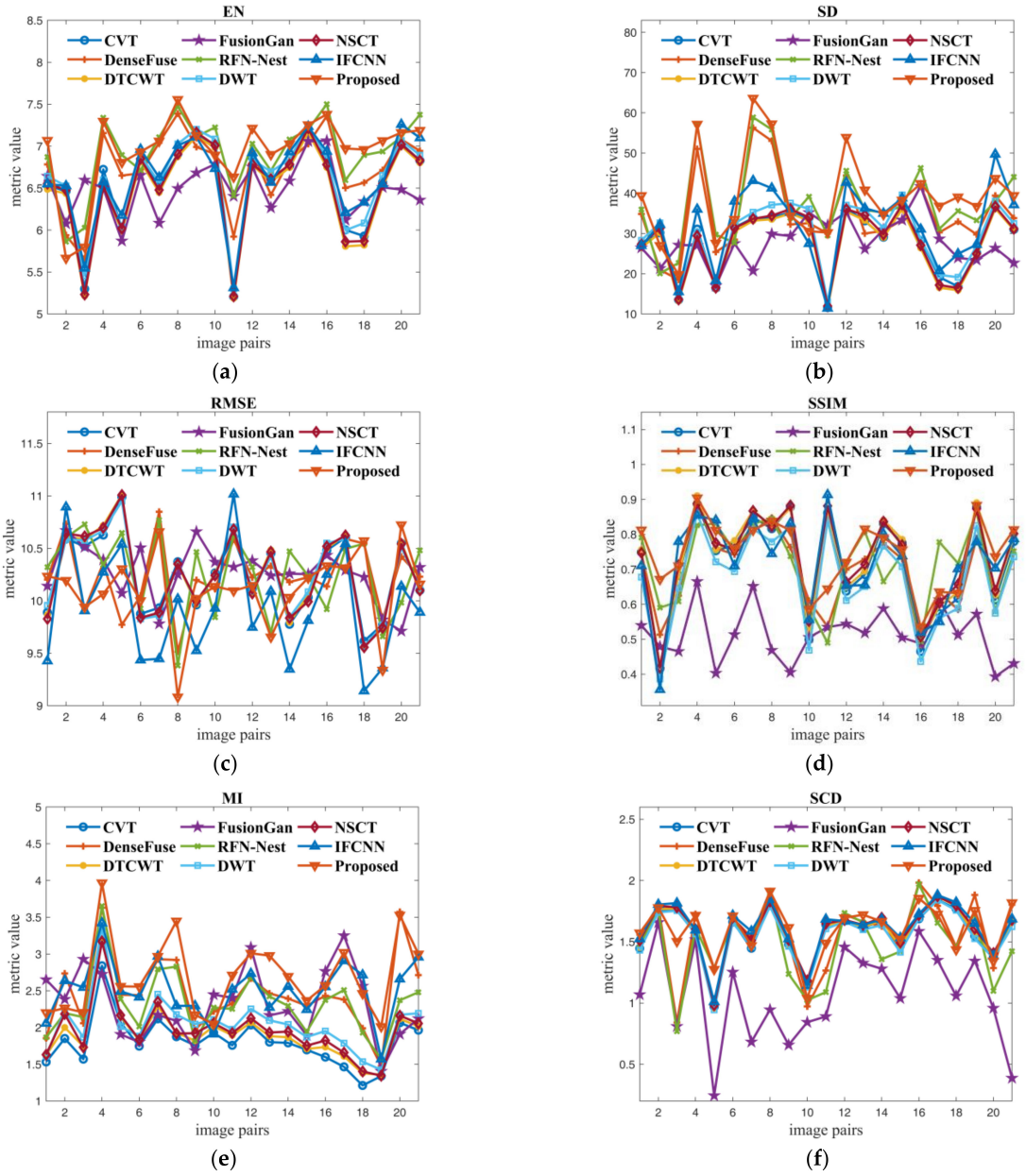

Figure 7. Quantitative comparison of the proposed method against eight existing fusion methods by using six metrics on 21 pairs of images from TNO dataset. (**a**–**f**) are the comparison results of the nine methods on EN, SD, RMSE, SSIM, MI and SCD respectively.

The average comparison results of the proposed method with eight other methods on the TNO dataset are given in Table 4. It can be seen that the proposed method has achieved the best value in four indexes (EN, SD, SSIM, MI), the second-best result in one index (RMSE), and the third-best result in one index (SCD). This demonstrates that with the help of focal self-attention, the proposed method outperforms the existing methods in transmitting and preserving information from the source images to the fused results. Moreover, since the multi-modal feature self-adaptive fusion strategy is efficient for feature fusion, the fused images obtained by the proposed method contain more natural and sharper content, which is more suitable for human perception.

Table 4. Quantitative results on 21 pairs of images from TNO dataset. The best values, the second-best values and the third-best values are indicated in bold, red and blue, respectively.

Method	EN	SD	SCD	SSIM	MI	RMSE
DWT	6.5964	29.6984	1.5552	0.6745	2.051	10.2507
NSCT	6.5107	29.1414	1.6018	0.7318	1.9575	10.2494
CVT	6.5371	28.1056	1.5735	0.7149	1.8108	10.2445
DTCWT	6.4773	27.4436	1.5794	0.7237	1.9163	10.2514
DenseFuse	6.7378	34.7623	1.5599	0.7001	2.4726	10.2377
FusionGan	6.4919	27.9282	1.0647	0.514	2.3137	10.2673
IFCNN	6.6265	31.869	**1.6153**	0.7155	2.5111	**9.939**
RFN-Nest	6.9271	37.7383	1.4686	0.7151	2.3238	10.2609
Proposed	**6.9519**	**39.3726**	1.6011	**0.7466**	**2.7028**	10.1066

4.4.2. Fusion Results Analysis on FLIR and LLVIP

Further, to verify the generalization performance and the fusion ability for complex scenes, we conduct comparative experiments on the FLIR and LLVIP datasets. The major scenes of FLIR and LLVIP are roads, including pedestrians, vehicles, and other targets. These images also include different lighting conditions, such as day and night. The intuitive comparison fusion results of the proposed method and the other eight methods on the FLIR and LLVIP are shown in Figure 8. It is observed that DWT, NSCT, CVT, and DTCWT cannot suppress noise, which results in poor visual effect. Densefuse does not perform well on FLIR, and some texture details of visible images are lost. The fused images obtained by FusionGan have very blurred edges. The images obtained by IFCNN have high brightness and a certain deformation. RFN-nest performs well on FLIR with a sharpened background and prominent edges of people and vehicles, but there are blurred edges and unobtrusive human targets on LLVIP. Compared with the above methods, our fusion framework has competitive performance on both FLIR and LLVIP datasets, with rich background texture, clear edge contour, salient thermal objects, and high contrast. The rich texture and clear edges are attributed to the ingenious design of the multi-modal self-adaptive fusion strategy, and the salient infrared objects and high contrast are achieved under the combined effect of the fusion strategy and saliency information.

For institutively comparing the fusion ability of the nine methods on the FLIR dataset, the average performances of these methods on the FLIR dataset are shown in Table 5. The best values, the second-best values, and the third-best values are indicated in bold, red, and blue. For metrics of EN, SD, SCD, and SSIM, RFN-nest achieves the best results because the model is trained on KAIST datasets similar to FLIR, so it can also have strong robustness when tested on FLIR. Our method is second only to RFN-nest but still far better than traditional methods because the self-adaptive fusion strategy helps the fusion of features to be more comprehensive and reduces information loss. Our method achieves the best results on MI, which indicates that the proposed method is robust to the transmission of information during the fusion process of complex scenes. This is because the transformer fusion block not only considers local features but also transmits the global context when fusing feature information, making the subsequent image reconstruction process more robust to obtain higher-quality results.

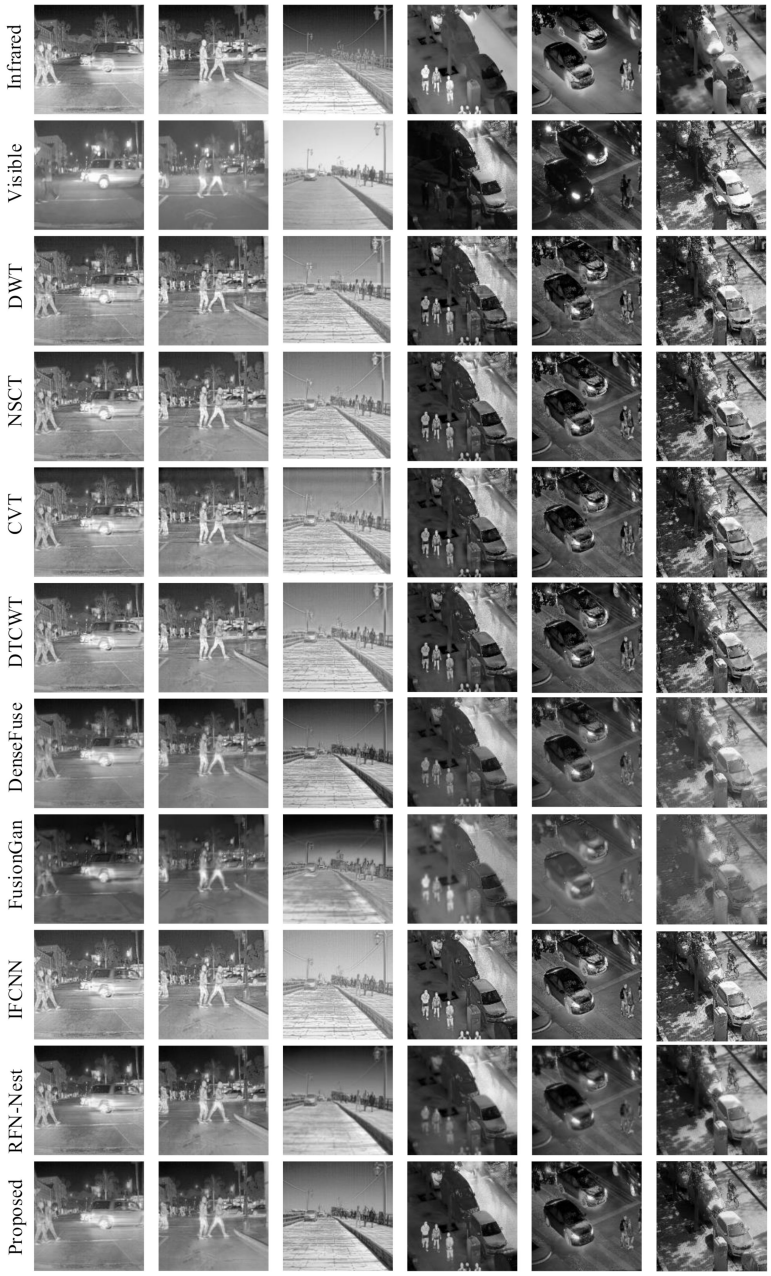

Figure 8. Subjective comparison between the proposed method and other state-of-art methods on 6 pairs of images from FLIR and LLVIP datasets. The first three pairs of images are selected from the FLIR dataset, and the last three pairs of images are selected from LLVIP dataset.

Table 5. Quantitative results on 50 pairs of images from FLIR dataset. The best values, the second-best values and the third-best values are indicated in bold, red and blue, respectively.

Method	EN	SD	SCD	SSIM	MI	RMSE
DWT	7.2733	43.5689	1.1043	0.5272	3.1864	10.1828
NSCT	7.2323	42.6921	1.1386	0.5722	3.1431	10.183
CVT	7.3018	43.8735	1.1106	0.5383	2.8208	10.1606
DTCWT	7.2096	41.9806	1.1051	0.5622	3.0607	10.1848
DenseFuse	7.4005	52.7755	1.2739	0.6014	3.7431	10.0585
FusionGan	7.3209	47.5473	0.5866	0.5947	3.3429	10.2062
IFCNN	7.185	40.4748	1.136	0.6112	3.3764	**9.6982**
RFN-Nest	**7.5439**	**58.9438**	**1.3638**	**0.7006**	3.5457	10.1004
Proposed	7.4811	52.4626	1.3064	0.6435	**3.7815**	10.103

Table 5 also shows that we can not achieve the best fusion performance for some images in FLIR. In order to find the reason for the decline in fusion performance, we put a pair of representative images in the third column in Figure 8 (termed Image 3). The objective evaluation results of Image 3 are shown in Table 6. Table 6 only shows the comparison between the proposed method and DenseFuse and RFN-Nest, which have good performance. Our method only achieves the best value on SCD, while other metrics are not optimal. There are many images in the FLIR dataset with open scenes and insignificant structural features, such as Image 3. Both DenseFuse and RFN-nest mainly constrain the pixels between the fused image and the source images and only consider local information during fusion processing, which can have better fusion performance for Figure 3. While the proposed method tends to fuse images more comprehensively in terms of structural features and saliency features and has more advantages in processing images with obvious structures and clear edges.

Table 6. Quantitative evaluation of images in the third column in Figure 8. This table only shows the comparison between the proposed method and DenseFuse and RFN-Nest which have good performance.

Method	EN	SD	SCD	SSIM	MI	RMSE
DenseFuse	7.6602	56.5542	1.455	0.8319	4.2070	10.3664
RFN-nest	7.7641	60.1371	1.4058	0.5231	3.8915	10.2250
Proposed	7.5851	50.2960	1.4921	0.7507	3.4256	10.3871

The average performances of these methods on the LLVIP dataset are shown in Table 7. In the test of the LLVIP dataset, our method obtains the best value in three indexes (SCD, SSIM, MI), the second-best result in two indexes (SD, RMSE), and the third best result in one index (EN). Our method has achieved the best comprehensive results among these comparative methods, which demonstrates that the proposed method can perform well on different datasets and can preserve abundant texture and salience information from the source images.

Table 7. Quantitative results on 50 pairs of images from LLVIP dataset. The best values, the second-best values and the third-best values are indicated in bold, red and blue, respectively.

Method	EN	SD	SCD	SSIM	MI	RMSE
DWT	7.1622	46.1461	1.3735	0.5756	2.834	10.0145
NSCT	7.1521	45.3417	1.4294	0.6318	2.7803	10.0127
CVT	7.1547	44.6028	1.3968	0.5984	2.5841	10.0227
DTCWT	7.1415	44.4009	1.3993	0.6177	2.701	10.0087
DenseFuse	7.1227	41.7651	1.4932	0.6027	3.2511	9.9555
FusionGan	6.7159	30.4424	0.7712	0.4887	2.7055	10.1182
IFCNN	**7.3332**	**48.2112**	1.4468	0.5971	3.1225	**9.7326**
RFN-Nest	7.2456	45.1502	1.5039	0.5533	2.8533	9.9285
Proposed	7.1779	46.9661	**1.5255**	**0.6875**	**3.2941**	9.9281

The above verification and comparison experiments show that the proposed MFST can generate fused images with a large amount of information and rich textures and is also robust to the fusion of complex scenes, which is mainly due to the following two points: (1) Automatic encoder for multi-modal feature extraction; (2) The design of multi-modal feature self-adaptive fusion strategy and transformer fusion block in the feature fusion stage greatly improve the efficiency of information transmission.

5. Conclusions

In order to improve the fusion quality of images and the fusion efficiency of multi-modal features, a novel infrared and visible image fusion method (MFST) is developed in this paper. The self-adaptive fusion strategy was firstly designed to more effectively

utilize the information of multi-modal features. Then, the focal self-attention mechanism facilitates the model to pay attention to both local and global information in the fusion process. Finally, the introduction of saliency information enables the fusion results to preserve more salient objects of infrared images. Three different infrared and visible image datasets were used to verify the effectiveness of the proposed method, and the results show that our method has a strong generalization ability and good ability to fuse complex scenes. Through the comparative experimental analysis with the current eight popular methods, it shows that the fusion quality and fusion efficiency of the method in this paper are better than the eight popular eight methods, which confirms the superiority of the proposed method.

However, the proposed method cannot perform the best on some images, which indicates that our model has much room for improvement. There are several key issues deserving to be further studied: (1) Are there other modal featuress that can be extracted? (2) How to lighten the model so that it can fuse higher resolution images? (3) Could choosing other FR-IQA metrics as the basis for the loss function improve the quality of the results? In the future, we will concentrate on these issues, improve and optimize the model, and make further contributions to this research topic.

Author Contributions: X.L. conceived of and designed the experiments and wrote the paper; H.G. performed the experiments and original draft preparation; Y.X. analyzed the data; Y.A. offered the data curation; D.G. provided the review and editing; Q.M. supervised the study and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of the People's Republic of China: 2018YFC0807504, B019030051; Ministry of Education of the People's Republic of China: 20101216855; The Key R&D Projects of Qingdao Science and Technology Plan: 21-1-2-18-xx.

Data Availability Statement: Not applicable.

Acknowledgments: We sincerely thank the authors of DWT, DTCWT, CVT, NSCT, DenseFuse, FusionGan, IFCNN, and RFN-Nest for providing their algorithm codes to facilitate the comparative experiments, and thanks to the authors of reference [38] for collecting the evaluation metrics used in the comparative experiments in this paper.

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

References

1. Hu, H.M.; Wu, J.; Li, B.; Guo, Q.; Zheng, J. An Adaptive Fusion Algorithm for Visible and Infrared Videos Based on Entropy and the Cumulative Distribution of Gray Levels. *IEEE Trans. Multimed.* **2017**, *19*, 2706–2719. [CrossRef]
2. Zhao, W.; Lu, H.; Wang, D. Multisensor Image Fusion and Enhancement in Spectral Total Variation Domain. *IEEE Trans. Multimed.* **2018**, *20*, 866–879. [CrossRef]
3. 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 (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9627–9636.
4. Kou, F.; Wei, Z.; Chen, W.; Wu, X.; Wen, C.; Li, Z. Intelligent Detail Enhancement for Exposure Fusion. *IEEE Trans. Multimed.* **2018**, *20*, 484–495. [CrossRef]
5. Arroyo, S.; 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]
6. 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]
7. 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]
8. 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.* **2022**, *60*, 1–14. [CrossRef]
9. Ying, J.; Shen, H.-L.; Cao, S.-Y. Unaligned Hyperspectral Image Fusion via Registration and Interpolation Modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
10. Zhu, Z.; Wei, H.; Hu, G.; Li, Y.; Qi, G.; Mazur, N. A Novel Fast Single Image Dehazing Algorithm Based on Artificial Multiexposure Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–23. [CrossRef]
11. Paramanandham, N.; Rajendiran, K. Infrared and Visible Image Fusion Using Discrete Cosine Transform and Swarm Intelligence for Surveillance Applications. *Infrared Phys. Technol.* **2018**, *88*, 13–22. [CrossRef]

12. Wang, G.; Li, W.; Gao, X.; Xiao, B.; Du, J. Functional and Anatomical Image Fusion Based on Gradient Enhanced Decomposition Model. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [CrossRef]
13. Li, G.; Lin, Y.; Qu, X. An Infrared and Visible Image Fusion Method Based on Multi-Scale Transformation and Norm Optimization. *Inf. Fusion* **2021**, *71*, 109–129. [CrossRef]
14. Jian, L.; Yang, X.; Zhou, Z.; Zhou, K.; Liu, K. Multi-Scale Image Fusion through Rolling Guidance Filter. *Future Gener. Comput. Syst.* **2018**, *83*, 310–325. [CrossRef]
15. Maqsood, S.; Javed, U. Multi-Modal Medical Image Fusion Based on Two-Scale Image Decomposition and Sparse Representation. *Biomed. Signal Process. Control* **2020**, *57*, 101810. [CrossRef]
16. Zhang, Q.; Liu, Y.; Blum, R.S.; Han, J.; Tao, D. Sparse Representation Based Multi-Sensor Image Fusion for Multi-Focus and Multi-Modality Images: A Review. *Inf. Fusion* **2018**, *40*, 57–75. [CrossRef]
17. Li, Q.; Han, G.; Liu, P.; Yang, H.; Wu, J.; Liu, D. An Infrared and Visible Image Fusion Method Guided by Saliency and Gradient Information. *IEEE Access* **2021**, *9*, 108942–108958. [CrossRef]
18. Zhang, X.; Ma, Y.; Fan, F.; Zhang, Y.; Huang, J. Infrared and Visible Image Fusion via Saliency Analysis and Local Edge-Preserving Multi-Scale Decomposition. *JOSA A* **2017**, *34*, 1400–1410. [CrossRef]
19. Li, H.; Wu, X.J. Infrared and Visible Image Fusion Using Latent Low-Rank Representation. *arXiv* **2022**, arXiv:1804.08992v5.
20. 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]
21. Vanmali, A.V.; Gadre, V.M. Visible and NIR Image Fusion Using Weight-Map-Guided Laplacian–Gaussian Pyramid for Improving Scene Visibility. *Sādhanā* **2017**, *42*, 1063–1082. [CrossRef]
22. Yan, H.; Zhang, J.-X.; Zhang, X. Injected Infrared and Visible Image Fusion via $_1$ Decomposition Model and Guided Filtering. *IEEE Trans. Comput. Imaging* **2022**, *8*, 162–173. [CrossRef]
23. Zhou, X.; Wang, W. Infrared and Visible Image Fusion Based on Tetrolet Transform. In Proceedings of the 2015 International Conference on Communications, Signal Processing, and Systems, Tianjin, China, 23–24 October 2015; pp. 701–708.
24. Yang, B.; Yang, C.; Huang, G. Efficient Image Fusion with Approximate Sparse Representation. *Int. J. Wavelets Multiresolution Inf. Process.* **2016**, *14*, 1650024.
25. Veshki, F.G.; Ouzir, N.; Vorobyov, S.A.; Ollila, E. Multimodal Image Fusion via Coupled Feature Learning. *Signal Process.* **2022**, *200*, 108637. [CrossRef]
26. 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]
27. Liu, Y.; Dong, L.; Xu, W. Infrared and Visible Image Fusion via Salient Object Extraction and Low-Light Region Enhancement. *Infrared Phys. Technol.* **2022**, *124*, 104223. [CrossRef]
28. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-Focus Image Fusion with a Deep Convolutional Neural Network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]
29. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and Visible Image Fusion with Convolutional Neural Networks. *Int. J. Wavelets Multiresolution Inf. Process.* **2017**, *16*, 1850018. [CrossRef]
30. 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**, 1. [CrossRef]
31. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. FusionDN: A Unified Densely Connected Network for Image Fusion. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12484–12491. [CrossRef]
32. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.-P. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef]
33. Yang, Y.; Kong, X.; Huang, S.; Wan, W.; Liu, J.; Zhang, W. Infrared and Visible Image Fusion Based on Multiscale Network with Dual-Channel Information Cross Fusion Block. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–7.
34. 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]
35. Li, H.; Wu, X.J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef] [PubMed]
36. 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]
37. 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]
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
39. 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 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929v2.
40. Liu, N.; Zhang, N.; Wan, K.; Shao, L.; Han, J. Visual Saliency Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 4722–4732.

41. 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 (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
42. Li, J.; Yan, Y.; Liao, S.; Yang, X.; Shao, L. Local-to-Global Self-Attention in Vision Transformers. *arXiv* **2021**, arXiv:2107.04735v1.
43. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal Self-Attention for Local-Global Interactions in Vision Transformers. *arXiv* **2021**, arXiv:2107.00641v1.
44. 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]
45. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
46. Toet, A. TNO Image Fusion Dataset. 2014. Available online: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029 (accessed on 10 December 2021).
47. 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 (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 3496–3504.
48. Ma, J.; Ma, Y.; Li, C. Infrared and Visible Image Fusion Methods and Applications: A Survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
49. Niu, Y.; Xu, S.; Wu, L.; Hu, W. Airborne Infrared and Visible Image Fusion for Target Perception Based on Target Region Segmentation and Discrete Wavelet Transform. *Math. Probl. Eng.* **2012**, *2012*, 275138. [CrossRef]
50. Lewis, J.J.; O'Callaghan, R.J.; Nikolov, S.G.; Bull, D.R.; Canagarajah, N. Pixel- and Region-Based Image Fusion with Complex Wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [CrossRef]
51. Nencini, F.; Garzelli, A.; Baronti, S.; Alparone, L. Remote Sensing Image Fusion Using the Curvelet Transform. *Inf. Fusion* **2007**, *8*, 143–156. [CrossRef]
52. Yin, S.; Cao, L.; Tan, Q.; Jin, G. Infrared and Visible Image Fusion Based on NSCT and Fuzzy Logic. In Proceedings of the 2010 IEEE International Conference on Mechatronics and Automation, Montreal, QC, Canada, 6–9 July 2010; pp. 671–675.
53. 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]
54. 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]