remote sensing

MDPI

Article

Self-Attention Progressive Network for Infrared and Visible Image Fusion

Shuying Li [1,2], Muyi Han [1], Yuemei Qin [1] and Qiang Li [3,*]

[1] School of Automation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; lishuying@xupt.edu.cn (S.L.); hanmuyi@stu.xupt.edu.cn (M.H.); qinyuemei@xupt.edu.cn (Y.Q.)
[2] Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[3] School of Artificial Intelligence, Northwestern Polytechnical University, Xi'an 710021, China
* Correspondence: qiangli@nwpu.edu.cn

Abstract: Visible and infrared image fusion is a strategy that effectively extracts and fuses information from different sources. However, most existing methods largely neglect the issue of lighting imbalance, which makes the same fusion models inapplicable to different scenes. Several methods obtain low-level features from visible and infrared images at an early stage of input or shallow feature extraction. However, these methods do not explore how low-level features provide a foundation for recognizing and utilizing the complementarity and common information between the two types of images. As a result, the complementarity and common information between the images is not fully analyzed and discussed. To address these issues, we propose a Self-Attention Progressive Network for the fusion of infrared and visible images in this paper. Firstly, we construct a Lighting-Aware Sub-Network to analyze lighting distribution, and introduce intensity loss to measure the probability of scene illumination. This approach enhances the model's adaptability to lighting conditions. Secondly, we introduce self-attention learning to design a multi-state joint feature extraction module (MSJFEM) that fully utilizes the contextual information among input keys. It guides the learning of a dynamic attention matrix to strengthen the capacity for visual representation. Finally, we design a Difference-Aware Propagation Module (DAPM) to extract and integrate edge details from the source images while supplementing differential information. The experiments across three benchmark datasets reveal that the proposed approach exhibits satisfactory performance compared to existing methods.

Keywords: deep learning; image fusion; infrared image; light-aware; self-attention learning

check for updates

Citation: Li, S.; Han, M.; Qin, Y.; Li, Q. Self-Attention Progressive Network for Infrared and Visible Image Fusion. *Remote Sens.* **2024**, *16*, 3370. https://doi.org/10.3390/rs16183370

Academic Editor: Salah Bourennane

Received: 30 July 2024
Revised: 3 September 2024
Accepted: 9 September 2024
Published: 11 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Due to the theoretical and technical limitations of hardware equipment, information obtained from a single modal sensor or a single shooting setting is often insufficient for effectively and comprehensively capturing the nuances of the imaging scene [1]. For instance, reflected light captures only the surface attributes of an object. Therefore, reflected light cannot reveal insights into the object interior or what lies beyond its immediate surface, offering only a partial view of the object's external characteristics. Given this constraint, it is necessary to develop related techniques to solve this condition. To achieve this, image fusion is proposed. Its aim is to combine data from multiple sources to create a composite image, which improves the accuracy of image analysis and interpretation by integrating the advantages of different sensors. In particular, the fusion of infrared and visible images provides complementary information and assistance for representation learning, thus enhancing scene depiction and visual perception. By leveraging this unique characteristic, the integration of image fusion technology significantly enhances the performance of diverse natural image tasks, including surveillance [2], target recognition [3], semantic delineation [4], etc.

Over the past few decades, significant advancements have been made in the development of traditional methods, with a particular focus on fusion strategies and theories

for integrating infrared and visible images. These methods can be largely categorized into five types: multi-scale transformation techniques [5,6], sparse representation strategies [7], subspace clustering algorithms [8], saliency-based methodologies [9], and optimization approaches [10]. Each category employs mathematical transformations to convert the source images into a specific domain, where they manually design activity levels and establish fusion rules to achieve image fusion. Although these methods achieve success in certain fusion tasks, the transformations used by the above methods are becoming increasingly complex. Frequently, the same transformations are applied to extract features from source images. However, this approach neglects the analysis of the unique characteristics inherent in infrared and visible images. Additionally, the limited selection of fusion rules and the handcrafted features often fail to address complex scenarios [11].

Currently, the development of deep learning is driving the creation of numerous deep fusion methods. These methods are primarily categorized into three types based on different network architectures. They are based on autoencoders (AEs) [12,13], convolutional neural networks (CNNs) [14,15], and generative adversarial networks (GANs) [16–18]. Deep learning technology usually outperforms traditional fusion methods in terms of fusion performance. However, deep learning also faces new challenges. One critical issue is the difficulty in selectively utilizing pertinent information to guide the training of the fusion framework, owing to the lack of specific fusion labels. While some end-to-end models seek to address this by weighting the pixel values of the source and fused images, they encounter difficulties by treating all regions equally in constructing the loss function [19]. Furthermore, the field of image fusion largely overlooks the issue of unbalanced lighting. For instance, daytime visible images exhibit high resolution and rich texture, whereas nighttime visible images suffer from a loss of texture detail. When the identical fusion model is applied across two scenarios, it may yield suboptimal results, thereby impacting the performance of the fusion.

To address the previously mentioned challenges, we introduce a light-aware subnetwork that utilizes light intensity loss as a guiding mechanism for enhancing the fusion network learning capabilities. In the fusion network, our proposed transformer technique integrates both contextual information mining and self-attention learning into a unified architecture. Subsequently, we propose a detail enhancement module that can utilize both infrared and visible images information, effectively enhancing edge depiction by extracting common and complementary features. In summary, this paper presents three significant contributions:

- We develop a progressive image fusion framework that explores the information contained within two modalities. This framework can autonomously learn advanced image features based on lighting intensity, and adaptively fuse complementary and common information, thereby effectively fusing meaningful information in the source image in an all-weather environment.
- A multi-state joint feature extraction module is proposed to enable context mining between keywords and self-attentive learning of feature maps in an intra-clustered manner. The module integrates these elements into a unified architecture that merges static and dynamic contextual representations into a consistent output.
- We introduce a differential module with learnable parameters for the activation function. It can dynamically adapt to the distribution of input data, effectively extracting and integrating key information from the source images while supplementing differential information.
- The rest of this paper is organized as follows: Section 2 reviews the related works based on techniques for infrared and visible images. Section 3 describes the proposed method in detail, including multi-state joint feature extraction, Difference-Aware Propagation Module and loss function. The experiments are thoroughly analyzed and discussed in Section 4. Finally, Section 5 provides a comprehensive summary of this work.

2. Related Work

Currently, infrared and visible images fusion techniques can be mainly categorized into two groups: traditional methods and deep learning-based methods.

2.1. Traditional Methods

Traditional methods employ mathematical transformations to convert the source image into a corresponding transformed domain. In this domain, image fusion is guided by manually designing activity levels measurement or fusion rules. They can be categorized into three types based on their fundamental principles: multi-scale transform-based methods, sparse representation-based methods and subspace clustering-based methods. Multi-scale transform-based methods [20] focus on analyzing images at various scales to capture fine details and contextual information across different resolutions. Typically, initial image decomposition uses techniques such as wavelet transform [21], Laplace pyramid [22] and non-downsampled contourlet transform [23]. These decomposition methods can extract detailed and overall structural information of an image at various scales. For instance, Tang et al. [4] use a method in which the source image is decomposed into a series of base layers and detail layers through the application of a simple weighted least squares filter. Then, they propose a sub-window variance filter for the fusion of detail layers. For the base layers, they have developed a fusion strategy that integrates visual saliency mapping with the principle of adaptive weight allocation. We are inspired by these advanced methods, and incorporate convolutions of different scales into the fusion network to capture details and contextual information.

Sparse representation-based methods comprise various techniques, such as joint sparse representation as demonstrated by Wu et al.'s [24] work on infrared image fusion and potentially low-rank representation explored by Li et al. [25]. These methods aim to construct an overcomplete dictionary from numerous high-quality images. Subsequently, the sparse representation coefficients of the source images are derived using this learned dictionary and then fused based on a pre-defined fusion rule. Inspired by this paper, we introduced joint loss, which optimizes the loss of multiple tasks through joint optimization. The model can better integrate information between different tasks and improve the understanding of task correlation. To improve the performance and applicability of sparse representation methods, researchers propose several improvements to sparse representation methods. These include the introduction of additional constraints within sparse representation and the utilization of various sparse induced functions, such as low-rank decomposition methods [26], structural information methods [27], etc.

Subspace clustering-based methods comprise independent component analysis (ICA) [28], principal component analysis (PCA) [29] and non-negative matrix factorization (NMF) [30]. These methods can accurately perform modeling for images and enhance the capture of structural details and features. As a result, superior fusion outcomes are achieved. Recently, several refined methods based on subspace clustering have been proposed, such as integrated subspace clustering fusion [31], sparse subspace clustering [32], etc. By employing these improved methods, the fusion process achieves higher precision and detail retention. Although the above methods achieve success in certain fusion tasks, the transformations used in traditional methods are too complex. Additionally, the selection of fusion rules is limited and the handcrafted features often fail to address complex scenarios.

2.2. Deep Learning-Based Methods

Deep learning-based methods utilize neural networks to extract features from multiple input images and effectively fuse these features. As a result, a higher-quality output image is produced. These methods enable end-to-end learning, multi-level feature extraction and cross-modal fusion. They can be categorized into two types based on their fundamental principles: image fusion based on CNNs and image fusion based on GANs.

Image fusion methods based on CNNs have achieved significant progress over time. Initially, Liu et al. [33] only utilized a pre-trained neural network to produce weight maps

and extract features, followed by a pyramid model to complete the fusion process. This approach inherently limits the effectiveness of the fusion results. With further research, Zhang et al. [34] introduced an end-to-end framework by modifying the loss function to adjust different hyperparameters for various fusion tasks. Subsequently, Li et al. [35] developed the RGB-induced detail enhancement (RDE) module along with the depth cross-modal feature modulation (CFM) module to facilitate the transfer of additional information from RGB to HSI. These modules can provide more direct and meaningful representations, thereby further facilitating edge recovery. To avoid the limitations of information loss and fusion of feature information, Karim et al. [36] proposed a multilayer triple dense network (DN) for IR and visible (VIS) image fusion, which completely exploits salient features and utilizes the residual features from the combination of input images adaptively. However, existing image fusion algorithms do not consider lighting factors in the modeling process. Tang et al. [37] add lighting perception to the image fusion network to highlight the intensity distribution of the target. Although the above methods have achieved good performance, the specificity of the fusion task and the absence of labels are still obstacles. Considering these problems, we shift our focus to develop a model using a semi-supervised approach in our paper.

The unique challenge for image fusion has driven research on the unsupervised distribution estimation using GANs. Ma et al. [18] were the first to use an adversarial strategy by setting up a competition between the fused image and the visible images to retain more textural details. However, this initial approach destroyed the data distribution between visible and infrared images. To address this issue, the Dual Conditional Generative Adversarial Network (DCGAN) [17] was introduced to incorporate infrared images into the adversarial process. On this basis, Li et al. [38] proposed a refined GAN framework with a multi-scale mechanism to improve performance. It guides the network with a specialized loss function, which makes it easier to concentrate on specific regions of interest. To address the imbalanced contrast and texture in the fused image, a novel GAN with multi-class constraints was developed and transformed the fusion process into a multi-distribution estimation task [39]. In addition, in order to extract accurate information in scenes with abnormal darkness or brightness fluctuations, Sakkos et al. [40] proposed a triple multi-task generative adversarial network that effectively simulates the semantic relationship between dark and bright images and performs end-to-end binary segmentation. Although deep learning-based methods have achieved significant success in the field of image fusion, they still have some limitations. Firstly, current deep learning-based models do not effectively select and utilize the information, which limits the extraction and utilization of key features during the training of the fusion network framework. Secondly, existing models primarily rely on visible images under daytime or high-light conditions. In fact, visible images lose important texture details in low-light conditions. This leads to performance degeneration. Finally, current models do not fully consider the inherent differences in the characteristics of infrared and visible images. This limits their ability to effectively utilize the complementary information present in the two types of images.

3. Method

This section introduces the overview of the proposed network, namely the Self-Attention Progressive Network. Then, the multi-state joint feature extraction, Difference-Aware Propagation Module and loss function are presented in detail.

3.1. Overview

Figure 1 illustrates the proposed framework. The network architecture comprises a multi-state joint feature extraction module (MSJFEM), Difference-Aware Propagation Module (DAPM) and image reconstruction module (IRM). To enhance the contextual perception capability of multi-modal images, we adopt enhanced self-attention learning to design a MSJFEM that extracts deep features by leveraging additional contextual information between input keywords. On this basis, the extracted deep features are processed differ-

entially through the DAPM to better highlight the edge details of the image. Considering noise and artifacts in the fusion process, IRM is proposed to further refine and reconstruct the details of the images. Additionally, a light-aware sub-network is integrated into framework to improve the model's adaptability to lighting conditions. Therefore, we introduce the light intensity loss to measure the probability of scene illumination. By combining Transformer and CNNs, this framework employs a designed loss function to autonomously learn critical image features based on light intensity, such as texture details and unique targets. As a result, it effectively fuses meaningful information from the source images in various conditions.

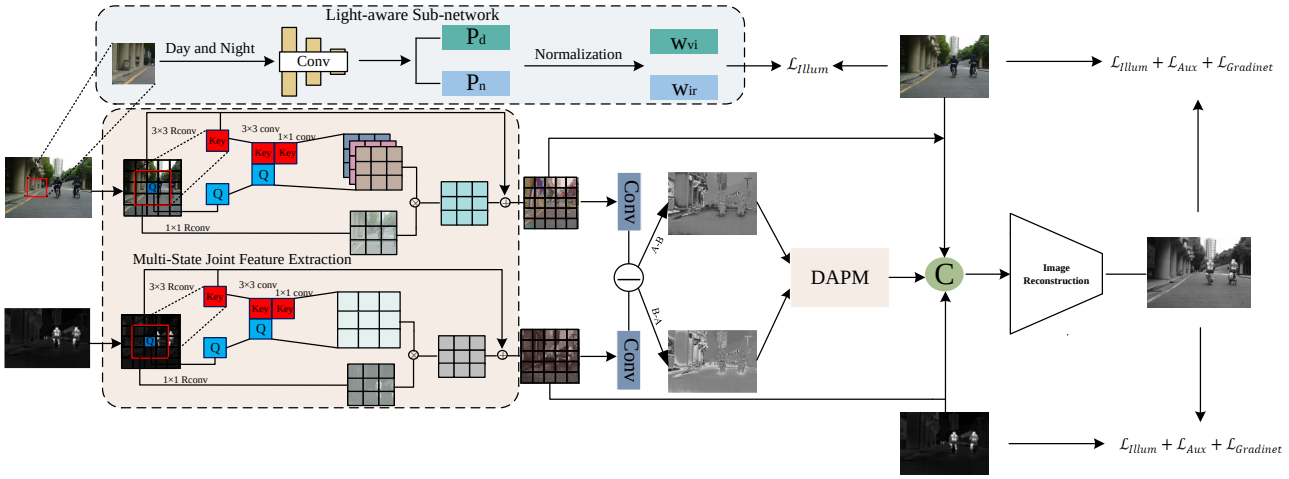


Figure 1. Overall workflow of the proposed method. A multi-state joint feature extraction module is used to process infrared and visible light images by constructing a light-aware sub-network to predict the lighting conditions. Then, a Difference-Aware Propagation Module is used to enhance the image edge details to generate the final fused image.

3.2. Multi-State Joint Feature Extraction

In the traditional self-attention mechanism, individual queries and keys are employed at each location to generate the attention matrix. As a result, the mechanism does not fully exploit the rich context information available between neighboring keys. To address this issue, the multi-state joint feature extraction module is proposed. This module improves self-attention learning utilizing extra contextual information between input keys. It also integrates context mining with the self-attention learning process within a unified architecture. Here, a new multi-head self-attention block is proposed, whose framework is shown in Figure 2. Specifically, given an input feature map $X \in R^{H \times W \times C}$, this module transforms X into the value $V = XW_V$ and key $k = XW_K$ through the embedding matrix W_V and W_K, and the queries are not processed. Meanwhile, the embedding matrix is implemented in space as a 3 × 3 convolution. Then, the input key is context encoded through a 3 × 3 convolution. This process reflects the static context between local neighboring keys. The learned context keycontent m is utilized as a static representation and then cascades with Q. The concatenation is followed by depthwise separable convolutions to extract and integrate spatial and channel features. This reduces many parameters and the computational complexity of the module. In addition, we introduce reflection-padding convolutions to deal with the feature map boundaries more effectively. The process can be defined as

$$P = R_{1\times1}(R_{3\times3}(m \odot Q)), \tag{1}$$

where P denotes the features obtained from the above process, $\odot$ denotes cascade and $R(\cdot)$ denotes convolution.

To enhance self-attention learning by mining additional guidance from static context m, a Softmax function is applied to the channel dimension, resulting in an attention matrix.

Then, a matrix multiplication is carried out between the attention matrix and the value V to generate the weighted attention feature map n. The process can be defined as

$$n = V * Soft\max(P), \tag{2}$$

where $*$ denotes a local matrix multiplication operation. The output generated by MSJFM is the fusion of static and dynamic context of the input feature map represented by an add residual connection.

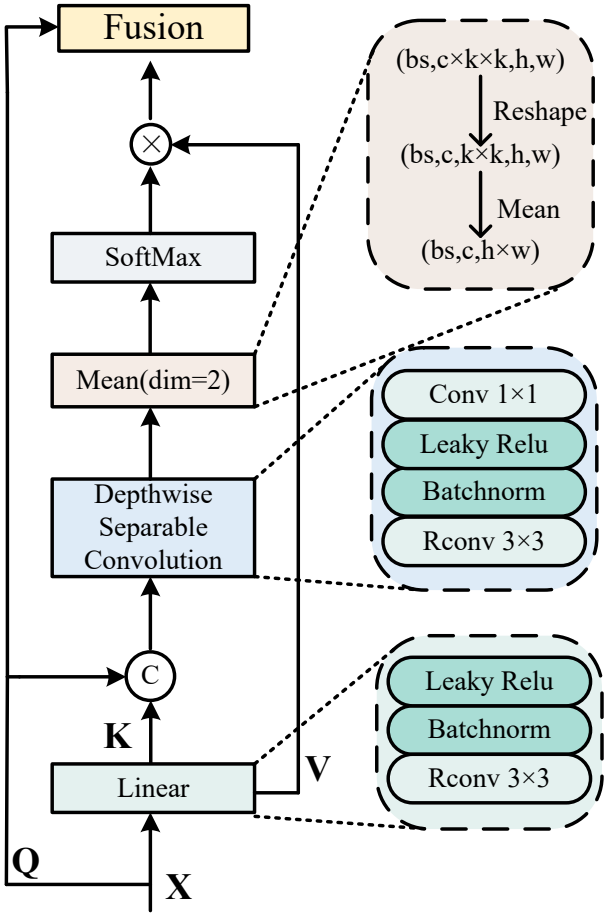


Figure 2. Detailed view of the multi-state joint feature extraction module.

3.3. Difference-Aware Propagation Module

To better integrate complementary and common features, the Difference-Aware Propagation Module (DAPM) is proposed, whose flowchart is depicted in Figure 3. Specifically, depth information obtained from infrared and visible images using the MSJFEM is used as input for this module. In this module, cross-modal difference calculation is performed to obtain complementary features. The channel weighting process is employed to integrate complementary information. The process can be defined as

$$F_{ir} = F_{ir} \odot \zeta(G(F_{vi} - F_{ir})) \otimes (F_{vi} - F_{ir}), \tag{3}$$

$$F_{vi} = F_{vi} \odot \zeta(G(F_{ir} - F_{vi})) \otimes (F_{ir} - F_{vi}), \tag{4}$$

where $\odot$ denotes channel cascade. $\otimes$ denotes channel multiplication. ζ represents activation function, and G represents global average pooling. In particular, the module uses a parameters learnable Swish activation function that it can be dynamically adjusted according to the distribution of the input data, automatically adapt to the feature distribution of different modalities, and balance the information differences between different modalities.

Firstly, the initial parameters and the active value range of the Swish activation function are given. Then, the complementary features are put into the global average pooling and compressed to a vector. In this process, the Swish function automatically adjusts its shape and self-optimizes based on the initial parameters and active values. Finally, the optimized channel weights are multiplied with the complementary features, and the resulting features are cascaded with the original ones to facilitate cross-modal information interaction and enhance the utilization of complementary information. Among them, complementary information reflects the complementary features of different modalities. Meanwhile, through strategies such as global average pooling and cascading features, the DAPM module can effectively integrate information from different modalities and maintain balance between them.

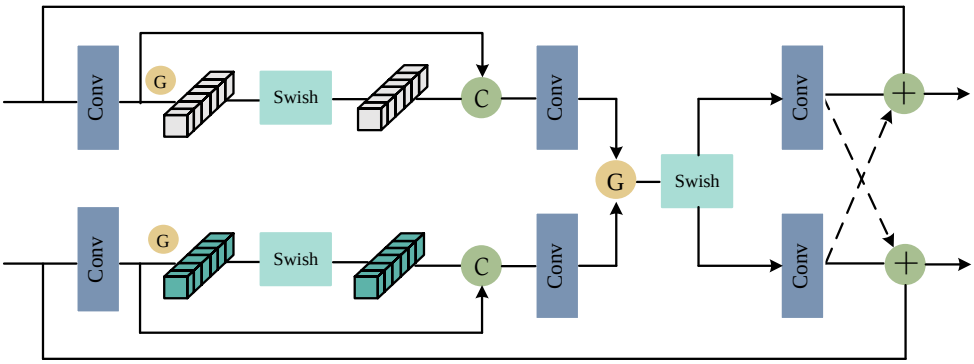


Figure 3. Schematic structure of the Difference-Aware Propagation Module (DAPM).

3.4. Light-Aware Sub-Network

The light-aware sub-network consists of four convolutional blocks and two fully connected blocks. The first two convolutional blocks have a 3×3 convolutional layer with batch normalization and a corrected linear unit (ReLU) activation layer. To simplify the network, a max pool was added in the subsequent convolutional blocks. The last two fully connected blocks include max pooling, batch normalization and ReLU activation layers to improve the network generalization ability and solve the problem of gradient vanishing. Given that the goal of this network is to distinguish between night sky and daytime conditions, it is crucial to set the final neuron count to 2. From this, we can obtain the illumination probability. In order to further enhance the sensitivity of the network to different lighting conditions, we normalize the illumination probability to obtain the light-aware weight, which is adapted to different lighting conditions through the illumination allocation mechanism.

3.5. Loss Function

3.5.1. Loss Function of Light-Aware Sub-Network

The performance of the fusion process significantly depends on the precision of the light-aware sub-network. It belongs to a binary classification task, which is used to compute the probability values of the images belonging to day and night. Therefore, to guide the training of the light-aware sub-network, we adopted a cross-entropy loss fusion method, which is defined as

$$\mathcal{L}_{ls} = -z \log \partial(y) - (1 - z) log(1 - \partial(y)), \tag{5}$$

where z represent the labels of the day and night images, y denotes the classification output of the light-aware sub-network (p_n, p_d), and ∂ represents the Softmax function of normalized output probabilities.

3.5.2. Loss Function of Self-Attention Progressive Network

The loss function dictates the kind of information preserved in the fused image, as well as the relative importance assigned to various types of information. The Self-Attention

Progressive Network overall objective function comprises illumination loss, auxiliary intensity loss, and gradient loss function, which is defined as

$$\mathcal{L}_{SAP} = t_1\mathcal{L}_{illum} + t_2\mathcal{L}_{aux} + t_3\mathcal{L}_{gradient}, \tag{6}$$

where t_1, t_2 and t_3 are the balance factors. In the image fusion process, to alleviate the problem of detail loss in visible images under environments with large differences in light intensity, we introduce illumination loss to enhance the detail content of visible under adverse lighting conditions. The auxiliary intensity loss guides the system to dynamically maintain the optimal intensity distribution. This ensures a balanced intensity distribution between infrared and visible images. Guided by gradient loss, the fusion outcome maintains clear edge details while preserving the texture features inherent in the original images.

The illumination loss can promote the Self-Attention Progressive Network to adaptively integrate contextual information according to lighting conditions. It is composed of light-aware weights and intensity loss. The illumination perception weights serve to adjust the weight restrictions on the fused image, whereas the intensity loss functions as a metric to quantify pixel-level disparities between the fused image and its constituent source images. The illumination loss is defined as

$$W_{ir} = \frac{p_n}{p_d + p_n}, \tag{7}$$

$$W_{vi} = \frac{p_d}{p_d + p_n}, \tag{8}$$

$$\mathcal{L}_{illum} = W_{ir} \cdot \mathcal{L}_{ls}^{ir} + W_{vi} \cdot \mathcal{L}_{ls}^{vi}, \tag{9}$$

where W_{ir} and W_{vi} denote the light-aware weights. To obtain more contextual information, the two source images exhibit different intensity distributions under varying lighting conditions. Consequently, the weight constraints of the fused image are adjusted based on the lighting environment. $\mathcal{L}_{ls}^{ir}$ and $\mathcal{L}_{ls}^{vi}$ represent the intensity loss of the infrared and visible images, respectively, and the intensity loss is defined as

$$\mathcal{L}_{ls}^{ir} = \frac{1}{HW}\left\|I_f - I_{ir}\right\|_1, \tag{10}$$

$$\mathcal{L}_{ls}^{vi} = \frac{1}{HW}\left\|I_f - I_{vi}\right\|_1, \tag{11}$$

where $\|\cdot\|_1$ represents the L1 norm of the vector, I_f denotes the fused image, and I_{ir} and I_{vi} represents the source images.

In illumination loss, although the intensity information from the source image is dynamically preserved according to the real-time lighting conditions, it cannot ensure that the fused image maintains the optimal intensity distribution. Therefore, the auxiliary intensity loss is introduced, which is defined as

$$\mathcal{L}_{aux} = \frac{1}{HW}\left\|I_f - \max(I_{vi}, I_{ir})\right\|_1, \tag{12}$$

where $\max(\cdot)$ denotes the maximum value in the element-by-element selection.

While maintaining the optimal intensity distribution, ensure that the fused image retains rich texture details. As a result, we incorporate gradient loss to enhance the image texture detail retention, which is defined as

$$\mathcal{L}_{gradient} = \frac{1}{HW}\left\|\left|\nabla I_f\right| - \max(|\nabla I_{vi}|, |\nabla I_{ir}|)\right\|_1, \tag{13}$$

where ∇ denotes the Sobel operator and $|\cdot|$ denotes the calculation of the absolute value.

4. Experiment

4.1. Datasets

4.1.1. Msrs Dataset [37]

The dataset is a novel infrared-visible image fusion multispectral collection, derived from the MFNet dataset. It comprises 1444 pairs of images, where each image has a corresponding color visible and an infrared image. These images are captured during the day and night in various traditional traffic scenes, such as campuses, streets and rural areas.

4.1.2. TNO Dataset [41]

The dataset contains 261 image pairs, which offer visible light, near-infrared and long-wave infrared nighttime images of various military and surveillance scenes. These images depict different object targets (such as people, vehicles) against various backgrounds (e.g., rural, urban).

4.1.3. RoadScene Dataset [42]

The dataset comprises 221 infrared-visible image pairs, which are generated from FLIR (Forward-Looking Infrared) video after preprocessing steps such as alignment and cropping. It includes a rich variety of scenes, such as roads, pedestrians, etc.

4.2. Evaluation Metrics and Comparison Methods

To showcase the superiority of proposed method, nine approaches for infrared and visible images fusion are employed, including SSR-Laplacian [22], FusionGAN [18], RFN-Nest [13], SEAFusion [43], SwinFusion [44], U2Fusion [42], TarDAL [45], UMF-CMGR [46], and CDDFuse [47]. Based on training methods, these approaches can be categorized into two groups: supervised and unsupervised. The supervised methods include RFN-Nest, SEAFusion, SwinFusion, and TarDAL, while the remaining techniques fall under the unsupervised category.

To evaluate the performance, six metrics are utilized, i.e., Entropy (EN), Mutual Information (MI), Average Gradient (AG), Structural Similarity Index (SSIM), Visual Information Fidelity (VIF) and Q_{abf}. From the perspective of statistical information measurement, EN is used to measure the complexity or uncertainty between the source image and the fused image information, while MI is employed to quantify the similarity between the visible images and the fused image. From the perspective of mathematical statistics, AG is employed to assess the sharpness of edges and the quantity of detailed information present in the fused image. In terms of image quality assessment, SSIM is an indicator used to measure the similarity between two images and VIF is employed to assess the effect of image distortion on visual information content. Both reflect the human visual system perception of image quality. Q_{abf} is used as a measure of the degree to which detailed edge features from the original image is transferred to the fused image. Moreover, higher values of MI, EN, AG, SSIM, VIF and Q_{abf} signify superior fusion performance.

4.3. Implementation Details

The MSRS dataset serves as the basis for training both the light-aware sub-network and the Self-Attention Progressive Network. Given the limited number of training samples, we randomly crop 64 patches from each image. Each patch undergoes augmentation through random flipping, rotation, and rolling. Prior to inputting the patches into the network, all images are normalized to the range of [0, 1].

Specifically, we train the light aware sub-network and the Self-Attention Progressive Network in sequence. Firstly, the cropped visible light image is used to train a light-aware sub-network to obtain light-aware weights. Then, utilize these light-aware weights to construct illumination loss during the training process of the Self-Attention Progressive Network. We adopt ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to optimize the both networks. The initial learning rate is established as 10^{-4} and is subsequently reduced through an exponential decay during the training process. The light-aware sub-network

is configured with a batch size of 128 and 100 training epochs. For the Self-Attention Progressive Network, the batch size is set to 32 with 30 training epochs. For Equation (6), we set $t_1 = 3$, $t_2 = 7$, and $t_3 = 50$, which are obtained through cross-validation and iterative adjustment. To preserve color information, the visible image is first converted to the YCbCr color space, and the Y, Cb, and Cr channels are separated. Then, a variety of fusion techniques are utilized to combine the Y channel of visible and infrared images. Finally, the merged Y channel is reorganized with the original Cb and Cr channels and transformed back into the RGB color space [48].

4.4. Performance Comparison with Existing Approaches

To fully evaluate the performance of the proposed method, we perform quantitative as well as qualitative comparisons with nine other methods on the MSRS dataset, RoadScene dataset, and TNO dataset. Specifically, this model is trained exclusively on the MSRS dataset for comparative experiments and then generalized to both the RoadScene and TNO datasets for assessing generalization capabilities.

4.4.1. Quantitative Evaluation

Table 1 presents the quantitative outcomes of six evaluation metrics for 150 image pairs from the MSRS dataset. It can be seen that the proposed method shows superiority in some metrics. The best EN metric means that the proposed method obtains more information from the source image, which indirectly reflects that the model can extract more information from the source image after adding a light-aware sub-network, especially in night scenes. The best AG and Q_{abf} metrics indicate that more edge information is retained in the fusion results, benefiting from the use of reflection filled convolution and the proposed DAPM. In addition, higher VIF and SSIM indicate that the fused image has satisfactory visual effects.

Thirty-five image pairs are randomly selected from the RoadScene dataset and twenty-five image pairs from the TNO dataset for direct testing within the proposed method. The quantitative results are presented in Tables 2 and 3. It can be seen that the proposed method shows superiority in some metrics. Higher EN and MI metrics indicate that greater extraction of information from the source images, and the quality of the images is better. This indirectly reflects that the proposed model pays more attention to the coordination between the EN and MI. The proposed method exhibits average performance in terms of the AG metric. This is relatively reasonable, as the two datasets mainly contain daylight scenes, and the proposed method tends to adjust the intensity of nighttime images to reduce the contrast of the fused images. In addition, the higher VIF and SSIM values indicate that the fused image is more similar to the source image and has a satisfactory visual effect.

Table 1. Evaluation metrics for 10 algorithms on the MSRS dataset. Bold text highlights the top-performing results and underlined text represents the second-best outcomes.

Model	En	MI	AG	VIF	SSIM	Q_{abf}
SSR-Laplacian	3.437	1.203	2.131	0.138	0.217	0.134
FusionGAN	5.459	2.104	1.345	0.358	0.415	0.127
RFN-Nest	6.113	2.709	1.683	0.557	0.621	0.266
SEAFusion	6.755	4.056	3.425	0.953	0.933	0.647
SwinFusion	6.714	4.267	3.201	0.959	0.945	0.595
U2Fusion	5.637	2.467	2.635	0.535	0.729	0.376
TarDAL	3.573	1.391	3.061	0.177	0.122	0.135
UMF-CMGR	5.748	2.214	2.251	0.441	0.531	0.289
CDDFuse	<u>6.761</u>	**5.165**	<u>3.436</u>	**1.037**	<u>0.963</u>	<u>0.648</u>
Ours	**6.822**	<u>4.382</u>	**3.742**	<u>0.993</u>	**1.034**	**0.678**

Table 2. Evaluation metrics for 10 algorithms on the RoadScene dataset. Bold text highlights the top-performing results and underlined text represents the second-best outcomes.

Model	En	MI	AG	VIF	SSIM	Q_{abf}
SSR-Laplacian	6.779	1.683	6.423	0.356	0.525	0.193
FusionGAN	7.279	2.645	3.559	0.338	0.617	0.286
RFN-Nest	7.223	2.276	3.361	0.419	0.735	0.329
SEAFusion	7.304	2.954	7.027	0.591	0.837	0.469
SwinFusion	6.942	<u>3.572</u>	4.249	0.683	0.811	<u>0.486</u>
U2Fusion	6.991	2.129	5.553	0.442	0.808	0.455
TarDAL	7.324	1.844	**13.481**	0.343	0.525	0.263
UMF-CMGR	6.829	2.366	3.767	0.476	<u>0.889</u>	0.429
CDDFuse	**7.609**	3.101	<u>6.826</u>	<u>0.639</u>	0.861	0.462
Ours	<u>7.349</u>	**3.577**	5.488	**0.693**	**0.891**	**0.573**

Table 3. Evaluation metrics transposed for 10 algorithms on the TNO dataset. Bold text highlights the top-performing results and underlined text represents the second-best outcomes.

Model	En	MI	AG	VIF	SSIM	Q_{abf}
SSR-Laplacian	5.934	1.071	5.213	0.474	0.503	0.174
FusionGAN	6.234	2.011	1.953	0.289	0.527	0.165
RFN-Nest	6.931	1.781	2.443	0.514	0.773	0.327
SEAFusion	6.927	2.521	5.332	0.662	0.923	0.446
SwinFusion	6.836	3.078	3.908	0.658	**1.052**	<u>0.521</u>
U2Fusion	6.924	1.825	4.923	0.551	0.939	0.412
TarDAL	6.477	1.049	**19.231**	0.351	0.421	0.195
UMF-CMGR	6.422	1.791	2.827	0.524	<u>1.033</u>	0.392
CDDFuse	**7.265**	<u>2.546</u>	5.151	<u>0.703</u>	0.967	0.473
Ours	<u>6.976</u>	**3.084**	**5.513**	**0.708**	0.984	**0.582**

4.4.2. Qualitative Evaluation

To demonstrate that the proposed network can fuse meaningful information from source images in different environments, one daytime image and one nighttime image on the MSRS dataset are selected for qualitative evaluation. The visual results are presented in Figures 4 and 5.

In day scenes, visible images contain abundant texture information. Thermal radiation information from infrared images can be used as complementary information to visible images. This suggests that the fusion algorithm must maintain the texture details of the visible images, enhance the salience of targets, and minimize spectral distortion. As shown in Figure 4, FusionGAN, SSR-Laplacian and TarDAL do not retain the detailed information of visible images well. Although the above comparison networks can significantly fuse the texture details of the visible images with the target details in the infrared image, FusionGAN, RFN-Nest, SEAFusion, U2Fusion and TarDAL have the drawbacks of blurry or incomplete infrared target edges. Only SwinFusion, CDDFuses and the proposed method completely retain the contour information of the person, as shown by the red boxes. In addition, the green boxes are used to illustrate background areas affected by varying degrees of spectral pollution. While TarDAL is less affected by spectral pollution in the background, it also preserves the texture of the visible images to a lesser degree. Although CDDFuses and the proposed method have similar architectures, we achieve clearer outlines compared with CDDFuses.

In night scenes, thermal radiation information in infrared images contains distinct targets, while limited detail information from visible images can serve as a supplement to the infrared images. As illustrated in Figure 5, all networks effectively maintain the prominent targets from the infrared image. However, SSR-Laplacian, RFN-Nest, U2Fusion and TarDAL are unable to clearly reveal background details in dark environments due to the absence of visible images detail information. Although FusionGAN and UMF-CMGR can reveal background information in the dark, they exhibit significant blur, as shown by

the green box. Additionally, SEAFusion and SwinFusion excel at preserving background information. However, they are unable to retain the details of both the tree trunks and the road fences simultaneously. This phenomenon is shown by the red boxes.

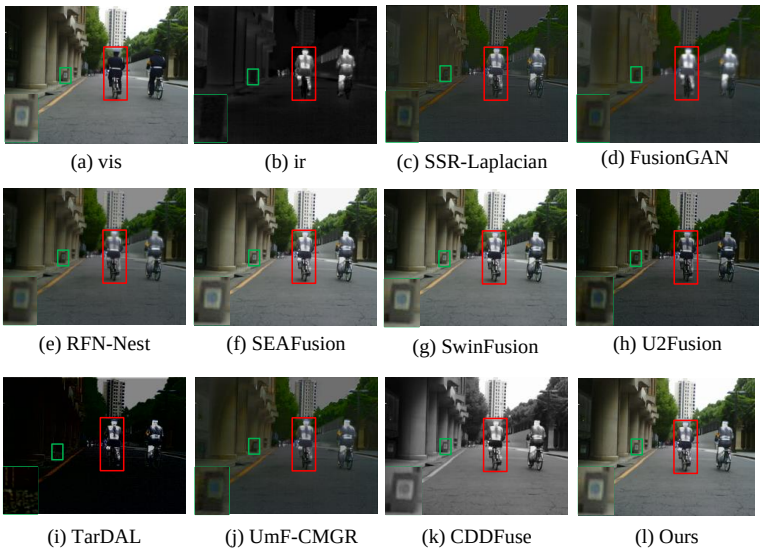


Figure 4. Visual results during the day on the MSRS dataset. (**a**,**b**) are the source images, (**c**–**k**) are nine advanced methods, and (**l**) is our proposed method.

The visual results of different methods on the TNO dataset and the RoadScene dataset are shown in the Figures 6 and 7. Although all networks maintain the distribution of the intensity of critical targets, the person in SSR-Laplacian, FusionGAN, RFN-Nest and UMF-CMGR appears blurred, as shown by the red boxes. The images in FusionGAN, RFN-Nest, UMF-CMGR and TarDAL exhibit uneven contrast, with a significant difference in contrast between bright and dark regions. Additionally, the backgrounds in FusionGAN, SSR-Laplacian, RFN-Nest and TarDAL are affected by varying degrees of spectral pollution, which compromises the visual effects, and this issue is shown by the green box. The method not only adjusts the image contrast but also preserves more texture and edge details while reducing spectral pollution in the background.

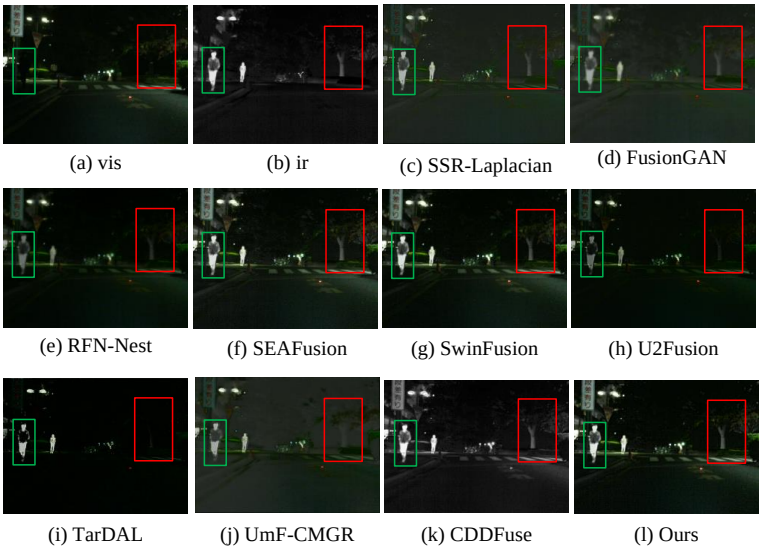


Figure 5. Visual results at night on the MSRS dataset. (**a**,**b**) are the source images, (**c**–**k**) are nine advanced methods, and (**l**) is our proposed method.

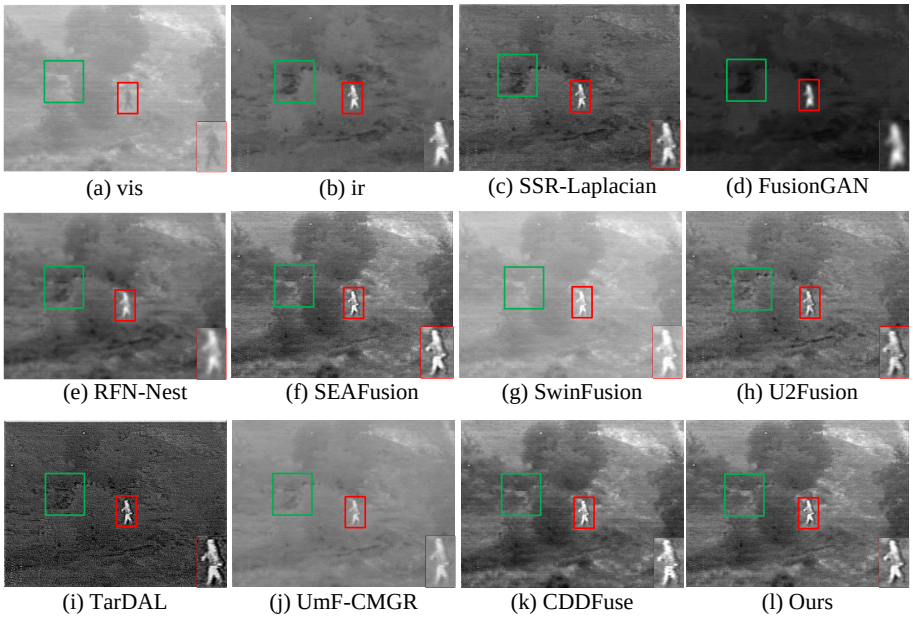


Figure 6. Visual results on the TNO dataset. (**a**,**b**) are the source images, (**c**–**k**) are nine advanced methods, and (**l**) is our proposed method.

In summary, our experimental results do not show any issues such as poor contrast, thermal target degradation, texture blurring, and spectral contamination in the three datasets mentioned above. This indicates that the proposed network exhibits superior performance in light complementarity, as well as in preserving texture and edge details. This is attributed to the proposed MSJFEM and DAPM. However, under extremely low-light conditions or when images are severely blurred, the method does not achieve the best results. The light-aware sub-network we introduced is designed to balance lighting.

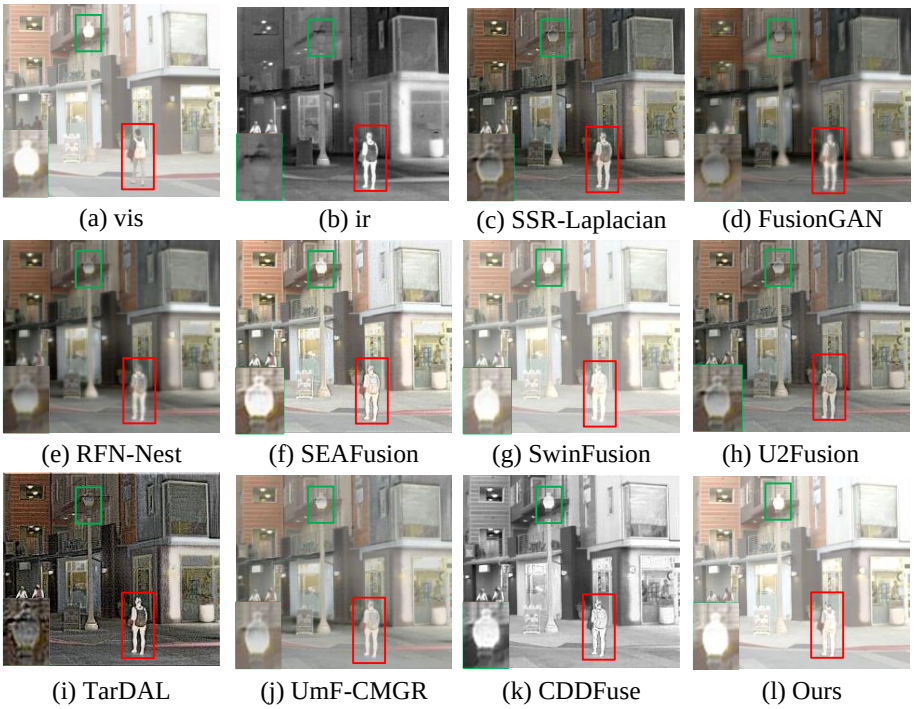


Figure 7. Visual results on the RoadSence dataset. (**a**,**b**) are the source images, (**c**–**k**) are nine advanced methods, and (**l**) is our proposed method.

4.5. Efficiency Comparison

Running efficiency is one of the important factors for evaluating model performance. In Table 4, we provide the average running time of 10 methods on the MSRS, RoadScene, and TNO datasets. From the table, it can be seen that deep learning methods have significant advantages in operation, while traditional methods take longer to fuse images. Although the transformation method we use contains a large number of parameters, our technology still achieves the optimal fusion effect while accelerating processing speed. In addition, due to the need to process the static and dynamic context of the image during the multi-state joint feature extraction stage, our running time is slightly inferior compared to other algorithms. In particular, CDDFuse has a similar structure to our method; however, our method still has significant advantages in terms of computational speed.

Table 4. The running time of 10 algorithms on three datasets. Bold text highlights the top-performing results and underlined text represents the second-best outcomes.

Time	MSRS	RoadScene	TNO
SSR-Laplacian	1.303 ± 0.131	1.213 ± 0.108	1.813 ± 0.114
FusionGAN	1.021 ± 0.108	0.7821 ± 0.256	0.874 ± 0.336
RFN-Nest	1.421 ± 0.436	0.885 ± 0.271	1.542 ± 0.724
SEAFusion	0.546 ± 0.239	0.673 ± 0.359	**0.434 ± 0.264**
SwinFusion	2.452 ± 0.136	1.734 ± 0.352	2.218 ± 0.214
U2Fusion	1.134 ± 0.376	0.745 ± 0.254	1.335 ± 0.579
TarDAL	<u>0.489 ± 0.112</u>	<u>0.588 ± 0.234</u>	0.579 ± 0.321
UmF-CMGR	**0.379 ± 0.273**	**0.534 ± 0.369**	<u>0.528 ± 0.224</u>
CDDFuse	1.512 ± 0.298	0.875 ± 0.157	1.871 ± 0.312
Ours	0.653 ± 0.121	0.603 ± 0.098	0.611 ± 0.116

4.6. Ablation Study

4.6.1. Study of Multi-State Joint Feature Extraction

The Self-Attention Progressive Network model designs a key module to improve the capability of feature extraction, namely, multi-state joint feature extraction. To verify the effectiveness of this module, we test the model by removing or adding this module on the MSRS dataset, TNO dataset and RoadScene dataset. Table 5 describes the model performance with different module combinations on different datasets.The design integrates context information mining and self-attention learning into a single architecture through multi-state joint feature extraction. It is beneficial for feature representation and enhances the learning of contextual details. From Table 5, it can be observed that the six evaluation metrics of the model with the addition of multi-state joint feature extraction are all higher than those without this module, especially in MI and Q_{abf}, with an average improvement of 35% and 61%, respectively. This suggests that the module effectively preserves information and enhances the learning of edge details.

Table 5. Ablation study on MSRS dataset. Bold indicates the best result.

Components				MSRS Dataset 640 × 480 Pixels					
Baseline	MSJFE	DAPM	IGL	EN	MI	AG	VIF	SSIM	Q_{abf}
✓		✓	✓	6.585	3.443	3.173	0.611	0.648	0.447
✓	✓		✓	6.786	4.328	**3.827**	0.969	0.754	0.665
✓	✓	✓		6.802	4.312	3.476	0.966	1.007	0.658
✓	✓	✓	✓	**6.822**	**4.382**	3.742	**0.993**	**1.034**	**0.678**

4.6.2. Study of Difference-Aware Propagation Module

To enhance feature representation by incorporating complementary and differential information between different modalities, the Difference-Aware Propagation Module is designed to facilitate cross-modal complementary information interaction. To validate the effectiveness of this module, we conduct experiments on the MSRS dataset, TNO dataset and RoadScene dataset by either removing or adding the module to the model. Table 5 reveals that the AG metric without the module is slightly higher than that of the fused image, which is reasonable as the module automatically balances with other modules during the fusion process, especially when there is no illumination loss constraint. In particular, the average improvement of 17% in SSIM indicates that our module performs well in retaining complementary and differential information.

4.6.3. Study of Bidirectional Difference-Aware Propagation Module

There are typically two methods for obtaining image differences: bidirectional subtraction (DAPM) and absolute value subtraction (BDA). To further validate the effectiveness of bidirectional subtraction in the Difference-Aware Propagation Module, we conduct experiments by applying both methods to three datasets while keeping other parameter settings constant. As shown in Table 6, fusion images produced using bidirectional subtraction currently exhibit better visual quality, sharper object edges, and the retention of more information within the fused images. The slight decrease in AG measurement results can be attributed to the bidirectional subtraction ability to preserve directional information of the differences. For instance, the visible images exhibit unclear targets in the nighttime scenes of the MSRS dataset. Additionally, more information from the infrared images is preserved, which results in lower AG measurement outcomes. In the TNO dataset containing more daytime scenes, AG measurement results slightly increase. Although there is a slight decrease in information quantity measurement on the RoadSence dataset, the visual quality of the images is significantly improved.

Table 6. Study of bidirectional difference-aware module on three benchmark datasets. Bold indicates the best result.

Dataset	Method	EN	MI	AG	VIF	SSIM	Q_{abf}
MSRS	Baseline+Basic module+DAPM	**6.822**	**4.382**	3.742	**0.993**	**1.034**	**0.678**
	Baseline+Basic module+BDA	6.821	4.219	**3.854**	0.966	0.749	0.664
TNO	Baseline+Basic module+DAPM	6.976	**3.084**	4.513	**0.708**	**0.984**	**0.582**
	Baseline+Basic module+BDA	**6.983**	3.054	**4.607**	0.653	0.972	0.573
RoadScene	Baseline+Basic module+DAPM	7.349	3.467	**5.488**	**0.693**	**0.981**	**0.573**
	Baseline+Basic module+BDA	**7.374**	**3.657**	5.286	0.639	0.807	0.543

4.6.4. Study of Illumination Guidance Loss

Visible images often lack detail under low-light conditions, which diminishes the key information that fusion networks can extract. To address this issue, the Illumination Guidance Loss is designed to guide the training of the Self-Attention Progressive Network. To validate the effectiveness of the Illumination Guidance Loss, we test the model on the MSRS dataset, TNO dataset and MSRS dataset by removing or adding this loss while keeping all other parameter and configurations constant. As shown in Table 5, the network performance in terms of information quantity and image quality metrics is superior when the Illumination Guidance Loss is included. This indicates that the loss function can adaptively extract important information for different lighting conditions, thereby effectively preventing spectral pollution and texture blurring. The slight decrease in AG measurement can be attributed to the Difference-Aware Propagation Module, which does not have the constraint of this loss and thus plays a more significant role.

Overall, the performance of this module combination surpasses all others, as evident

from the experimental results presented in Tables 5 and 6. It demonstrates the efficiency and excellence of the proposed module.

5. Conclusions

In this paper, we propose a Self-Attention Progressive Network to address the low-illumination problem in visible and infrared image fusion. This network measures the probability of scene lighting by constructing a loss of illumination to enhance the model adaptability to lighting conditions. Meanwhile, a multi-state joint feature extraction module is designed to achieve context mining between keywords and self attention learning of feature maps. On this basis, we introduce a Difference-Aware Propagation Module to better integrate complementary and common features. The experimental results indicate that our proposed method exhibits superiority over existing approaches in both quantitative and qualitative assessments. In the future, we will explore complementary relationships in spatial and planar domains to enhance feature extraction further.

Author Contributions: Methodology, S.L.; resources, S.L.; software, S.L.; supervision, Q.L. and Y.Q.; writing—original draft, M.H.; writing—review and editing, Q.L. and Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160401, in part by the National Natural Science Foundation of China under Grant 62301385, and in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-YBXM-130.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. Due to privacy, the data cannot be made public.

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

References

1. Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; Ma, J. Image fusion meets deep learning: A survey and perspective. *Inf. Fusion* **2021**, *76*, 323–336. [CrossRef]
2. Guo, Y.; Wu, X.; Qing, C.; Liu, L.; Yang, Q.; Hu, X.; Qian, X.; Shao, S. Blind Restoration of a Single Real Turbulence-Degraded Image Based on Self-Supervised Learning. *Remote Sens.* **2023**, *15*, 4076. [CrossRef]
3. Wang, R.; Wang, Z.; Chen, Y.; Kang, H.; Luo, F.; Liu, Y. Target Recognition in SAR Images Using Complex-Valued Network Guided with Sub-Aperture Decomposition. *Remote Sens.* **2023**, *15*, 4031. [CrossRef]
4. Tang, H.; Liu, G.; Qian, Y.; Wang, J.; Xiong, J. EgeFusion: Towards Edge Gradient Enhancement in Infrared and Visible Image Fusion with Multi-Scale Transform. *IEEE Trans. Comput. Imaging* **2024**, *10*, 385–398. [CrossRef]
5. Li, Q.; Yuan, Y.; Wang, Q. Multi-Scale Factor Joint Learning for Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5523110.
6. Ji, C.; Zhou, W.; Lei, J.; Ye, L. Infrared and Visible Image Fusion via Multiscale Receptive Field Amplification Fusion Network. *IEEE Signal Process. Lett.* **2023**, *30*, 493–497. [CrossRef]
7. Deng, C.; Chen, Y.; Zhang, S.; Li, F.; Lai, P.; Su, D.; Hu, M.; Wang, S. Robust dual spatial weighted sparse unmixing for remotely sensed hyperspectral imagery. *Remote Sens.* **2023**, *15*, 4056. [CrossRef]
8. Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; Feng, R. Contrastive Multiview Subspace Clustering of Hyperspectral Images Based on Graph Convolutional Networks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5510514. [CrossRef]
9. Wang, X.; Guan, Z.; Qian, W.; Cao, J.; Wang, C.; Yang, C. Contrast Saliency Information Guided Infrared and Visible Image Fusion. *IEEE Trans. Comput. Imaging* **2023**, *9*, 769–780. [CrossRef]
10. Wang, Z.; Cao, B.; Liu, J. Hyperspectral image classification via spatial shuffle-based convolutional neural network. *Remote Sens.* **2023**, *15*, 3960. [CrossRef]
11. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [CrossRef]
12. Yan, H.; Su, S.; Wu, M.; Xu, M.; Zuo, Y.; Zhang, C.; Huang, B. SeaMAE: Masked Pre-Training with Meteorological Satellite Imagery for Sea Fog Detection. *Remote Sens.* **2023**, *15*, 4102. [CrossRef]
13. 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]
14. Fan, L.; Yuan, J.; Niu, X.; Zha, K.; Ma, W. RockSeg: A Novel Semantic Segmentation Network Based on a Hybrid Framework Combining a Convolutional Neural Network and Transformer for Deep Space Rock Images. *Remote Sens.* **2023**, *15*, 3935. [CrossRef]

15. Li, Q.; Gong, M.; Yuan, Y.; Wang, Q. Symmetrical feature propagation network for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
16. Palsson, B.; Ulfarsson, M.O.; Sveinsson, J.R. Synthesis of Synthetic Hyperspectral Images with Controllable Spectral Variability Using a Generative Adversarial Network. *Remote Sens.* **2023**, *15*, 3919. [CrossRef]
17. 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]
18. 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]
19. Xiao, B.; Xu, B.; Bi, X.; Li, W. Global-Feature Encoding U-Net (GEU-Net) for Multi-Focus Image Fusion. *IEEE Trans. Image Process.* **2021**, *30*, 163–175. [CrossRef]
20. Chen, J.; Li, X.; Luo, L.; Mei, X.; Ma, J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf. Sci.* **2020**, *508*, 64–78. [CrossRef]
21. Liu, S.; Yan, A.; Huang, S. Seismic Data Denoising Based on DC-PCNN Image Fusion in NSCT Domain. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 7502205. [CrossRef]
22. Wu, R.; Yu, D.; Liu, J.; Wu, H.; Chen, W.; Gu, Q. An improved fusion method for infrared and low-light level visible image. In Proceedings of the ICCWAMTIP, Chengdu, China, 15–17 December 2017; IEEE: New York, NY, USA, 2017; pp. 147–151.
23. Zhang, Q.; Maldague, X. An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing. *Infrared Phys. Technol.* **2016**, *74*, 11–20. [CrossRef]
24. Wu, M.; Ma, Y.; Fan, F.; Mei, X.; Huang, J. Infrared and visible image fusion via joint convolutional sparse representation. *JOSA A* **2020**, *37*, 1105–1115. [CrossRef] [PubMed]
25. Li, H.; Wu, X.J.; Kittler, J. MDLatLLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef]
26. Zhao, W.; Rong, S.; Li, T.; Feng, J.; He, B. Enhancing underwater imagery via latent low-rank decomposition and image fusion. *IEEE J. Ocean. Eng.* **2022**, *48*, 147–159. [CrossRef]
27. Sun, Y.; Lei, L.; Liu, L.; Kuang, G. Structural Regression Fusion for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4504018. [CrossRef]
28. Shen, S.; Wang, X.; Wu, M.; Gu, K.; Chen, X.; Geng, X. ICA-CNN: Gesture Recognition Using CNN with Improved Channel Attention Mechanism and Multimodal Signals. *IEEE Sens. J.* **2023**, *23*, 4052–4059. [CrossRef]
29. Xia, Z.; Chen, Y.; Xu, C. Multiview PCA: A Methodology of Feature Extraction and Dimension Reduction for High-Order Data. *IEEE Trans. Cybern.* **2022**, *52*, 11068–11080. [CrossRef]
30. Li, X.; Zhang, X.; Yuan, Y.; Dong, Y. Adaptive Relationship Preserving Sparse NMF for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5504516. [CrossRef]
31. Lv, J.; Kang, Z.; Wang, B.; Ji, L.; Xu, Z. Multi-view subspace clustering via partition fusion. *Inf. Sci.* **2021**, *560*, 410–423. [CrossRef]
32. Chen, Y.; Li, C.G.; You, C. Stochastic sparse subspace clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision, Seattle, WA, USA, 13–19 June 2020; pp. 4155–4164.
33. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolut. Inf. Process.* **2018**, *16*, 1850018. [CrossRef]
34. 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.
35. Li, Q.; Gong, M.; Yuan, Y.; Wang, Q. RGB-induced feature modulation network for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [CrossRef]
36. Karim, S.; Tong, G.; Li, J.; Yu, X.; Hao, J.; Qadir, A.; Yu, Y. MTDFusion: A Multilayer Triple Dense Network for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2023**, *73*, 5010117. [CrossRef]
37. 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]
38. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans. Multimedia* **2020**, *23*, 1383–1396. [CrossRef]
39. 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*, 1–14. [CrossRef]
40. Sakkos, D.; Ho, E.S.; Shum, H.P. Illumination-aware multi-task GANs for foreground segmentation. *IEEE Access* **2019**, *7*, 10976–10986. [CrossRef]
41. Toet, A. The TNO multiband image data collection. *Data Brief* **2017**, *15*, 249–251. [CrossRef]
42. 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]
43. 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]
44. 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]

45. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
46. Wang, D.; Liu, J.; Fan, X.; Liu, R. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv* **2022**, arXiv:2205.11876.
47. 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, Vancouver, BC, Canada, 24 June 2023; pp. 5906–5916.
48. Ram Prabhakar, K.; Sai Srikar, V.; Venkatesh Babu, R. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4714–4722.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.