sensors

MDPI

Article

FERFusion: A Fast and Efficient Recursive Neural Network for Infrared and Visible Image Fusion

Kaixuan Yang [1,2,3,4], Wei Xiang [1,2], Zhenshuai Chen [1,2,3,4] and Yunpeng Liu [1,2,*]

[1] Key Laboratory of Opto-Electronic Information Processing, Chinese Academy of Sciences, Shenyang 110016, China; yangkaixuan@sia.cn (K.Y.); xiangwei@sia.cn (W.X.); chenzhenshuai@sia.cn (Z.C.)
[2] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: ypliu@sia.cn

Abstract: The rapid development of deep neural networks has attracted significant attention in the infrared and visible image fusion field. However, most existing fusion models have many parameters and consume high computational and spatial resources. This paper proposes a fast and efficient recursive fusion neural network model to solve this complex problem that few people have touched. Specifically, we designed an attention module combining a traditional fusion knowledge prior with channel attention to extract modal-specific features efficiently. We used a shared attention layer to perform the early fusion of modal-shared features. Adopting parallel dilated convolution layers further reduces the network's parameter count. Our network is trained recursively, featuring minimal model parameters, and requires only a few training batches to achieve excellent fusion results. This significantly reduces the consumption of time, space, and computational resources during model training. We compared our method with nine SOTA methods on three public datasets, demonstrating our method's efficient training feature and good fusion results.

Keywords: infrared-visible image fusion; transformer; deep learning

Citation: Yang, K.; Xiang, W.; Chen, Z.; Liu, Y. FERFusion: A Fast and Efficient Recursive Neural Network for Infrared and Visible Image Fusion. *Sensors* **2024**, *24*, 2466. https://doi.org/10.3390/s24082466

Academic Editors: Daquan Yang and Jinhui Chen

Received: 5 March 2024
Revised: 7 April 2024
Accepted: 8 April 2024
Published: 11 April 2024

1. Introduction

Due to technical limitations and the impact of the shooting environment, a single image captured by the same device often fails to provide a comprehensive description of the entire scene [1]. Therefore, image fusion techniques are emerging, in which infrared and visible image fusion is the most widely used. Infrared and visible image fusion (IVIF) produces fused images with complementary characteristics and richer information than either modality alone [2–4]. The resulting images are visually appealing and, more importantly, beneficial for practical applications such as medical diagnostics [5], remote sensing [6], automotive assistance [7], video surveillance [8], and wildfire monitoring [9].

Traditional infrared and visible image fusion methods are dedicated to finding the optimal representation of common features across modalities and designing appropriate weights for fusion [10]. For instance, multi-scale transform methods start by breaking down source images into features represented at multiple scales [11–14]. Similarly, sparse representation-based models begin by dividing source images into patches using the sliding window technique. Subsequently, these image patches are processed by sparse coding to derive sparse representation coefficients [15–17]. Low-rank representation (LRR) methods are also utilized to extract saliency features from source images, too [18,19]. After obtaining image features or representation coefficients, they will be reconstructed to produce the final fusion results according to delicately designed fusion rules. These fusion algorithms generally have the advantages of fast fusion speed and weak data dependence. However, manual feature selection and extraction are complex and cannot obtain good performance [20].

Deep learning has achieved tremendous success in various artificial intelligence applications in recent years, owing to its powerful non-linear fitting and feature extraction capabilities [3]. Researchers utilize public IVIF datasets or create [21,22], design, and build their own various deep neural network models for learning modal features and fusion strategies. After extensive training, these models can effectively extract information from infrared and visible images, generating information-rich fused images.

Unfortunately, many of the current advanced algorithms require a vast amount of space to store network parameters, as shown in Table 1, and their slow processing speeds make it difficult to serve downstream visual tasks. Compared to traditional approaches such as subspace analysis and sparse representation, deep learning-based methods have achieved notable improvements in fusion effects. However, their development has encountered bottlenecks: deep learning methods require the stacking of numerous convolutional modules to learn the common features of infrared and visible images, which have significant appearance differences; training these massive networks also necessitates a large number of strictly matched image pairs. Although there are public datasets like MSRS [23], M3FD [21], and LLVIP [24], they are still insufficient for models with millions or even tens of millions of parameters, such as RFN-Nest [25], CDDFuse [26], PIAFusion [23].

Table 1. Parameters of current advanced fusion models. The model we propose has the smallest number of parameters.

Model	Parameters *	Model	Parameters	Model	Parameters
DenseFuse	74,193	GANMcC	1,864,129	ReCoNet	7527
MFEIF	371,140	SDNet	67,091	LRRNet	196,816
PMGI	42,017	RFN-Nest	19,165,952	CDDFuse	1,188,272
U2Fusion	659,217	TarDAL	296,577	MetaFusion	811,714
DIDFuse	260,935	PIAFusion	1,266,595	**Ours**	**5885**

* All values in the 'Parameters' column represent the number of trainable parameters in each model.

LRRNet [27] formulates the fusion task mathematically, and establishes a connection between its optimal solution and the network architecture that can implement it. It proposes a lightweight fusion network which avoids the time-consuming empirical network design by a trial-and-test strategy. However, the model has a large number of parameters, which limits further improvements in its speed. ReCoNet also explores a lightweight fusion network, inspired by traditional fusion knowledge priors [28]. It utilizes a simple max pooling channel and an average pooling channel, enabling the model to focus more on the textural features within the image. This approach is straightforward and effective, yet it might not be as flexible in feature extraction since the convolutional module in the model processes images treated with a simple attention mechanism rather than the source images. Moreover, due to their inherent characteristics, convolutional neural networks have a limited receptive field and lack the capability to perceive global features. This limitation prevents the establishment of long-range dependencies between features and leads to poor quality of the final fused image.

On the other hand, from the architectural perspective, visual transformers [29], with their unique self-attention mechanism that encodes global positional relationships, exhibit excellent global feature extraction capabilities. However, the computational cost of methods based on transformers is high, leaving room for further improvements considering the efficiency and performance trade-off in image fusion architectures. Therefore, we contemplate building upon ReCoNet by incorporating transformer modules [30] to enhance the model's feature extraction capabilities and employing techniques like depthwise separable convolutions to reduce the model's parameter count.

Overall, we employ two separate branches to extract features from infrared and visible images, respectively. Following [31,32], the exchange of a variety of information between different feature extraction channels can pre-fuse gradient and intensity information, serving as an enhancement of information before the next extraction. Hence, we have also

introduced a parameter-sharing module between the two branches. To capture cross-modal, long-distance dependencies, we leverage transformer as our parameter-sharing module's backbone. After obtaining the features, we use a convolutional module to further fuse and integrate information, ultimately outputting the fused image. Considering that different image features may have different scales, using traditional convolution layers with the same kernels cannot effectively represent source image information. In addition, to ensure the model's lightweight nature, we employ three-channel dilated convolutions with different receptive fields for further processing of the features. The activation function for all layers is ReLU, except for the last layer, which uses Tanh [1].

To sum up, we proposes a lightweight recursive network, which can be trained on the public data set MSRS in 30 batches to fit, efficiently extract source image features, and generate fusion images with good visual effects. Specifically, we designed an attention module that uses traditional fusion knowledge prior guidance and channel attention to track the respective salience areas of the infrared and visible images to fit the network quickly. We also use an attention module with shared parameters to promote the early fusion of features; finally, we use a parallel dilated convolution module to learn features of different scales with different receptive fields. We iteratively train a set of parameters of this simple fusion module. This cyclic process reduces the number of network parameters and iteratively improves the image fusion quality.

2. Related Work

2.1. Infrared and Visible Image Fusion Based on Deep Learning

Due to the powerful nonlinear fitting capabilities, neural networks have been widely applied in infrared and visible image fusion, achieving performance far superior to traditional methods. Currently, the methods of infrared and visible image fusion based on deep learning can generally be divided into four types: CNN-based methods [33–36], GAN-based methods [37–41], AE-based [1,42–45] methods, and transformer-based [32,46–49] methods. CNN-based methods tend to focus on the design of loss functions, forcing the model to generate images that contain as much information from the source images as possible. GAN-based methods utilize an adversarial game between the generator and the discriminator to optimize the model. The generator produces images, and the discriminator judges whether the images are authentic or generated. When the discriminator cannot accurately judge, it is considered that the generated images meet the requirements. GAN-based models have a strong generation capability and can produce entirely new images, but their training process is complex and not robust enough [50–52]. AE-based methods force the decoder to generate images as close to the source images as possible during the training phase. When the two images are sufficiently close (optimal training loss), it is considered that the encoder can extract features from the source images well. In the testing phase, some specific fusion rules are applied to the results of the encoder for fusion and then decoded by the decoder to obtain the fused image. transformer-based methods have become popular in recent years, introducing attention mechanisms to encode the global positional relationships of images and overcoming the limitations of the convolution operation's receptive field, but they also bring a high computational burden.

In recent years, some scholars have not limited themselves to the task of image fusion itself but have combined it with upstream and downstream tasks to guide image fusion or make it more robust and targeted. ReCoNet [28] introduces a geometric correction module to perform geometric compensation on the input pairs of infrared and visible images, thereby improving the model's robustness. Similarly, RFNet [53] trains both multimodal image correction and fusion in a coarse-to-fine manner. SeAFusion [54] combines image fusion with downstream task of semantic segmentation, constructing a semantic loss that allows high-level semantic information to flow to the image fusion module, improving the effects of image fusion. IRFS [55] combines image fusion with saliency target detection, using the fused image generated by the fusion subnetwork as a third modality to drive

precise prediction of the saliency map. Both enhance each other and are optimized together, achieving good experimental results.

2.2. Vision Transformer

The transformer was initially proposed by Vaswani and applied in the field of natural language processing [56]. It was first introduced to computer vision in 2020 [57]. IFT [58] was the first to introduce transformer into the IVIF field, proposing a multi-scale fusion strategy based on transformer to consider local and global information. Subsequently, PPT Fusion [59] improved upon the transformer by designing a patch transformer that converts images into a series of patches and then performs position encoding on each patch to extract local representations. It also designed a pyramid transformer to extract global information from the images.

YDTR [49] introduces a Y-shaped dynamic transformer architecture for infrared and visible image fusion. It designs a dynamic transformer module to extract local and global information simultaneously. SwinFusion [32] presents a universal image fusion method based on multi-modal long-range learning and the Swin transformer. It introduces self-attention-based intra-modal and inter-modal fusion units to capture global relationships within and between modalities effectively. CDDFuse [26] incorporates attention mechanisms into the image fusion framework. It uses CNN to extract modality-specific fine-grained features from infrared and visible images and employs the transformer's long and short-term attention to extract modality-shared features, resulting in improved feature extraction and enhanced visual quality of the fused images. DATFuse [47] proposes an end-to-end dual transformer model to avoid manually designing complicated activity-level measurement and fusion strategies.

Until now, numerous transformer-based models have achieved excellent results in tasks such as image classification [60], object detection [61], segmentation [62], tracking [63,64], and multimodal learning [65]. For instance, in low-level computer vision tasks (e.g., denoising, super-resolution, and deraining), a pre-trained image processing transformer has outperformed CNN models [66]. Ref. [67] proposed a cross-scale mixing attention transformer-based model for hyperspectral and multispectral image fusion and classification. FD-Net [68] designed a feature distillation network for oral squamous cell carcinoma lymph node segmentation, which can effectively assist pathologists in disease screening and reduce workload. However, the application of transformers for mobile networks(with limited model size) significantly lags behind that for large networks. This is mainly because the computational overhead brought by most transformer models is not affordable for mobile networks.

Considering the heavy computational load of spatial self-attention, Wu et al. proposed a lightweight transformer architecture called LT for natural language processing tasks on mobile devices [69]. The model's parameters are significantly reduced without sacrificing accuracy by incorporating long-short-range attention and a flattened feedforward network. Restormer [30] enhances the representation capabilities of transformers for high-resolution images by improving gated convolutional networks and multi-head attention modules. We drew inspiration from Restormer's improvements on transformer architecture, adapting its concepts from image restoration to the task of image fusion.

3. The Proposed Method

3.1. Tri-Phase Attention Module

Texture features, such as edges, targets, and contours, play a crucial role in the fusion process. However, as the network deepens, texture features gradually degrade, leading to blurred targets and loss of details in the fused image. Existing work has focused on designing various attention mechanisms or increasing network width (such as dense and residual connections) to address this issue. In fact, some attention mechanisms have difficulty characterizing contextual features from the source images [28]; the increasingly large model architectures lead to a vast demand for computational resources and memory.

We propose a tri-phase attention module that accelerates network convergence, suppresses information loss through an attention layer with traditional fusion knowledge priors, and enhances feature extraction capabilities through a channel attention layer.

The module includes a maximum pooling layer, an average pooling layer, a transformer attention layer, and a bias-free convolution layer. The maximum pooling and average pooling draw inspiration from traditional fusion strategies. The maximum and average values of each pixel position in the two images are calculated and then stacked together with the output of the transformer attention layer to serve as the input for the convolution layer.

Let $\mathcal{A}$ represent the tri-phase attention layer, and I_a and I_b represent the two input images, respectively. The following equation can represent this process:

$$\mathcal{A}(I_a, I_b) = \theta_{\mathcal{A}} * [\max(I_a, I_b), \text{avg}(I_a, I_b), \text{trans}(I_a, I_b)] \quad (1)$$

where $*$ represents the convolution operation, $\theta_{\mathcal{A}}$ represents the parameters of the convolution layer, and [] represents the concatenation operation. As shown in Figure 1, the network calculates the attention maps σ_x, σ_y from the input image set $\{x, u, y\}$ according to the following equation:

$$\sigma_x = \mathcal{A}_x(x, u_i) \quad \sigma_y = \mathcal{A}_y(y, u_i) \quad (2)$$

where $\mathcal{A}_x$ and $\mathcal{A}_y$ represent the infrared and visible light attention layers, respectively, and u_i denotes the fusion result obtained from the previous iteration.

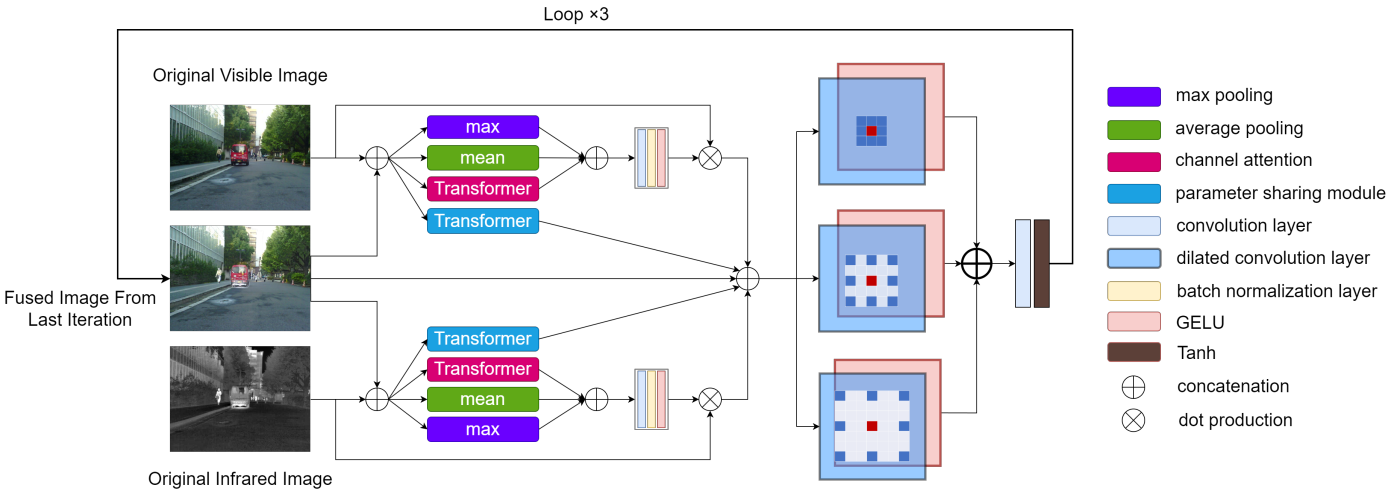

Figure 1. Overall flowchart of the model.

Thanks to the tri-phase attention layer, our network can focus more on the features of each modality. Unlike traditional bi-phase attention layers (which only use max pooling and average pooling) [28], the tri-phase attention layer not only allows the network to converge quickly but also, with the introduction of channel attention, enables the network to more flexibly perceive the relationships between different channels, thereby enhancing the fusion effect [30].

3.2. Transformer Block

The structure of our transformer module is shown in Figure 2. It generally comprises two modules: a multi-head transposed attention module and a gated forward propagation module. Below, we will discuss the computational principles of these two modules in detail.

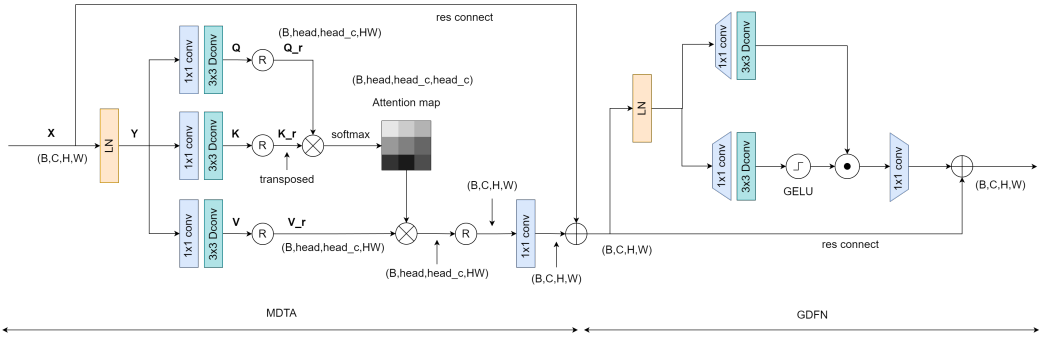

Figure 2. Structural diagram of the parameter sharing module and the transformer module.

3.2.1. Multi-Dconv Head Transposed Attention

In traditional transformers, the computational load primarily originates from the self-attention layer. The time and space complexity of the dot product between queries and keys grows quadratically with the input's spatial resolution. For instance, for an image of size $W \times H$, the complexity would be $\mathcal{O}(W^2H^2)$, which is unacceptable for high-resolution images. However, the multi-head transposed attention module has a complexity that grows linearly, focusing on computing attention between channels rather than between spatial locations. This aligns perfectly with our image fusion task, which concentrates on the relationships between images of different modalities.

On the other hand, depthwise convolution further reduces the computational load and the number of parameters. Depthwise convolution performs convolution operations on each input data channel rather than convolutions across all channels.

For the input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, after layer normalization, we obtain $\mathbf{Y} = \mathrm{LN}(\mathbf{X})$. Then, through 1×1 convolution and 3×3 depthwise convolution, we compute our query $(\mathbf{Q})$, key $(\mathbf{K})$, and value $(\mathbf{V})$ matrices:

$$\mathbf{Q} = \mathbf{W}_d^Q \mathbf{W}_p^Q \mathbf{Y} \tag{3}$$

$$\mathbf{K} = \mathbf{W}_d^K \mathbf{W}_p^K \mathbf{Y} \tag{4}$$

$$\mathbf{V} = \mathbf{W}_d^V \mathbf{W}_p^V \mathbf{Y} \tag{5}$$

where $\mathbf{W}_p^{(\cdot)}$ and $\mathbf{W}_d^{(\cdot)}$ represent the 1×1 convolution and the 3×3 depthwise convolution, respectively. Next, we reshape the query and key matrices so that their dot product produces a transposed attention map of size $\mathbb{R}^{C \times C}$, rather than a traditional attention map of size $\mathbb{R}^{HW \times HW}$. We first reshape the three matrices:

$$\mathbf{Q}_r = \mathrm{reshape}(\mathbf{Q}) \tag{6}$$

$$\mathbf{K}_r = \mathrm{reshape}(\mathbf{K}) \tag{7}$$

$$\mathbf{V}_r = \mathrm{reshape}(\mathbf{V}) \tag{8}$$

We obtain $\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r \in \mathbb{R}^{C \times HW}$, and then compute the attention map:

$$\mathrm{Attention}(\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r) = \mathrm{Softmax}\left(\mathbf{Q}_r \times \mathbf{K}_r^T / \alpha\right) \times \mathbf{V}_r \tag{9}$$

where the parameter α is a learnable scale parameter used to control the magnitude of $\mathbf{Q}_r \times \mathbf{K}_r^T$ before applying the softmax function. Thus, the output of the entire Multi-Dconv Head Transposed Attention (MDTA) module is obtained as follows:

$$\hat{\mathbf{X}} = \mathbf{W}_p \ \mathrm{Attention}\ (\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r) + \mathbf{X} \tag{10}$$

3.2.2. Gated-Dconv Feedforward Network

For the input tensor $\hat{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$, the gated forward propagation layer can be described as:

$$\mathbf{Z} = \mathbf{W}_p^0 \,\text{Gating}\,(\hat{\mathbf{X}}) + \hat{\mathbf{X}} \tag{11}$$

$$\text{Gating}(\hat{\mathbf{X}}) = \phi\left(\mathbf{W}_d^1\mathbf{W}_p^1(\text{LN}(\hat{\mathbf{X}}))\right) \odot \mathbf{W}_d^2\mathbf{W}_p^2(\text{LN}(\hat{\mathbf{X}})) \tag{12}$$

where $\odot$ represents the dot product, ϕ represents the GELU activation function, and LN stands for layer normalization.

3.3. Parameter Sharing Module

The structure of our parameter-sharing module is the same as that of the transformer module. Unlike the transformer attention branches, the parameters of the parameter-sharing module are shared between the infrared and visible light branches. This module can promote the early fusion of features, thereby improving the fusion effect. Specifically, for the ith iteration,

$$\text{base}_{ir} = \mathbf{trans}_{base}[x, u_i] \tag{13}$$

$$\text{base}_{vi} = \mathbf{trans}_{base}[y, u_i] \tag{14}$$

$$f_{in}^i = \left[x * \sigma_x, u_i, y * \sigma_y, \text{ base}_{ir}, \text{ base}_{vi}\right] \tag{15}$$

where x, y, u_i represent the infrared image, visible light image, and the fused image in the ith iteration, respectively; $\mathbf{trans}_{base}$ represents the parameter sharing module, [] represents the concatenation operation, $*$ denotes the dot product, and f_{in}^i indicates the input to the dilated convolution layer in the ith iteration.

3.4. Dilated Convolution Layer

We employ parallel dilated convolution layers to extract and fuse image features effectively. A set of dilated convolution layers with different dilation factors can expand the receptive field without losing neighborhood information. The three convolution layers across three channels all have the same 3×3 size convolution kernels, but they have varying receptive fields due to different dilation factors. As shown in Figure 1, the dilation rates from top to bottom are $1, 2, 3$, respectively, with the receptive fields of the three parallel convolution channels being $3 \times 3, 5 \times 5$, and 7×7 respectively.

To express this more formally, we denote f_{in}^i as the input to the dilated convolution layer in the ith iteration. The output feature map f_{out}^i of the recursive parallel dilated convolution layers can be updated step by step through the following formula:

$$f_{\text{out}}^i = \left\{\mathcal{C}^k\left(f_{\text{in}}^i\right)\right\}_{k\in\{1,2,3\}} \tag{16}$$

$$\mathcal{C}^k\left(f_{in}^i\right) = \theta_{\mathcal{C}}^k * f_{\text{in}}^i + b_{\mathcal{C}}^k \tag{17}$$

where, $\theta_{\mathcal{C}}^k$ and $b_{\mathcal{C}}^k$ respectively represent the weight and bias of the convolution layer with a dilation rate of k.

3.5. Recursive Learning

We adopted a recursive structure to replace traditional multilayer convolution to extract features from the source images from coarse to fine in a manner, as shown in Figure 3. The loop structure receives the fused result from the previous iteration as input for each cycle. Due to the lack of intermediate fusion results for the first iteration, we directly initialize the fused image as either the maximum or average value of the input infrared and visible light images. Compared to the linear stacking method, this approach can save 66% of parameters and mitigate the problem of information loss in deep networks. Due

to the model's minimal number of parameters and high processing speed, it is especially suitable for use on mobile devices.

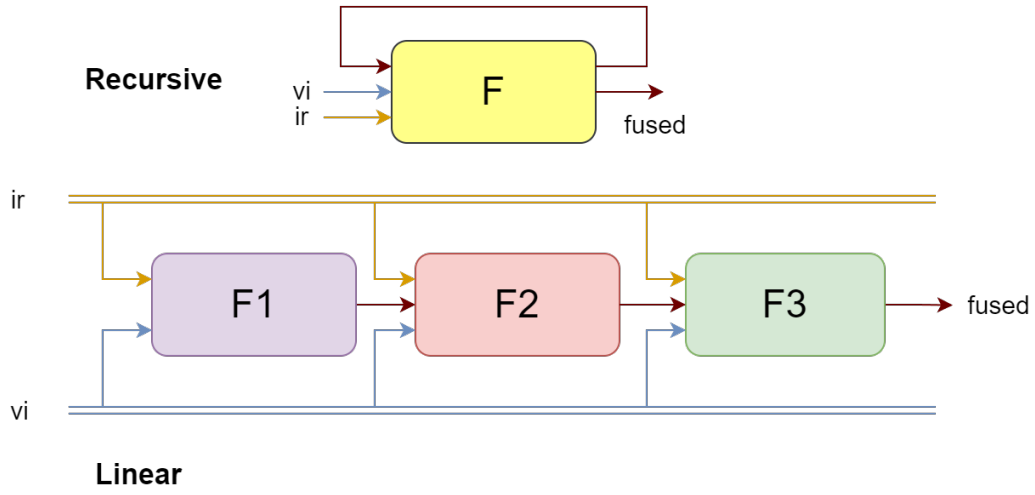

Figure 3. Schematic diagram of cyclic architecture and linear architecture.

3.6. Loss Functions

Our fusion loss function consists of three loss terms. The structural similarity loss $\mathcal{L}_{\text{SSIM}}$ is used to maintain the structure of the source images from the perspectives of illumination, contrast, and structural information. The intensity loss $\mathcal{L}_{\text{int}}$ is used to preserve the pixel brightness values of the source images. The gradient loss $\mathcal{L}_{\text{grad}}$ forces the fused image to keep consistent gradient information with the source images. Thus, our total loss function $\mathcal{L}_{\text{total}}$ can be expressed as:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{SSIM}} + \beta\mathcal{L}_{\text{int}} + (1-\alpha-\beta)\mathcal{L}_{\text{grad}} \tag{18}$$

where, α and β are hyperparameters. Specifically, we hope that the fused image can have the same fundamental structure as the source images; hence, the structural similarity loss $\mathcal{L}_{\text{SSIM}}$ is defined as:

$$\mathcal{L}_{\text{SSIM}} = (1-\text{SSIM}(u,x)) + (1-\text{SSIM}(u,y)) \tag{19}$$

Similarly, the distribution of pixel intensity values in the fused image should achieve some balance between the infrared and visible light images; hence, the intensity loss $\mathcal{L}_{\text{int}}$ can be defined as follows:

$$\mathcal{L}_{\text{int}} = \|u-x\|_1 + \|u-y\|_1 \tag{20}$$

where $\|\cdot\|_1$ represents the l_1 norm. For gradient information, highlighted targets tend to have richer gradient information in the infrared or visible light images. For example, at night, infrared images highlight pedestrians, vehicles, and other targets, while other parts of the scene generally have lower intensity values; visible light images show more gradient information in well-lit areas. Therefore, if the fused image wants to retain the information from both source images, the gradient should equal the maximum gradient of the two source images. Hence, our gradient loss $\mathcal{L}_{\text{grad}}$ is designed as follows:

$$\mathcal{L}_{\text{grad}} = \left\|\text{grad}_u - \max\left(\text{grad}_x, \text{grad}_y\right)\right\|_1 \tag{21}$$

$$\text{grad}(I) = \|\,\text{sobelx}(I)\|_1 + \|\,\text{sobely}(I)\|_1 \tag{22}$$

where sobelx and sobely are Sobel operators used to extract the gradient information of the image in two directions.

4. Experiments and Results

4.1. Dataset and Preprocessing

- Dataset: To compare model performance, we selected three public datasets for our experiments: MSRS [23], TNO [70], and RoadScene [71]. Our training set is the MSRS (Multispectral Road Scene) dataset, which includes 1083 pairs of infrared and visible images. The test set consists of the MSRS dataset (20 pairs), the TNO dataset (20 pairs), and the RoadScene dataset (40 pairs). The TNO dataset contains grayscale images, while the MSRS and RoadScene datasets contain color images.
- Evaluation metrics: In this paper, we selected seven metrics as our evaluation criteria: information entropy (EN) [72], standard deviation (SD) [73], spatial frequency (SF), mutual information (MI), sum of the correlations of differences (SCD) [74], visual information fidelity (VIF) [75], and gradient-based fusion performance (Qabf) [76]. EN is a reference-free metric that measures the richness of information in the fused image from an information theory perspective. The higher the EN index, the richer the information contained in the fused image, which usually means the better the fusion effect. SD is also an independent indicator, usually used to evaluate the grayscale distribution and contrast of the fused image. As the variability in pixel intensities increases (which is reflected by an increase in SD), the image's contrast also increases, making the visual content more distinguishable and engaging. SF denotes image detail clarity and spatial variation. The larger the SF, the richer the texture and edges. SF is an evaluation index based on image gradients. It evaluates the edge information in the fused image by comprehensively calculating the gradient in the horizontal spatial direction and the gradient in the vertical spatial direction of the fused image. MI determines the degree to which the fused image retains the source image information by calculating the sum of the dependencies between the fused image and the infrared image and visible image, respectively. The larger the MI value, the more information the fused image retains of the source image. SCD is a relatively nuanced metric designed to evaluate the quality of fused images by measuring the correlation of sum differences between them. Essentially, SCD assesses the spatial relationship and variance in intensity between pixels in different images, providing a measure of how changes in one image correspond to changes in another. This metric is particularly useful when comparing the information content and spatial detail of fused images to their source images. VIF is a metric that quantifies the quality of an image by assessing how accurately it preserves visual information from a reference image, based on the human visual system's perception. Qabf is an evaluation index that calculates the saliency information from the source image retained in the fused image in each window from the perspectives of image contrast, clarity, and information entropy. The value of this indicator ranges from 0 to 1. The higher the values of the aforementioned metrics, the better the quality of the fused image. The comprehensive application of these metrics can thoroughly assess the quality of the fused images.
- RGB image processing: Our model can directly take color images as input for training and prediction without requiring manual code adjustments. Inspired by [28,34,77], we read color images in RGB format, then convert the images to the YCrCb space and extract the Y channel for fusion. After obtaining the fusion result, we use it as the new Y channel, combine it with the original Cr and Cb channels, and obtain the final color fused image. Our model can adaptively handle both color and grayscale images without the need for additional preprocessing operations.
- Training details: Parameters are updated using the Adam optimizer, with a learning rate of 0.001 and training for 30 epochs. The values of α and β are set to 0.5 and 0.1, respectively, and the model iterates three times. The training dataset is the MSRS dataset. Unless otherwise specified, all experiments are conducted on a computer with a single Nvidia RTX3090 GPU.

4.2. Qualitative Comparison

From Figure 4a, it can be seen that methods such as DenseFuse [78], PMGI [31], DIDFuse [79], SDNet [80], and U2Fusion [81] exhibit apparent ghosting phenomena around the edges of trees, whereas our method can avoid these flaws more effectively. The overall brightness of images produced by SDNet and U2Fusion is too high, while on the contrary, the overall brightness of images from PMGI, DIDFuse, and GANMcC [38] is too low (roads are not visible in Figure 4a). The LRRNet [27] method tends to reduce the contrast of targets, which is disadvantageous for target detection and tracking. Compared to ReCoNet, our method makes the pedestrians in Figure 4a,c and the stairs in (e) more clearly visible.

Figure 4. *Cont.*

Figure 4. Comparison of visual effects with 9 models on the TNO dataset. (**a**) Kaptein 1123; (**b**) Marne 04; (**c**) two men in front of house; (**d**) soldier behind smoke; (**e**) Marne 07.

From Figure 5, it can be seen that the overall brightness of images from PMGI and SDNet is too high, while on the contrary, the overall brightness of images from DIDFuse and ReCoNet is too low. In MFEIF and GANMcC, the edges of pedestrian targets are blurred; in DIDFuse, U2Fusion, ReCoNet, and LRRNet, the contrast between infrared salient targets and the environment is lower. Meanwhile, our method can produce highlighted targets with precise contours and results with moderate brightness.

Figure 5. *Cont.*

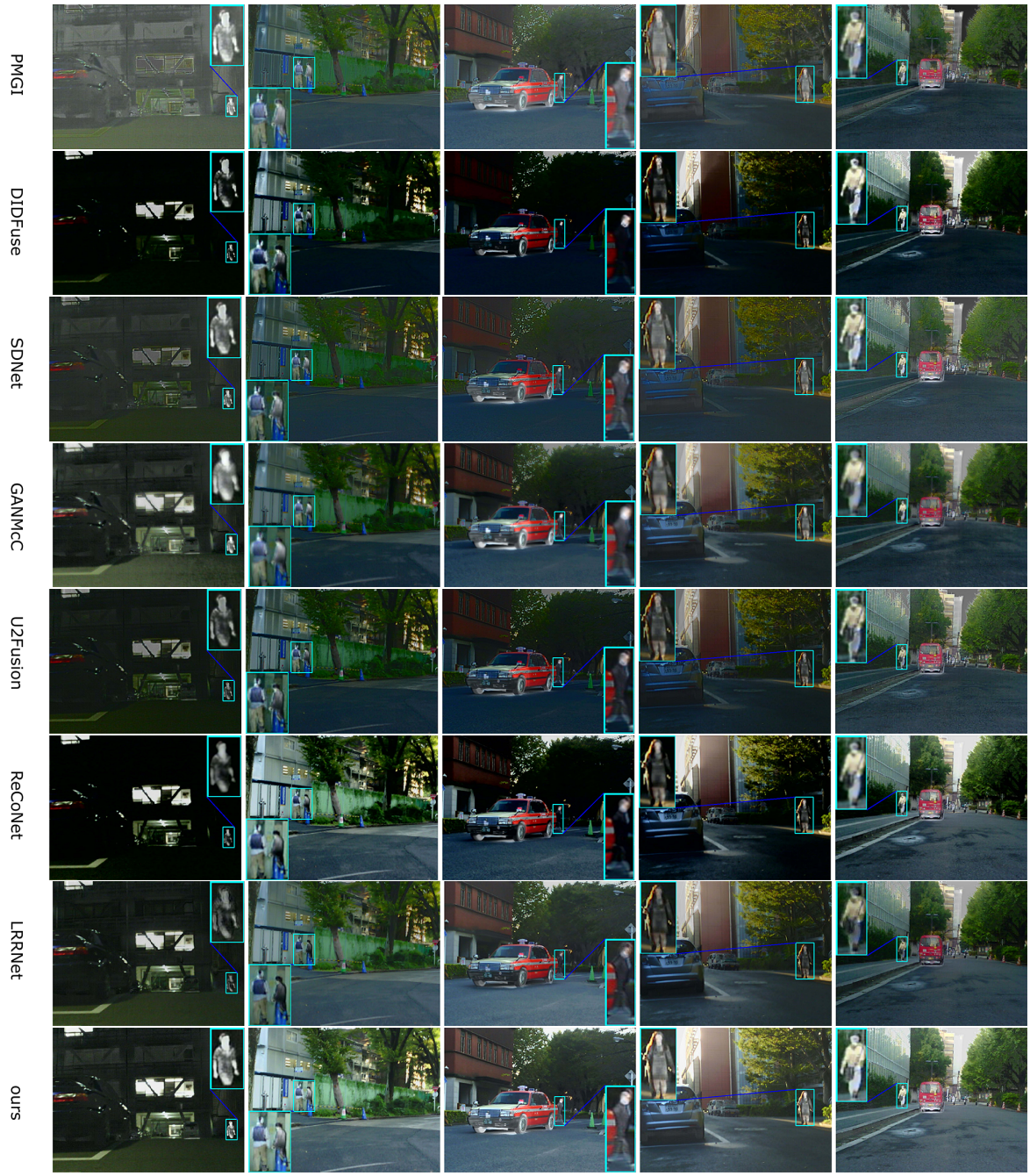

Figure 5. Comparison of visual effects with 9 models on the MSRS dataset. (**a**) 00051N; (**b**) 00085D; (**c**) 00131D; (**d**) 00169D; (**e**) 00357D.

From Figure 6, it can be seen that the overall brightness of images from MFEIF [14], SDNet, U2Fusion, and ReCoNet is too high, while PMGI and GANMcC have an overall brightness that is too low and the targets are relatively blurry. In the images from DIDFuse and SDNet, the edges of the leaves are filled with a large amount of white ghosting, and there is considerable noise. Our method can produce fused images with moderate brightness which prominently display infrared targets and are relatively clear.

Figure 6. Comparison of visual effects with 9 models on the RoadScene dataset. (**a**) FLIR 00306; (**b**) FLIR 00497; (**c**) FLIR 01463; (**d**) FLIR 04269; (**e**) FLIR 04302.

4.3. Quantitative Comparison

We selected seven evaluation metrics to comprehensively assess our method against nine other advanced fusion methods across three public datasets, focusing on the richness of information in the fused images (EN, MI, SF, Qabf), image contrast (SD), structural similarity to the original images (SCD), and visual effects (VIF), as shown in Tables 2–4.

Table 2. Comparison with 9 SOTA fusion methods on the TNO dataset; 20 pairs of images are selected for testing, and the results are averaged and rounded to two decimal places. ↑ indicates that the larger the indicator value, the better the fusion effect. **Bold** and *underline*, respectively, indicate the best and second-best results.

	EN ↑	SD ↑	SF ↑	MI ↑	SCD ↑	VIF ↑	Qabf ↑
DenseFuse	6.86	35.59	9.33	1.54	1.81	0.62	0.42
MFEIF	6.72	33.16	7.54	1.73	1.74	0.64	0.45
PMGI	6.85	32.97	9.00	1.43	1.72	0.57	0.37
DIDFuse	7.09	**47.47**	11.78	1.62	**1.83**	0.60	0.41
GANMcC	6.72	32.85	6.44	1.54	1.71	0.50	0.26
U2Fusion	6.15	23.95	8.25	1.14	1.51	0.50	0.39
SDNet	6.29	23.98	9.86	1.28	1.46	0.48	0.41
ReCoNet	6.89	40.24	8.07	1.66	1.70	0.54	0.38
LRRNet	7.07	42.13	10.18	1.72	1.63	0.57	0.37
ours	**7.13**	44.24	**12.07**	**2.00**	1.70	**0.71**	**0.47**

Table 3. Comparison with 9 SOTA fusion methods on the MSRS dataset; 20 pairs of images are selected for testing, and the results are averaged and rounded to two decimal places. ↑ indicates that the larger the indicator value, the better the fusion effect. **Bold** and *underline*, respectively, indicate the best and second-best results.

	EN ↑	SD ↑	SF ↑	MI ↑	SCD ↑	VIF ↑	Qabf ↑
DenseFuse	6.46	40.48	10.29	**2.68**	1.50	**0.87**	**0.63**
MFEIF	5.92	35.34	8.67	2.04	1.58	0.72	0.54
PMGI	5.85	19.62	8.73	1.33	0.84	0.59	0.36
DIDFuse	4.54	33.68	10.95	1.40	1.15	0.30	0.22
GANMcC	6.16	28.75	6.06	1.73	1.42	0.64	0.32
U2Fusion	4.80	21.30	7.60	1.28	1.04	0.45	0.34
SDNet	5.02	16.96	8.32	1.15	0.89	0.44	0.32
ReCoNet	5.27	**43.20**	10.72	1.77	1.41	0.51	0.42
LRRNet	6.19	32.14	8.83	2.11	0.89	0.55	0.46
ours	**6.50**	41.93	**11.73**	2.37	**1.74**	0.86	0.61

Table 4. Comparison with 9 SOTA fusion methods on the RoadScene dataset; 40 pairs of images are selected for testing, and the results are averaged and rounded to two decimal places. ↑ indicates that the larger the indicator value, the better the fusion effect. **Bold** and *underline*, respectively, indicate the best and second-best results.

	EN ↑	SD ↑	SF ↑	MI ↑	SCD ↑	VIF ↑	Qabf ↑
DenseFuse	7.26	48.57	12.68	2.40	1.51	0.62	**0.48**
MFEIF	6.84	35.74	8.84	2.30	1.66	0.61	0.44
PMGI	7.20	45.04	10.54	2.24	1.68	0.56	0.40
DIDFuse	7.18	46.57	14.01	2.07	**1.76**	0.58	0.44
GANMcC	**7.30**	47.96	9.41	1.97	1.74	0.52	0.34
U2Fusion	6.58	30.82	11.40	1.91	1.42	0.52	0.45
SDNet	7.04	39.10	12.82	2.38	1.40	0.55	0.46
ReCoNet	6.82	37.90	8.61	2.33	1.56	0.54	0.37
LRRNet	7.11	43.98	12.85	2.16	1.65	0.53	0.38
ours	7.25	**48.85**	**14.75**	**2.50**	1.62	**0.67**	0.44

From Table 2, we can observe that MFEIF performs well on metrics such as MI, VIF, and Qabf, indicating that its fused images are information-rich and visually appealing, as shown in Figure 4. DIDFuse performs well on EN, SD, SF, and SCD, suggesting its fused images are information-rich with higher contrast. Our method outperforms others on metrics like EN, SF, MI, VIF, and Qabf. It is second-best on the SD metric, indicating that the fused images produced by our method contain abundant information, have good visual effects, and possess high contrast, enabling the prominent display of infrared targets.

From Table 3, it can be seen that MFEIF and DIDFuse, which perform well on the TNO dataset, cannot effectively process images from the MSRS dataset. In contrast, DenseFuse is the best in terms of MI, VIF, and Qabf metrics, and second-best on the EN metric, indicating its superior ability to fuse information from color images. ReCoNet performs best on the SD metric, meaning its fused images have higher contrast, while our method excels in all seven metrics.

From Table 4, it is observed that DenseFuse performs well in metrics such as EN, SD, MI, VIF, and Qabf, indicating that its fused images contain rich information and have good visual effects. DIDFuse performs well on SF and SCD, suggesting its images contain rich texture details. However, as seen in Figure 6, its images are filled with a significant amount of noise at the edges of leaves and contain many white false edges in the sky in (d), leading to a potential misjudgment by the SF and SCD metrics regarding image quality. A similar issue occurs with GANMcC, where the EN metric might mistakenly consider noise as valid information. Our method performs best in SD, SF, MI, and VIF metrics, demonstrating that, compared to other methods, it produces images with less noise, more explicit images, moderate brightness, and good visual effects.

4.4. The Effectiveness of Recursive Training

We modified the architecture to a linear stacking approach to verify the recursive training methods' effectiveness and compared it with our model. The training dataset is the MSRS dataset, and the test dataset consists of 20 pairs of images from the MSRS test set. See Tables 5–7.

Table 5. Comparison of two architectures when training on the MSRS dataset, with a batch size of 16, training for 30 epochs.

	GPU Memory (MB)	Model Parameters	Average Training Time per Batch (s)
Linear	61.66	17,655	47.0
Recursive	61.27	5885	45.8

Table 6. Comparison of two architectures when testing on the MSRS dataset.

	GPU Memory (MB)	Average Time for Fusing Each Pair of Images (s)
Linear	17.635	0.050
Recursive	17.545	0.051

Table 7. Comparison of the fusion effects of two architectures on the MSRS dataset.

	EN	SD	SF	MI	SCD	VIF	Qabf
Linear	6.34	38.89	9.67	2.16	1.59	0.75	0.55
Recursive	6.41	39.75	11.47	2.29	1.67	0.81	0.60

In summary, the recursive training approach can significantly reduce the number of model parameters and is beneficial in preventing the loss of information due to overly deep models. However, there is no substantial improvement in GPU memory usage and runtime. This conclusion greatly differs from that of ReCoNet [28], which claims that a

cyclical architecture reduces the time consumption by about 15%, the number of parameters by 33%, and GPU memory usage by 42%. For models with a depth of 3, the parameter count should be reduced by about 66%, as we have calculated.

4.5. Discussion of the Iteration in Attention Module

Figure 7 shows the impact of the attention module's iterations on the fusion results. The attention map will be dot-multiplied with the image of the corresponding modality. After the concatenation operation, it will be used as the input of the dilated convolution module. It can be observed that with the progression of iterations, the infrared channel pays more attention to vehicles and pedestrians, while the visible light channel focuses more on bushes, the big tree on the right and in the distance. The information focused on by the two channels complements each other, extracting different features and laying the foundation for subsequent feature fusion to obtain a more information-rich fused image.

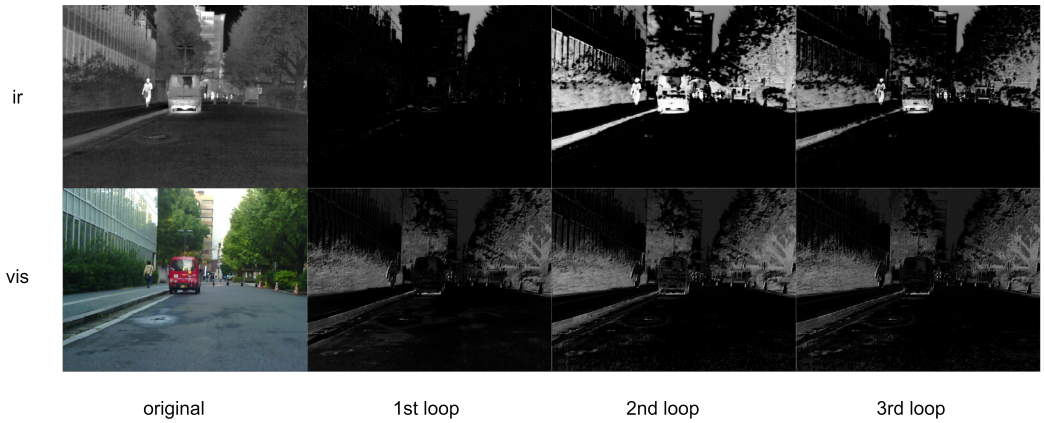

Figure 7. Visual result of our recurrent learning mechanism.

4.6. Ablation Studies

To demonstrate the effectiveness of the proposed module, we conducted ablation experiments on the MSRS dataset, with the results shown in Table 8. 'ori' refers to the model that lacks both the transformer attention branch and the shared feature fusion module, with the ReCoNet model serving as a reference; '+base' denotes the model obtained by removing the transformer attention branch; '+trans' represents the model with the shared feature fusion module removed. All models were trained on the MSRS dataset for 30 batches and underwent partial parameter tuning to demonstrate the model's capabilities fully.

Table 8. Ablation experiment results on the MSRS dataset. Bold indicates the best result.

	EN	SD	SF	MI	SCD	VIF	Qabf
ori	6.36	39.87	10.12	2.00	1.71	0.77	0.57
+base	6.45	42.01	11.51	2.34	1.72	0.86	0.62
+trans	6.49	42.73	**11.96**	2.40	**1.76**	0.87	0.62
ours	**6.61**	**43.52**	11.80	**2.99**	1.66	**0.92**	**0.63**

Adding either the transformer attention branch or the shared feature fusion module alone can significantly enhance the performance of the original model, with improvements across all seven metrics. When both modules are incorporated, there is a slight decrease in the model's performance on SF and SCD. However, the fused images contain richer information (EN, MI increase), have greater contrast (SD increase), and offer improved visual effects (VIF increase).

4.7. Fusion Time Test

To better demonstrate the actual running speed of our model, we compared the fusion speed of 11 state-of-the-art fusion models using 20 pairs of images from the TNO dataset, with the results shown in Figure 8.

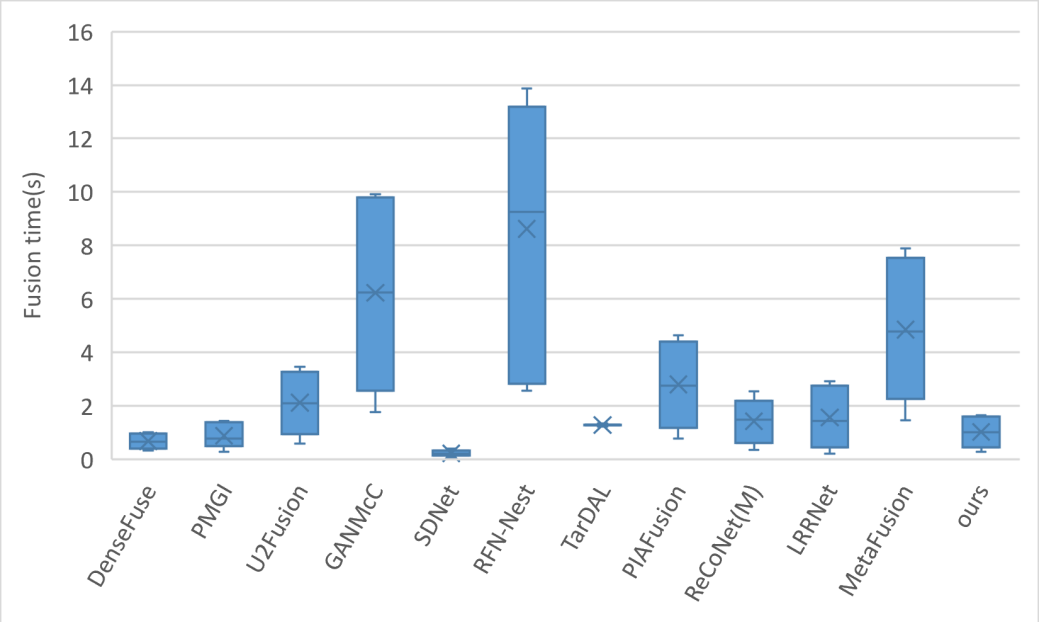

Figure 8. Comparison of fusion time for 20 pairs of images on the TNO dataset. The fusion time represents the average time taken to fuse each pair of images.

In the reproduction of U2Fusion, code was added to convert image precision from float32 to uint8, which slightly increased the fusion time for U2Fusion compared to the original. In the ReCoNet model, the m-Register registration module was activated instead of using the pre-trained model provided by the authors. For the replication of TarDAL, the tardal-dt pre-trained model given by the authors was used, and a timing code was added to its fusion script. The timing started at the beginning of the fusion process and ended before saving the image (excluding image saving time), with the timing setup for other models being similar. DenseFuse, by default, crops images to a resolution of 256 × 256, reducing runtime; we turn off image cropping in this study. MetaFusion defaults to cropping images to a resolution of 512 × 384; we modify it to keep the input and output resolution consistent for comparison with other algorithms. Considering that some models' code frameworks are outdated and could not utilize GPU acceleration on our machine, all our models were run on the same computer with an Intel Core i7-9750H CPU @ 2.60 GHz processor, and GPU acceleration was turned off to ensure a consistent hardware platform for model execution.

It can be observed that our model operates quickly, processing a pair of images on average every second, but is slightly slower than methods like DenseFuse, SDNet, and PMGI in terms of speed. An interesting point to note is that the aforementioned three models are based on the TensorFlow code framework, whereas our model is built on PyTorch. Therefore, the slower performance of our model might be related to the internal implementations of the deep learning framework, although this requires further research. Although our model has the smallest number of parameters, the computational method of the transformer and the model's recursive calling may also limit further speed improvements.

5. Conclusions

This paper proposes a lightweight and training-efficient infrared and visible image fusion model. The model only requires about 15 min of training on an NVIDIA GeForce RTX 3090 GPU, without the need for the demanding hardware specifications and potentially extensive tuning time costs associated with other larger models. Its end-to-end training

approach is straightforward and effective, avoiding the complexity of two-stage training processes like CDDFuse and the instability issues associated with GAN-based models. Current researchers are keen on designing broader and deeper models, for instance, by incorporating self-attention mechanisms of transformer and dense connections to improve fusion effects. However, the performance improvement might be unacceptable compared to the added expense. As demonstrated in this paper, a few simple max pooling and average pooling layers can achieve satisfactory fusion results. Despite its minimal parameter count, the model's processing speed has yet to reach its optimum. We consider further exploring the relationship between model parameters and processing speed in the future.

Author Contributions: Conceptualization, W.X.; methodology, K.Y.; software, K.Y.; validation, K.Y.; writing—original draft preparation, K.Y.; writing—review and editing, W.X., Z.C. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Infrared vision theory and method] grant number [2023-JCJQ-ZD-011-12], and the APC was funded by [Infrared vision theory and method].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The TNO dataset can be found at https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029 (accessed on 27 November 2023), the MSRS dataset at https://github.com/Linfeng-Tang/MSRS (accessed on 26 November 2023) and the RoadScene dataset at https://github.com/hanna-xu/RoadScene (accessed on 27 November 2023).

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

References

1. Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; Ma, J. DIVFusion: Darkness-free infrared and visible image fusion. *Inf. Fusion* **2023**, *91*, 477–493. [CrossRef]
2. Ma, W.; Wang, K.; Li, J.; Yang, S.X.; Li, J.; Song, L.; Li, Q. Infrared and visible image fusion technology and application: A review. *Sensors* **2023**, *23*, 599. [CrossRef] [PubMed]
3. Zhang, X.; Demiris, Y. Visible and infrared image fusion using deep learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10535–10554. [CrossRef] [PubMed]
4. Singh, S.; Singh, H.; Bueno, G.; Deniz, O.; Singh, S.; Monga, H.; Hrisheekesha, P.; Pedraza, A. A review of image fusion: Methods, applications and performance metrics. *Digit. Signal Process.* **2023**, *137*, 104020. [CrossRef]
5. Sayyadi Tooranloo, H.; Saghafi, S. Assessing the risk of hospital information system implementation using IVIF FMEA approach. *Int. J. Healthc. Manag.* **2021**, *14*, 676–689. [CrossRef]
6. Pandit, V.R.; Bhiwani, R. Image fusion in remote sensing applications: A review. *Int. J. Comput. Appl.* **2015**, *120*, 22–32.
7. Gu, Y.; Wang, X.; Zhang, C.; Li, B. Advanced driving assistance based on the fusion of infrared and visible images. *Entropy* **2021**, *23*, 239. [CrossRef] [PubMed]
8. Wang, C.; Yang, G.; Sun, D.; Zuo, J.; Wang, E.; Wang, L. Frequency domain fusion algorithm of infrared and visible image based on compressed sensing for video surveillance forensics. In Proceedings of the 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Shenyang, China, 20–22 October 2021; pp. 832–839.
9. Ciprián-Sánchez, J.F.; Ochoa-Ruiz, G.; Gonzalez-Mendoza, M.; Rossi, L. FIRe-GAN: A novel deep learning-based infrared-visible fusion method for wildfire imagery. *Neural Comput. Appl.* **2023**, *35*, 18201–18213. [CrossRef]
10. Luo, Y.; Luo, Z. Infrared and visible image fusion: Methods, datasets, applications, and prospects. *Appl. Sci.* **2023**, *13*, 10891. [CrossRef]
11. Liu, Y.; Wu, Z.; Han, X.; Sun, Q.; Zhao, J.; Liu, J. Infrared and visible image fusion based on visual saliency map and image contrast enhancement. *Sensors* **2022**, *22*, 6390. [CrossRef] [PubMed]
12. Lu, S.; Ding, Y.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. Multiscale feature extraction and fusion of image and text in VQA. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 54. [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. Liu, J.; Fan, X.; Jiang, J.; Liu, R.; Luo, Z. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 105–119. [CrossRef]
15. Li, L.; Lv, M.; Jia, Z.; Ma, H. Sparse representation-based multi-focus image fusion method via local energy in shearlet domain. *Sensors* **2023**, *23*, 2888. [CrossRef] [PubMed]

16. Yang, Y.; Zhang, Y.; Huang, S.; Zuo, Y.; Sun, J. Infrared and visible image fusion using visual saliency sparse representation and detail injection model. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5001715. [CrossRef]
17. Li, X.; Tan, H.; Zhou, F.; Wang, G.; Li, X. Infrared and visible image fusion based on domain transform filtering and sparse representation. *Infrared Phys. Technol.* **2023**, *131*, 104701. [CrossRef]
18. Qi, B.; Bai, X.; Wu, W.; Zhang, Y.; Lv, H.; Li, G. A novel saliency-based decomposition strategy for infrared and visible image fusion. *Remote Sens.* **2023**, *15*, 2624. [CrossRef]
19. Li, H.; Wu, X.J. Infrared and visible image fusion using latent low-rank representation. *arXiv* **2018**, arXiv:1804.08992.
20. Liu, J.; Dian, R.; Li, S.; Liu, H. SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. *Inf. Fusion* **2023**, *91*, 205–214. [CrossRef]
21. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
22. Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; Fan, X. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 8115–8124.
23. 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]
24. 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.
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. Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; Van Gool, L. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5906–5916.
27. Li, H.; Xu, T.; Wu, X.J.; Lu, J.; Kittler, J. LRRNet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11040–11052. [CrossRef] [PubMed]
28. Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; Luo, Z. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 539–555.
29. 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.
30. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
31. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; Ma, J. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12797–12804. [CrossRef]
32. 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]
33. Long, Y.; Jia, H.; Zhong, Y.; Jiang, Y.; Jia, Y. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Inf. Fusion* **2021**, *69*, 128–141. [CrossRef]
34. Cheng, C.; Xu, T.; Wu, X.J. MUFusion: A general unsupervised image fusion network based on memory unit. *Inf. Fusion* **2023**, *92*, 80–92. [CrossRef]
35. Liu, J.; Wu, Y.; Wu, G.; Liu, R.; Fan, X. Learn to search a lightweight architecture for target-aware infrared and visible image fusion. *IEEE Signal Process. Lett.* **2022**, *29*, 1614–1618. [CrossRef]
36. Li, H.; Cen, Y.; Liu, Y.; Chen, X.; Yu, Z. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans. Image Process.* **2021**, *30*, 4070–4083. [CrossRef] [PubMed]
37. 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]
38. Ma, J.; Zhang, H.; Shao, Z.; Liang, P.; Xu, H. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5005014. [CrossRef]
39. Zhou, H.; Hou, J.; Zhang, Y.; Ma, J.; Ling, H. Unified gradient-and intensity-discriminator generative adversarial network for image fusion. *Inf. Fusion* **2022**, *88*, 184–201. [CrossRef]
40. Fu, Y.; Wu, X.J.; Durrani, T. Image fusion based on generative adversarial network consistent with perception. *Inf. Fusion* **2021**, *72*, 110–125. [CrossRef]
41. Wang, Z.; Shao, W.; Chen, Y.; Xu, J.; Zhang, L. A cross-scale iterative attentional adversarial fusion network for infrared and visible images. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3677–3688. [CrossRef]
42. Li, Q.; Han, G.; Liu, P.; Yang, H.; Chen, D.; Sun, X.; Wu, J.; Liu, D. A multilevel hybrid transmission network for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5016114. [CrossRef]

43. Xu, H.; Gong, M.; Tian, X.; Huang, J.; Ma, J. CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. *Comput. Vis. Image Underst.* **2022**, *218*, 103407. [CrossRef]
44. Wang, X.; Hua, Z.; Li, J. PACCDU: Pyramid attention cross-convolutional dual UNet for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5024416. [CrossRef]
45. Wang, J.; Xi, X.; Li, D.; Li, F. FusionGRAM: An infrared and visible image fusion framework based on gradient residual and attention mechanism. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5005412. [CrossRef]
46. Zhao, H.; Nie, R. Dndt: Infrared and visible image fusion via densenet and dual-transformer. In Proceedings of the 2021 International Conference on Information Technology and Biomedical Engineering (ICITBE), Nanchang, China, 24–26 December 2021; pp. 71–75.
47. 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]
48. Li, J.; Zhu, J.; Li, C.; Chen, X.; Yang, B. CGTF: Convolution-guided transformer for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5012314. [CrossRef]
49. Tang, W.; He, F.; Liu, Y. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Trans. Multimed.* **2022**, *25*, 5413–5428. [CrossRef]
50. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
51. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the 29th Conference on Advances in Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
52. Arjovsky, M.; Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv* **2017**, arXiv:1701.04862.
53. Xu, H.; Ma, J.; Yuan, J.; Le, Z.; Liu, W. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19679–19688.
54. Tang, L.; Yuan, J.; Ma, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
55. Wang, D.; Liu, J.; Liu, R.; Fan, X. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Inf. Fusion* **2023**, *98*, 101828. [CrossRef]
56. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
57. 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 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
58. Vs, V.; Valanarasu, J.M.J.; Oza, P.; Patel, V.M. Image fusion transformer. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3566–3570.
59. Fu, Y.; Xu, T.; Wu, X.; Kittler, J. Ppt fusion: Pyramid patch transformerfor a case study in image fusion. *arXiv* **2021**, arXiv:2107.13967.
60. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
61. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
62. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
63. Lin, L.; Fan, H.; Zhang, Z.; Xu, Y.; Ling, H. Swintrack: A simple and strong baseline for transformer tracking. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16743–16754.
64. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
65. Hu, R.; Singh, A. Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1439–1449.
66. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
67. Gao, Y.; Zhang, M.; Wang, J.; Li, W. Cross-scale mixing attention for multisource remote sensing data fusion and classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5507815. [CrossRef]
68. Zhang, X.; Li, Q.; Li, W.; Guo, Y.; Zhang, J.; Guo, C.; Chang, K.; Lovell, N.H. FD-Net: Feature distillation network for oral squamous cell carcinoma lymph node segmentation in hyperspectral imagery. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1552–1563. [CrossRef] [PubMed]
69. Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; Han, S. Lite transformer with long-short range attention. *arXiv* **2020**, arXiv:2004.11886.

70. Toet, A. TNO Image Fusion Dataset. 2014. Available online: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029/2 (accessed on 27 November 2023).
71. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. FusionDN: A unified densely connected network for image fusion. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
72. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522.
73. Eskicioglu, A.M.; Fisher, P.S. Image quality measures and their performance. *IEEE Trans. Commun.* **1995**, *43*, 2959–2965. [CrossRef]
74. Aslantas, V.; Bendes, E. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-Int. J. Electron. Commun.* **2015**, *69*, 1890–1896. [CrossRef]
75. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [CrossRef]
76. Piella, G.; Heijmans, H. A new quality metric for image fusion. In Proceedings of the Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), Barcelona, Spain, 14–17 September 2003; Volume 3, pp. III–173.
77. Deng, X.; Dragotti, P.L. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3333–3348. [CrossRef] [PubMed]
78. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef] [PubMed]
79. Zhao, Z.; Xu, S.; Zhang, C.; Liu, J.; Zhang, J.; Li, P. DIDFuse: Deep image decomposition for infrared and visible image fusion. In Proceedings of the IJCAI, Yokohama, Japan, 11–17 July 2020; pp. 970–976.
80. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [CrossRef]
81. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef] [PubMed]