drones

MDPI

Article

DFA-Net: Multi-Scale Dense Feature-Aware Network via Integrated Attention for Unmanned Aerial Vehicle Infrared and Visible Image Fusion

Sen Shen [1,†], Di Li [2,†], Liye Mei [2,†], Chuan Xu [2], Zhaoyi Ye [2], Qi Zhang [2], Bo Hong [3], Wei Yang [3] and Ying Wang [3,*]

[1] School of Weapon Engineering, Naval Engineering University, Wuhan 430032, China; 1712021019@nue.edu.cn
[2] School of Computer Science, Hubei University of Technology, Wuhan 430068, China; 102101028@hbut.edu.cn (D.L.); meiliye@hbut.edu.cn (L.M.); 20200064@hbut.edu.cn (C.X.); 102101051@hbut.edu.cn (Z.Y.); 102111136@hbut.edu.cn (Q.Z.)
[3] School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, China; hongbo@wsyu.edu.cn (B.H.); yangwei403@wsyu.edu.cn (W.Y.)
* Correspondence: wangying@wsyu.edu.cn
† These authors contributed equally to this work.

Abstract: Fusing infrared and visible images taken by an unmanned aerial vehicle (UAV) is a challenging task, since infrared images distinguish the target from the background by the difference in infrared radiation, while the low resolution also produces a less pronounced effect. Conversely, the visible light spectrum has a high spatial resolution and rich texture; however, it is easily affected by harsh weather conditions like low light. Therefore, the fusion of infrared and visible light has the potential to provide complementary advantages. In this paper, we propose a multi-scale dense feature-aware network via integrated attention for infrared and visible image fusion, namely DFA-Net. Firstly, we construct a dual-channel encoder to extract the deep features of infrared and visible images. Secondly, we adopt a nested decoder to adequately integrate the features of various scales of the encoder so as to realize the multi-scale feature representation of visible image detail texture and infrared image salient target. Then, we present a feature-aware network via integrated attention to further fuse the feature information of different scales, which can focus on specific advantage features of infrared and visible images. Finally, we use unsupervised gradient estimation and intensity loss to learn significant fusion features of infrared and visible images. In addition, our proposed DFA-Net approach addresses the challenges of fusing infrared and visible images captured by a UAV. The results show that DFA-Net achieved excellent image fusion performance in nine quantitative evaluation indexes under a low-light environment.

Keywords: infrared and visible fusion; unmanned aerial vehicles; image fusion; multi-scale feature; unsupervised gradient estimation

check for updates

Citation: Shen, S.; Li, D.; Mei, L.; Xu, C.; Ye, Z.; Zhang, Q.; Hong, B.; Yang, W.; Wang, Y. DFA-Net: Multi-Scale Dense Feature-Aware Network via Integrated Attention for Unmanned Aerial Vehicle Infrared and Visible Image Fusion. *Drones* **2023**, *7*, 517. https://doi.org/10.3390/drones7080517

Academic Editor: Giordano Teza

Received: 27 June 2023
Revised: 1 August 2023
Accepted: 3 August 2023
Published: 6 August 2023

Copyright: © 2023 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

Unmanned aerial vehicle (UAV) infrared and visible image fusion is an important aspect of image fusion since it enables the efficient integration of information on the different features of infrared and visible images [1], contributing to improving the quality of visual information, strengthening target detection and recognition, and enhancing environmental perception, making it a crucial technique for various UAV applications. Among them, infrared images utilize thermal radiation differences to separate the feature object from the background, which can highlight the feature object and are more effective in low-light conditions, though the resolution is lower and less effective. In contrast, visible images are created by capturing the information provided by reflected light with the aid of sensors. Images of this type offer a higher level of spatial resolution and detail, but they

are more sensitive to harsh weather conditions by combining infrared and visible images. Combining the complementary information of these two sensors can produce images with greater features that will be useful for UAV target detection and RGB-T analysis, as well as understanding the scene. Due to this, developing an intelligent algorithm for UAV infrared and visible image fusion with high performance is of great practical importance.

Infrared image fusion methods are currently divided into traditional fusion methods [2–4] and deep-learning-based methods [5–7]. The traditional image fusion method can be categorized into three categories: multi-scale transformation methods (MST) [8–14], sparse representation methods (SR) [15–18], and hybrid methods [19–22]. Among them, multi-scale transformation-based methods are more commonly used, and the algorithms can be further divided into three categories: pyramid-transform-based image decomposition [23], wavelet-transform-based image decomposition, and multi-scale geometric decomposition [24]. In addition, there are some subspace-based methods such as independent component analysis (ICA), principal component analysis (PCA) [25], and non-negative matrix decomposition, which all project high-dimensional images into a low-dimensional subspace to capture the intrinsic structure of the source image. Traditional methods can acquire feature images quickly, but they also have the following drawbacks: (1) inconspicuous features and blurred images are common with manually extracted features due to the low robustness of the extraction process; (2) it is difficult to determine one general method of feature extraction that can be applied to all fusion tasks, since fusion performance is highly dependent on manually extracted features; (3) to perform infrared image fusion, different features may require manual design of different fusion strategies; and (4) traditional fusion methods are inefficient, involve complex models, and require a high level of computational complexity.

To overcome these drawbacks, deep learning is applied to the study of algorithms for infrared image fusion. At this stage, deep-learning-based image fusion methods are classified into two categories: fusion framework based on convolutional neural networks [26–30] and transformer-based fusion architecture [31–37]. In different scene conditions, one would design network frameworks for different infrared radiation fused images, for example, PIAFusion [38] can fuse meaningful information from source images 24 h a day by sensing light conditions, the SDNet [39] can perform multiple fusion tasks in real time, and FusionGAN [40] performs different resolution image fusion without the noise caused by infrared information. SeAFusion [41] can be deployed as a preprocessing module for advanced vision tasks to achieve real-time image fusion. The RXDNFuse [42] algorithm takes full advantage of the hierarchical features of the source images and does not require the manual design of image decomposition metrics and fusion rules. In the field of image fusion, deep learning image fusion algorithms have made significant advances, but there are still several challenges to overcome. On the one hand, fusion effect and fusion efficiency are always in conflict, and video fusion has strict requirements on fusion speed in practical applications, so lightening the model and realizing the low-light environment fusion of infrared and visible images will be an important development in the future. On the other hand, under complex conditions, infrared and visible image fusion is easily affected by the outside world and usually requires image preprocessing before fusion, so achieving fast and efficient image fusion under complex conditions is also an important issue.

To address the above problems, we propose a multi-scale dense feature-aware network via integrated attention for infrared and UAV visible image fusion (DFA-Net). Specifically, we design a dual-channel encoder for learning advanced features of visible and infrared images. In comparison to single-channel feature extraction, dual-channel feature extraction can yield a greater amount of feature information. [43] Moreover, we introduce the dense skip connection structure [44], making the decoder deeply fuse the features from different layers of the encoder, aiming at obtaining the multi-scale feature extraction of the detailed texture of the visible image and the significant target of the infrared image. Furthermore, we use an integrated attention fusion module [45] to refine features of different scales and fuse feature information of different scales, reducing the loss of feature information. Finally, we use an unsupervised gradient estimation and intensity loss to learn significant fusion

features from the UAV infrared and visible images, resulting in generating high-quality fusion images.

The main contributions of this study are summarized below:

1. We design a novel multi-scale dense feature-aware network via integrated attention, which can extract infrared image target features and visible detail texture features, and achieve excellent results in infrared and visible image fusion by multi-scale nesting methods.
2. We develop an integrated attention module for enhancing complementary features of both infrared and visible images, aiming at retaining richer detail information and focusing on salient features during fusion.
3. We combine intensity and gradient loss to refine the fused multi-source information and generate high-quality infrared and visible fused images.
4. We achieve excellent image fusion effects on UAV infrared and visible images.

2. Materials and Methods

2.1. Network Architecture

The traditional technique of image fusion relies mostly on manually designed rules to extract feature information, which does not take into account the rich and complex information contained in both infrared and visible images. As a result of the fusion, features are insufficiently extracted and information is not reproduced accurately. Consequently, we need to extract more detailed information from our fusion model regarding the image features. In order to enhance infrared image information representation, the first step is to identify salient target information. An infrared image mostly contains salient information, such as people, cars, or other objects with heat, and a multi-scale fusion frame is used to learn the data by focusing on the encoder. Additionally, in complex imaging environments, such as low-light environments, it is difficult to see visible detail textures. Therefore, the framework we designed must be able to extract the necessary visible detail information from the detected area and reconstruct and fuse the feature maps derived from the convolutional neural network to achieve the desired results. As a result, we introduce the integrated attention fusion module, which allows better extraction of channel features and further refinement of the feature map after it has been extracted.

As shown in Figure 1, the proposed DFA-Net consists of a dual-channel integrated network and a standard auto-encoder architecture. The image inputs for infrared and visible are separated into two branches. In this way, the same convolution filter can be used to extract the salient features of the two images. Dense skip connections are used by DFA-Net to store detailed texture information. In order to compensate for the loss of the deep position information of the decoder, the two branches are down-sampled separately, and the detailed features of fusion are transmitted to the decoder through the dense jump connection mechanism. Each node except the original node must be up-sampled to achieve dense connections throughout the backbone network.

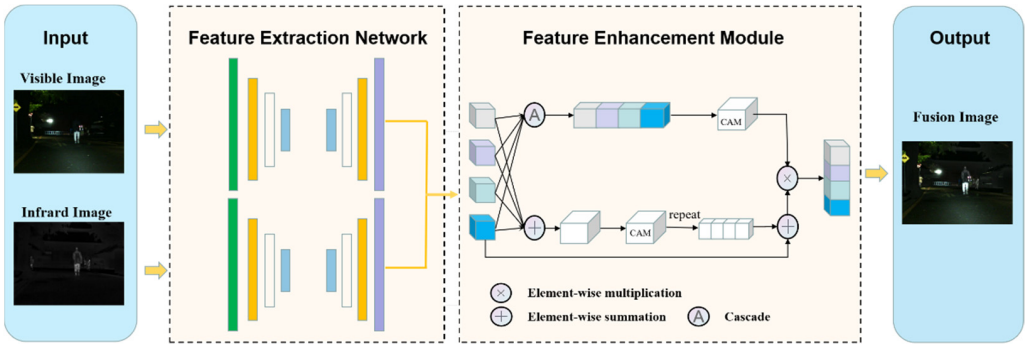

Figure 1. Multi-scale dense feature-aware network via integrated attention (input an infrared image and a visible image, and the feature enhancement module is an integrated attention fusion module).

The infrared and visible images are fed into the dual-channel encoder network. For the output of each node after down sampling, the child decoder restores it to the original scale. The fine positioning characteristics of the encoder are transmitted to the four sub-decoders via a jump connection, which means that the shallow positioning information is applied directly to the deep layer, thus maintaining fine-grained information. Additionally, in order to maximize the use of fine properties, a node in the shallower sub-decoder is coupled with a node in the deeper sub-decoder. In particular, as the encoder deepens, the number of channels in the feature map increases, while as the decoder deepens, the number of channels decreases. Figure 2 shows the feature extraction channel module. The nodes in the figure represent nested convolution blocks. As the color changes from dark to light, it represents the main feature information that the network has extracted. The downward arrow, upward arrow, and dashed arrow indicate 2 × 2 maximum pooling, 2 × 2 upward sampling, and skip connections. Through tensor connections, skip connections combine coding and decoding features in the channel dimension. To observe the complexity of the network more intuitively, we denote the output of node $X_{I,j}$ as $X_{i,j}$.

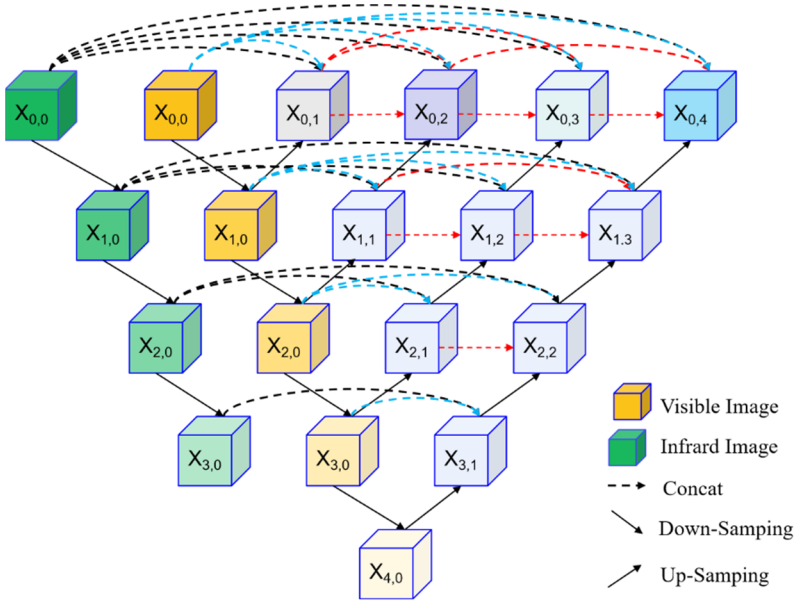

Figure 2. The proposed feature extraction channel module based on multi-scale dense feature-aware network ($X_{i,j}$ denotes the convolutional block, and the dashed lines represent dense connections).

The sub-encoder in the network model obtains multi-channel image feature information through the dual-channel feature extraction layer. A second step is to feed the acquired features into the dense connection module to retain the information about the features of the infrared and visible images as much as possible. In addition, we preprocessed the infrared images and visible images input into the batch standardized block, including a nested convolution kernel of 3, two-dimensional convolution with a step size of 1, two-dimensional batch norm, ReLU activation function with the output channel number of 64, and a number of channels to extract feature information of 128, 256, and 512. Then, an improved ensemble-channel attention module is introduced at the end of the encoder sub-network to focus the salient information of infrared and visible images from two aspects of channel and space and suppress the useless information. To ensure that all the salient features in the model can be utilized, the decoding sub-network is composed of full convolution, and the extracted feature information is reconstructed in the decoding sub-network to output the fusion results of infrared and visible images.

2.2. Integrated Attention Fusion

Considering that the complementary feature information of infrared and visible image fusion is easy to miss, we design an integrated attention fusion module (IAF) to enhance the

complementary features of infrared and visible, aiming to retain richer detailed information and focus on salient feature information in the fusion process. As shown in Figure 3, the IAF module supports plug-and-play, so that it can enhance the channel feature of the input feature graph and finally output through the integrated attention fusion module without changing the size of the input feature map.

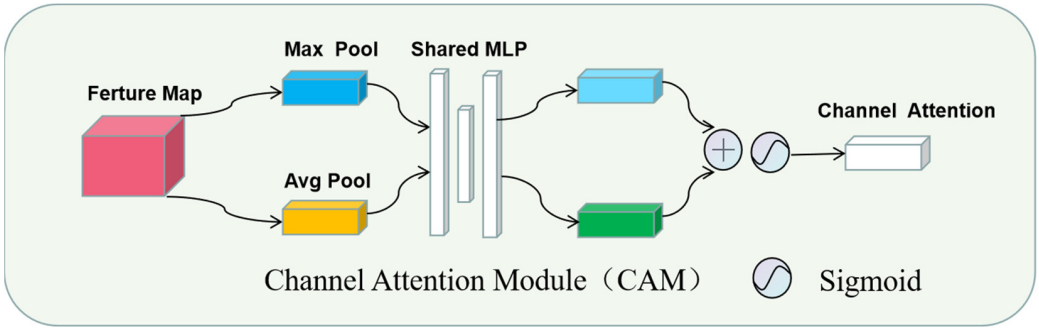

Figure 3. Integrated attention fusion module.

The process of integrating the attention fusion module is as follows: first, the feature map is input, and the average pooling and maximum pooling operations are carried out. Then, the feature extraction is carried out by using convolution to realize cross-channel interaction. There is a parallel operation between average pooling and maximum pooling, in addition to an additional information encoding mode, which improves the quality of the information obtained and makes the features of the fusion image more apparent.

$$\mathrm{CAM}(F) = \sigma(\mathrm{MLP}(\mathrm{Avgpool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \tag{1}$$

$$M_{\mathrm{int}ra} = \mathrm{CAM}(x^{0,1} + x^{0,2} + x^{0,3} + x^{0,4}) \tag{2}$$

$$F_{ensemble} = [x^{0,1}, x^{0,2}, x^{0,3}, x^{0,4}] \tag{3}$$

$$M_{\mathrm{int}er} = \mathrm{CAM}(F_{ensemble}) \tag{4}$$

$$\mathrm{IAF}(F_{ensemble}) = (F_{ensemble} + \mathrm{repeat}_{(4)}(M_{\mathrm{int}ra})) \otimes M_{\mathrm{int}er} \tag{5}$$

where σ represents Sigmoid function, MLP means multi-layer awareness, and Avgpool and MaxPool represent average and maximum pooling operations, respectively. $[\bullet]$ represents a set of feature mapping connections, $\mathrm{repeat}_{(n)}(\bullet)$ indicates a repeated attention operation, and $\otimes$ represents the dot product of matrix elements.

2.3. Fusion Loss

In this paper, our loss function consists of two types of loss terms, gradient loss L_{grad} and intensity loss L_{int}. The loss function of the fused image of infrared and visible is divided into two parts for the purpose of constraining the image. There are two types of loss: the intensity loss, which constrains the apparent intensity of the fused image with appropriate intensity based on the intensity information from the source image, and the gradient loss, which causes the fused image to contain richer information. We define the loss function as

$$L_{total} = L_{\mathrm{int}} + \lambda L_{grad} \tag{6}$$

Among them, the goal of L_{int} is to constrain the apparent intensity of the fused image, while L_{grad} aims to force the fused image to contain more texture detail. Here, λ is used to balance intensity losses and gradient losses. The intensity loss constraint fusion images

retain more useful pixel information, such as contrast, brightness, etc. At the same time, it can also make the visual effect of infrared–visible images look more natural and closer to the visible image. The intensity loss can be defined as follows:

$$L_{\text{int}} = \frac{1}{HW}\sum_{i}\sum_{j}\left(I_{fused_{i,j}} - I_{1_{i,j}}\right)^2 + \alpha\left(I_{fused_{i,j}} - I_{2_{i,j}}\right)^2 \tag{7}$$

where H and W are the height and width of the image, respectively; I_1 is the infrared image, I_2 is the visible image, and $I_{fused(\bullet)}$ is the infrared–visible fused image. The proportional setting strategy is used to adjust α, which constrains the pixel intensity distribution of infrared and visible images. Gradient loss L_{grad} reduces the loss of feature information by gradient descent so that the fused image has richer detail textures. We designed gradient loss L_{grad} to enhance the contrast of image details while compressing the overall dynamic range of the image. After the operation, the feature map can be reconstructed to obtain the enhanced image. The gradient loss can be defined as

$$L_{grad} = \frac{1}{HW}\sum_{i}\sum_{j} S_{1_{i,j}} \cdot \left(\nabla I_{fused_{i,j}} - \nabla I_{1_{i,j}}\right)^2 + S_{2_{i,j}} \cdot \left(\nabla I_{fused_{i,j}} - \nabla I_{2_{i,j}}\right)^2 \tag{8}$$

where $\nabla(\bullet)$ represents the Sobel operator. Moreover, based on the gradient level of the source image, $S(\bullet)$ represents a decision graph generated by the decision block. As a first step, the decision block performs Gaussian low-pass filtering on the source image in order to reduce the influence of noise on gradient judgments. In the next step, we use the Laplacian operator to find the gradient graph and generate the decision graph based on the gradient size on a pixel scale. The calculation formulas for the decision map $S_{1_{i,j}}$ and $S_{2_{i,j}}$ are as follows:

$$S_{1_{i,j}} = sign\left(\left|\nabla\left(L\left(I_{i,j}\right)\right)\right| - \min\left(\left|\nabla\left(L\left(I1_{i,j}\right)\right)\right|, \left|\nabla\left(L\left(I2_{i,j}\right)\right)\right|\right)\right) \tag{9}$$

$$S_{2_{i,j}} = 1 - S_{1_{i,j}} \tag{10}$$

where $\nabla(\bullet)$ is the Laplacian operator, $|\bullet|$ is the absolute value function, $L(\bullet)$ is the Gaussian low-pass filter function, $sign(\bullet)$ is the sign function, and $\min(\bullet)$ is the minimum function. And $S(\bullet)$ is also $H \times W$ in size. The proposed method uses a low-pass filter in both source images and selects pixels with large gradients, which ensures that the normal texture is very rarely estimated incorrectly. In summary, our multi-scale dense feature-aware network via integrated attention can achieve the optimal intensity distribution and maintain rich texture information under the mutual constraints of intensity loss and gradient loss. The target features in infrared images are more evident, and the detailed textures in visible images are more noticeable.

3. Experiment and Analysis

3.1. Data Preparation and Baselines

To comprehensively evaluate our proposed method, we adopt a public multi-spectral road scene MSRS dataset, which contains 1444 pairs of high-quality aligned infrared and visible images with 640 × 480 pixels, mainly describing road scenes. We use RGB images to obtain visible images and Y-channel infrared images for experiments. In order to obtain more accurate experimental results, we use 1083 images as the training set and 361 images as the test set. In this study, we compare our method with nine state-of-the-art approaches; these include two traditional methods, namely GTF [46] and MST-SR [47], an Ae-based method, RFN-Nest [48], two GAN-based methods, including FusionGAN and GANMcC [49], and four CNN-based approaches (namely IFCNN [50], U2Fusion [51], SDNet [39], and SeAFusion [41]). Finally, we tested our algorithm using data captured by a DJI Mavic 2 drone with a 12-megapixel camera sensor. Due to the UAV images containing

unaligned infrared and visible images, we apply a specific image-matching algorithm called "A UAV Image Matching Algorithm Considering Log-Polar Description and Position Scale Distance Feature" [48]. This algorithm is utilized to correlate infrared and visible images of the same location to reduce registration errors in UAV images.

3.2. Evaluation Metric

It is necessary to use different types of indicators for the quantitative evaluation of infrared and visible image fusion algorithms in order to determine their performance more accurately. Based on two evaluation indicators of information theory, it includes entropy (EN) [52], mutual information (MI) [53], human visual perception based on visual information fidelity (VIF) [54], spatial frequency (SF) [55], standard deviation (SD) [56] of two evaluation indexes based on image features, sum of correlation differences (SCD) [57] of two evaluation indexes based on image quality and average gradient (AG) [58], and the source image and the generated image evaluation index edge retention $Q^{AB/F}$ [59]. The evaluation indicators are described as follows.

(1) *Entropy*, *EN* is the average amount of information contained in each received message, also known as information entropy, source entropy, and average self-information. The index can only be used to reflect the information carried by the fusion image.

$$H(A) = -\sum_{a} P_A(a) \log p_A(a) \tag{11}$$

where a represents gray value, and $p_A(a)$ represents gray probability distribution. It is believed that fusion will be more effective if the *EN* is large, as it indicates that the image contains more information.

(2) *Mutual Information*, *MI*, represents the amount of information that can be extracted from a source image. The fusion effect is better with a higher *MI* value because the source images contain more information. *MI* is calculated according to the joint information entropy $H(A, B)$ and the information entropy $H(A)$ and $H(B)$ of the image:

$$MI(A, B) = H(A) + H(B) - H(A, B) \tag{12}$$

(3) *Visual Information Fidelity*, *VIF*, refers to a measurement method based on visual information fidelity, which is used to measure the quality of fused images. As the *VIF* value increases, the better the visual effect people will have on the fused image.

(4) *Spatial Frequency*, *SF*, is calculated by row frequency and column frequency to measure the spatial frequency information contained in the fusion image. Spatial frequency increases with the sharpness of the image. The formula for its calculation is as follows:

$$SF = \sqrt{RF^2 + CF^2} \tag{13}$$

$$CF = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} |H(i, j) - H(i, j-1)|^2} \tag{14}$$

$$RF = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} |H(i, j) - H(i, j-1)|^2} \tag{15}$$

(5) *Standard Deviation*, *SD*, represents how much an image's pixel value has changed relative to its average.

$$S_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{N}} \tag{16}$$

where x_i represents each individual data point in the dataset, $\overline{x}$ represents the average value of the dataset, and N represents the total number of data points in the dataset.

(6) *Gradient based Fusion Performance*, $Q^{AB/F}$, is a new objective non-reference quality evaluation method for fused images. The algorithm for obtaining $Q^{AB/F}$ uses local metrics to estimate the degree of representation of input important information in the fused image.

$$Q^{AB/F} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} Q^{AF}(n,m)w^{A}(n,m) + Q^{BF}(n,m)w^{B}(n,m)}{\sum_{i=1}^{N}\sum_{j=1}^{M}(w^{A}(i,j) + w^{B}(i,j))} \tag{17}$$

where Q^{AF} and Q^{BF} represent the residual value of the edge, n and m represent the intensity and orientation retention of the image edge, and $w^{A}(i,j)$ and $w^{B}(i,j)$ are the gradient intensity of the source image A and B, respectively.

(7) *Average Gradient*, *AG*, refers to the sharpness of an image and its ability to express information. A larger average gradient will result in a sharper image and a better fusion result, as indicated by this theory. Here is the calculation formula:

$$AG = \frac{1}{(M-1)(N-1)}\sum_{i=1}^{M-1}\sum_{i=1}^{N-1}\sqrt{\frac{\left(H(i+1,j) - H(i,j+1)^2 - H(i,j)^2\right)}{2}} \tag{18}$$

where H represents the fused image, and M and N represent the height and width of the image, respectively.

(8) *Sum of Correlation Differences*, *SCD*, measures the quality of images in image fusion. Based on this method, differential images are calculated using the source image and the fused image, and their correlation is evaluated. Rather than directly evaluating the correlation between the source image and the fused image, it calculates the quality of the fused image by considering the source image and its effects.

3.3. Experimental Result

Since traditional methods do not support the training and fusion of RGB images, we first convert visible images into YCbCr color space and then use different methods to merge infrared–visible images into Y channel and then convert them into RGB images. Finally, the visualization results are clearer. Since only visible images and PET images contain color information, the fused Y channel is mapped back into the RGB color space along with the Cb and Cr (chromaticity) channels of the visible image (or PET image). Typically, Cb and Cr are combined as follows for the fusion of visible and infrared images:

$$C_f = \frac{C_1(|C_1 - \tau|) + C_2(|C_2 - \tau|)}{|C_1 - \tau| + |C_2 - \tau|} \tag{19}$$

where C_1 and C_2 are Cr and Cb channels of the first image and the second source image, respectively, C_f is the result of the fusion channel, and τ is generally set to 128. Afterward, the fused image is converted back into RGB by performing a reverse transformation.

3.3.1. Visual Performance

We compare five low-light condition images on MSRS with nine different fusion methods, such as SeAFusion, GANMcC, MST-SR, and FusionGAN, in order to verify the benefits of the proposed DFA-Net algorithm. The fusion results are shown in Figure 4. In infrared–visible image fusion, infrared sensors can generate infrared images by reflecting or capturing the thermal radiation emitted by objects, making the target more prominent in the background. However, infrared images ignore the texture and are easily affected by

noise. In contrast, visible sensors capture reflected light information. It is often the case that visible images contain rich texture and structure information, but they are susceptible to light and weather-related degradation. Infrared–visible images can effectively synthesize the feature information of the target and the detailed texture information of visible, and obtain the fused image with more comprehensive information. In Figure 5, FusionGAN and GANMcC are not obvious in the infrared image because the single discriminator can easily cause modal imbalance in the training process. However, the thermal radiation tree information of GTF, RFN-Nest, and U2Fusion is not displayed. Compared with these methods, the edge structure of this paper is more apparent, and the texture details are more apparent, which indicates that enhanced channel feature extraction can retain the features of the source image better by integrating the integrated attention fusion module and achieving infrared–visible image fusion through enhanced channel feature extraction.

Figure 5 shows the fusion results in the fuzzy scene under low-light conditions. The visible image is a little fuzzy, and the target of the infrared image is prominent. GTF, FusionGAN, SDNet, U2Fusion, and GANMcC cannot retain visible texture information. Compared with the visible image, the car license plate becomes lighter, indicating that the visible detail texture is missing. Our method has produced fusion images in low-light environments that show clearly defined characters, evident thermal radiation features of the three individuals, clear images, and car license plates, all of which retain complementary information from both infrared and visible images.

As shown in Figure 6 below, the number of figures cannot be distinguished only through visible images, while the infrared image can recognize the outline of figures and obtain the number of figures under the condition of low light. The other nine comparison methods have different degrees of information loss on the whole, while the method in this paper has more comprehensive feature information, the light is clearer and brighter, and the information of the vehicle behind is also obvious.

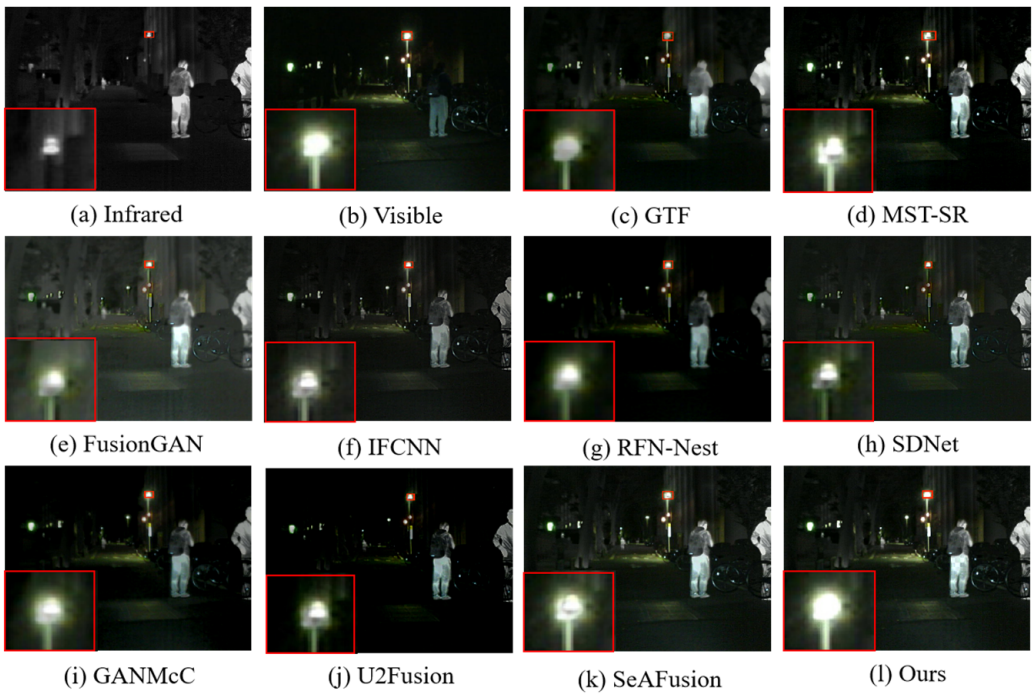

Figure 4. Qualitative comparison of DFA-Net with nine state-of-the-art methods on 01012N image from the MSRS dataset. For a clear comparison, we select a small area with abundant texture in each image and zoom in on it in the bottom right corner and highlight a salient region, as shown in the red box. (**a**) Infrared. (**b**) Visible. (**c**) GTF. (**d**) MST-SR. (**e**) FusionGAN. (**f**) IFCNN. (**g**) RFN-Nest. (**h**) SDNet. (**i**) GANMcC. (**j**) U2Fusion. (**k**) SeAFusion. (**l**) Ours.

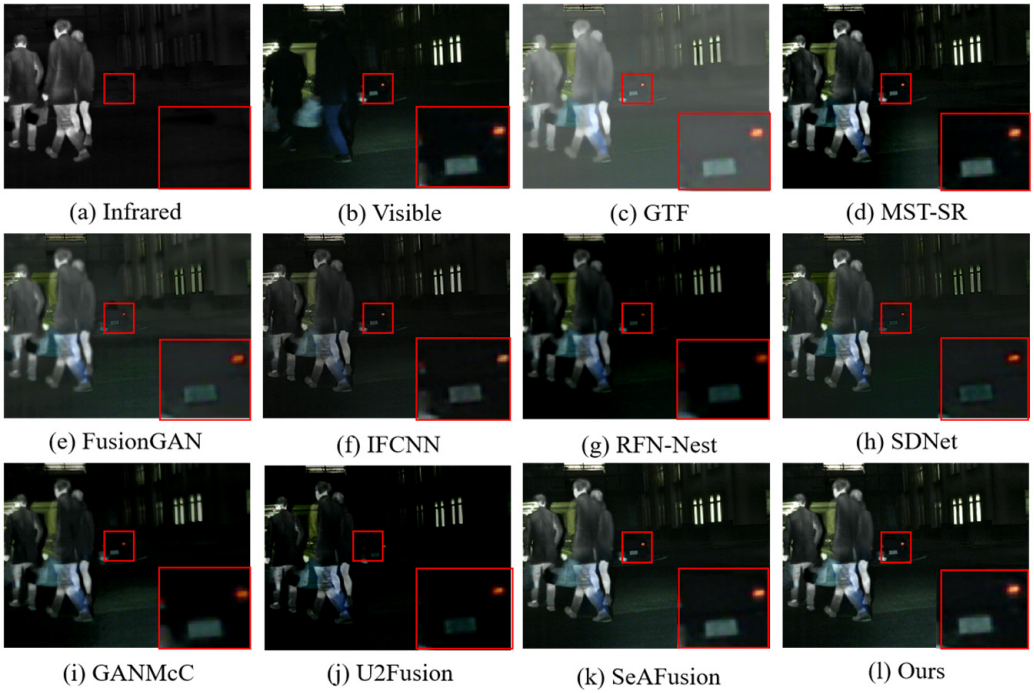

Figure 5. Qualitative comparison of DFA-Net with nine state-of-the-art methods on 00681N image from the MSRS dataset. For a clear comparison, we select a small area with abundant texture in each image and zoom in on the area in the bottom right corner and highlight a salient region, as shown in the red box. (**a**) Infrared. (**b**) Visible. (**c**) GTF. (**d**) MST-SR. (**e**) FusionGAN. (**f**) IFCNN. (**g**) RFN-Nest. (**h**) SDNet. (**i**) GANMcC. (**j**) U2Fusion. (**k**) SeAFusion. (**l**) Ours.

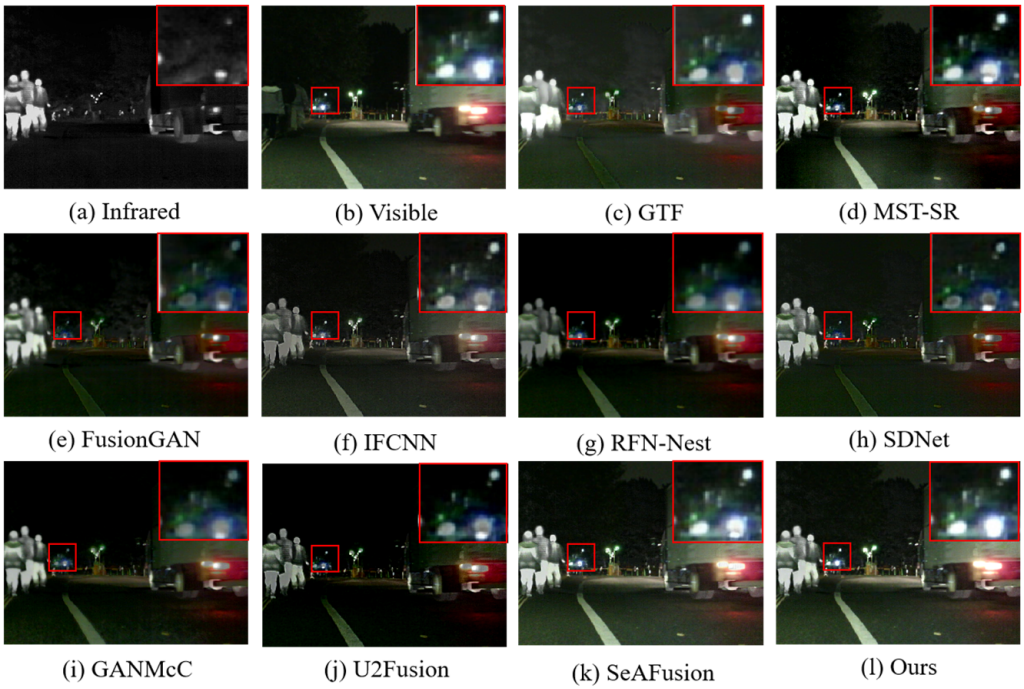

Figure 6. Qualitative comparison of DFA-Net with nine state-of-the-art methods on 00890N image from the MSRS dataset. For a clear comparison, we select a small area with abundant texture in each image and zoom in on the area in the bottom right corner and highlight a salient region, as shown in the red box. (**a**) Infrared. (**b**) Visible. (**c**) GTF. (**d**) MST-SR. (**e**) FusionGAN. (**f**) IFCNN. (**g**) RFN-Nest. (**h**) SDNet. (**i**) GANMcC. (**j**) U2Fusion. (**k**) SeAFusion. (**l**) Ours.

In low-light environments, vehicles will inevitably have lights of different colors at night. As shown in Figure 7, visible contains the yellow lights of street lamps and the red lights of vehicles. At this time, the fusion image needs bright lights and complete vehicle outlines. According to the images, it can be clearly seen that the depth features of GTF and IFCNN are a little fuzzy compared with the source image due to the fusion rules designed by the traditional manual but the image produced using our method is not. In the images not produced using our method, the brightness of the vehicle's taillight is dimmed to varying degrees.

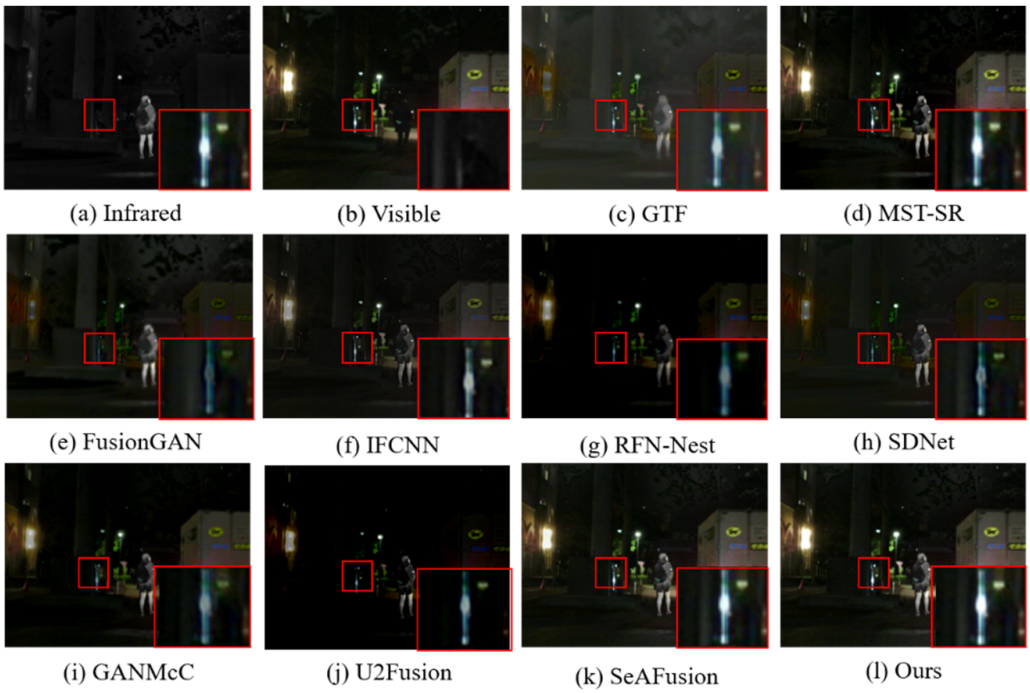

Figure 7. Qualitative comparison of DFA-Net with nine state-of-the-art methods on 00714N image from the MSRS dataset. For a clear comparison, we select a small area with abundant texture in each image and zoom in on the area in the bottom right corner and highlight a salient region, as shown in the red box. (**a**) Infrared. (**b**) Visible. (**c**) GTF. (**d**) MST-SR. (**e**) FusionGAN. (**f**) IFCNN. (**g**) RFN-Nest. (**h**) SDNet. (**i**) GANMcC. (**j**) U2Fusion. (**k**) SeAFusion. (**l**) Ours.

According to the comparative analysis of Figure 8, the infrared image contains less thermal radiation information, while the visible image contains more abundant content. The images of GTF, SDNet, and FusionGAN are somewhat blurred, and the above-ground implementation of FusionGAN is almost invisible, indicating that the image has poor edge retention ability. The brightness of the background light in U2Fusion, SDNet, and IFCNN is not high. The method we propose can effectively eliminate these defects. Under the condition of a low-light environment, the background light is bright and the white solid line on the ground is clear, and the fusion effect is better than in the other methods. In general, the DFA-Net algorithm provides certain advantages over other infrared image combining methods.

3.3.2. Quantitative Comparison

In order to evaluate the effectiveness of our experiment, we performed a quantitative comparison of the performance of our method against nine representative methods of image fusion on the MSRS dataset. The index results are shown in Table 1, which shows the mean values of the nine methods in the eight evaluation indexes. According to Table 1, among the eight indicators, seven indicators of DFA-Net are better than other indicators, and MI ranks second after SeAFusion.

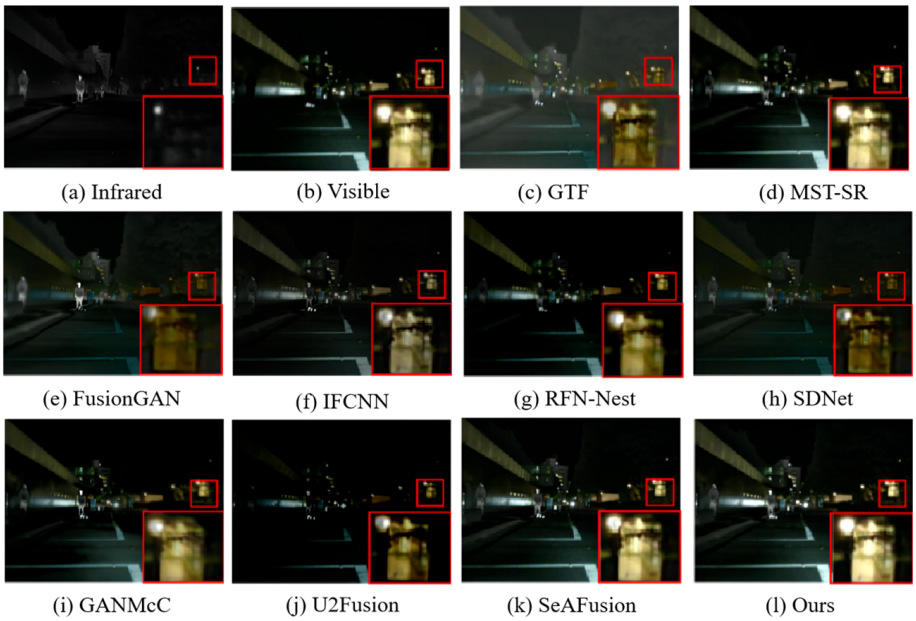

Figure 8. Qualitative comparison of DFA-Net with nine state-of-the-art methods on 01254N image from the MSRS dataset. For a clear comparison, we select a small area with abundant texture in each image and zoom in on the area in the bottom right corner and highlight a salient region, as shown in the red box. (**a**) Infrared. (**b**) Visible. (**c**) GTF. (**d**) MST-SR. (**e**) FusionGAN. (**f**) IFCNN. (**g**) RFN-Nest. (**h**) SDNet. (**i**) GANMcC. (**j**) U2Fusion. (**k**) SeAFusion. (**l**) Ours.

Table 1. Quantitative fusion results of 361 sets of infrared–visible images in 9 methods (red represents the best results; blue represents the second-best results).

	EN	**SF**	**SD**	**MI**	**VIF**	**AG**	**SCD**	$\mathbf{Q^{AB/F}}$
MST-SR	6.274500	0.042881	8.1	2.615585	0.805536	3.379710	1.302535	0.528150
GTF	5.488868	0.031165	6.3	1.703488	0.472943	2.369413	0.711759	0.401289
FusionGAN	5.549333	0.019308	6.3	1.879273	0.595634	1.642359	1.065234	0.159680
U2Fusion	4.819441	0.039653	6.5	1.813905	0.547874	2.976654	1.334482	0.390667
IFCNN	6.031980	0.039803	7.4	2.330893	0.694640	3.169491	1.291991	0.543784
RFN-Nest	5.586484	0.027077	7.7	2.183343	0.680637	2.264621	1.546296	0.392791
SDNet	5.428818	0.037694	6.1	1.754339	0.484199	2.982137	1.122136	0.406097
GANMcC	5.877195	0.024687	8.4	2.423083	0.694944	2.141018	1.459269	0.312746
SeAFusion	6.651394	0.043554	8.4	4.037259	0.985942	3.696791	1.685249	0.662335
Ours	6.741801	0.048172	8.5	3.75237	1.041419	3.837578	1.718825	0.689588

According to the data analysis in Table 1, DFA-Net has a good performance in the MSRS dataset. First of all, a higher VIF indicates a better visual effect, while an increase in EN and SF indicates more detail in the image. The fusion of the infrared image and the visible image is information complementary, so the preservation of background information and the fidelity of background information are equally important. Maximum $Q^{AB/F}$ indicates that more edge information has been transferred from the source image to the fused image. Additionally, the higher the level of retained gradient, the greater the SD, AG, and SCD, the better the image quality. Compared to other studies, the results of this paper have the highest quality, the characteristics of infrared images are more obvious, and the details and textures of visible images are more discernible.

To make the improvement effect of the evaluation index more evident, we also selected 100 pairs of low-light environment image pairs from the MSRS dataset for quantitative evaluation and drew the cumulative distribution map. Figure 9 shows the results of the comparison of these eight indicators by different methods. As can be seen from Figure 9, DFA-Net has obvious advantages in the indexes of EN, AG, SD, VIF, SF, $Q^{AB/F}$, and SCD

in the road scene dataset. This means that our fused images are more richly textured at the edges and have better visuals and better image quality. In addition, DFA-Net ranks second in MI, indicating that DFA-Net contains more mutual information and richer image information.

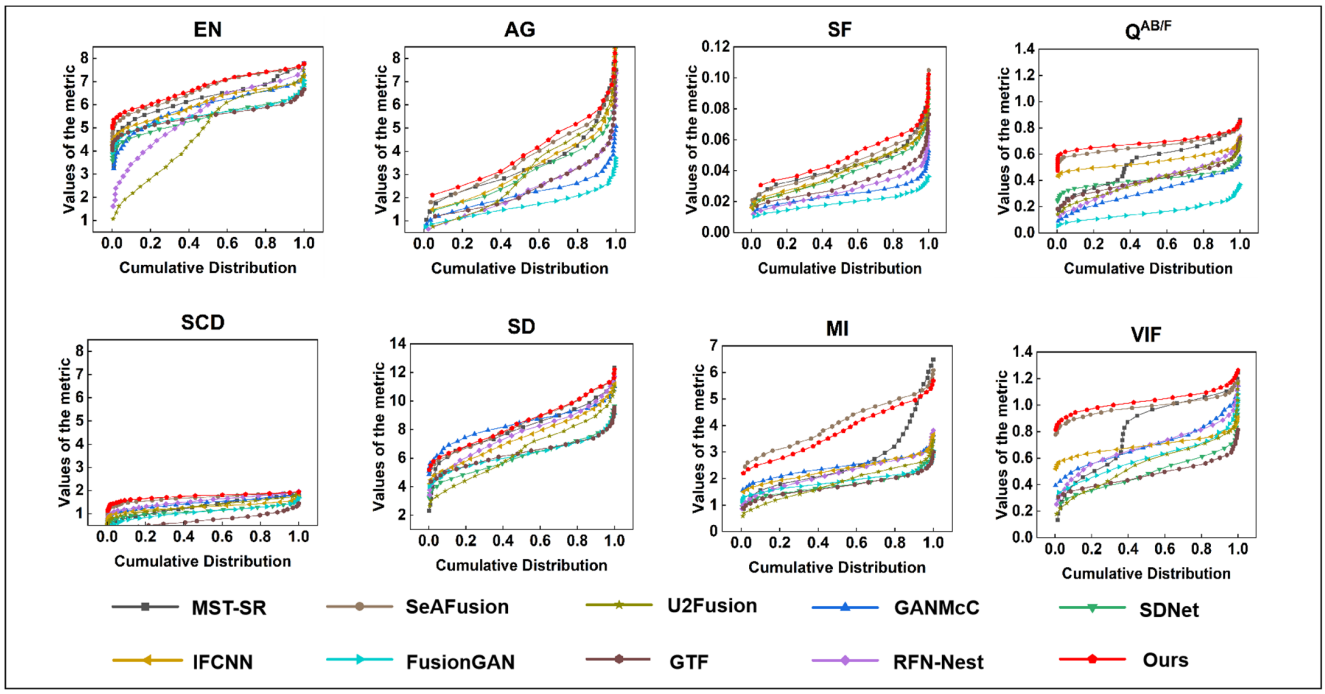

Figure 9. Quantitative comparison of 100 pairs of infrared–visible image fusion methods in MSRS dataset.

3.4. Ablation Study

We designed an intensity loss constraint fusion image to retain more useful pixel information. More specifically, we adjusted the epoch in order to change the proportion of intensity information from different images so that it can be applied to a variety of image fusion tasks. Therefore, we investigated the impact of different epochs on the model performance. The results are presented in Table 2.

Table 2. Computational results of MRSR dataset with different epochs (red represents the best results).

	EN	SF	SD	MI	VIF	AG	SCD	$Q^{AB/F}$
1	6.656426	0.045487	8.4	4.170775	0.988418	3.646758	1.621753	0.669267
5	6.705488	0.046563	8.5	3.844076	1.030301	3.758284	1.663681	0.690490
10	6.741384	0.048109	8.5	3.723383	1.044102	3.831187	1.726796	0.690936
15	6.773159	0.048611	8.5	3.764401	1.051541	3.831952	1.727661	0.691273
20	6.797420	0.049807	8.6	3.642573	1.060563	3.905604	1.751419	0.690326
25	6.788444	0.049672	8.5	3.588535	1.045185	3.887107	1.752573	0.691391
30	6.788444	0.049672	8.5	3.588535	1.045185	3.887107	1.752573	0.691391

In summary, a large number of qualitative and quantitative results show that the proposed DFA-Net algorithm has good robustness and good performance in various indexes under the conditions of efficiency, effect and multi-scale, etc. The reasons are summarized as follows: Firstly, our network structure adopts a double-branch integrated nested network. Secondly, we add the integrated attention fusion module to enhance the channel feature extraction and improve the quality of the fused image. Thirdly, we define an intensity loss, which constrains the overall apparent intensity of the fused image to

integrate useful image information more effectively. Finally, we use dense jump connections to enhance the description of detailed grains.

To prove that our integrated attention is effective, we tested the effect after training without adding integrated attention. In Table 3, we can see that there is a great improvement in all the indicators after adding integrated attention.

Table 3. Ablation results of MRSR dataset (mean and variance).

Metric	DFA-Net(-IAF)	DFA-Net(+IAF)
EN	6.01 ± 0.43	6.74 ± 0.45
SF	0.051385 ± 0.000089	0.048172 ± 0.000184
SD	7.4 ± 2.1	8.5 ± 2.7
MI	2.75 ± 0.47	3.75 ± 0.86
VIF	0.643 ± 0.01	1.0414 ± 0.0057
AG	2.18 ± 0.34	3.8 ± 2.0
SCD	1.294 ± 0.040	1.719 ± 0.018
$Q^{AB/F}$	0.4052 ± 0.0038	0.6896 ± 0.0025

3.5. Generalization Analysis

The MSRS dataset contains aligned infrared and visible images, so we can conduct discriminant analysis to compare performance across different samples in this article. In order to demonstrate the excellent generalization capability of DFA-Net under low-light conditions, we collect some infrared and visible images using the DJI drone pre-2.

Based on our UAV dataset, Figures 10 and 11 illustrate a comparison of infrared and visible image fusion results. Comparison of the fused image with the visible image demonstrates that the thermal radiation obtained by our method highlights the feature target with respect to the visual effects. Additionally, the background information in the fusion image is more accurate than in the infrared image, improving the image quality. The difference between the UAV image and the general image is that the angle of view is different. As shown in Figure 10, even in a low-light environment, the target features in the fused image are more prominent, and the influence of shadow on the target features is significantly reduced. It is apparent in Figure 11 that the texture information of the fused UAV image is richer than that of the visible image.

(a) Infrared (b) Visible (c) Ours

Figure 10. Some typical examples of UAV infrared and visible image fusion results on DFA-Net.

(a) Infrared (b) Visible (c) Ours

Figure 11. Some typical examples of UAV infrared and visible image fusion results on DFA-Net.

4. Discussion and Conclusions

The current methods for infrared–visible image fusion still face certain challenges. Extracting detailed texture features from visible light and background features from infrared under low-light conditions is difficult, and comprehensively extracting information from the source images is also challenging. We have developed a novel multi-scale dense feature-aware network with integrated attention (DFA-Net), capable of extracting target features from infrared images and detail texture features from visible images. Our network utilizes multi-scale nesting methods and has achieved excellent results in infrared and visible image fusion. Additionally, we have introduced an integrated attention module to enhance complementary features in both infrared and visible images, emphasizing the retention of richer detail information and salient features during the fusion process. To further improve the quality of the fused images, we have combined intensity and gradient loss for refining the multi-source information, resulting in high-quality infrared and visible fused images.

In addition, through qualitative and quantitative analysis, the algorithm designed in this paper demonstrates certain advantages compared to the other nine algorithms. The fusion of image features is more prominent, and the background texture information is richer. In particular, the performance in unmanned aerial vehicles (UAVs) is outstanding, effectively enhancing the complementarity of UAV images and improving UAV environment perception capabilities.

Author Contributions: Conceptualization, Y.W., L.M., S.S. and D.L.; methodology, S.S.; software, Z.Y.; validation, L.M., Y.W. and D.L.; formal analysis, W.Y.; investigation, C.X.; data curation, S.S.; writing—original draft preparation, C.X.; writing—review and editing, Q.Z.; visualization, Y.W.; supervision, D.L., W.Y. and Z.Y.; project administration, B.H.; funding acquisition, S.S. and W.Y. All authors have read and agreed to the published version of this manuscript.

Funding: This research is funded by the Scientific Research Foundation for Doctoral Program of Hubei University of Technology (BSQD2020056); Natural Science Foundation of Hubei Province (2022CFB501): University Student innovation and Entrepreneurship Training Program Project (202210500028).

Data Availability Statement: Not applicable.

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

References

1. Ehlers, M. Multisensor image fusion techniques in remote sensing. *ISPRS J. Photogramm. Remote Sens.* **1991**, *46*, 19–30. [CrossRef]
2. Liu, Y.; Liu, S.; Wang, Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* **2015**, *24*, 147–164. [CrossRef]
3. Burt, P.J. A gradient pyramid basis for pattern-selective image fusion. *Proc. SID* **1992**, *23*, 467–470.
4. Yang, Y.; Que, Y.; Huang, S.; Lin, P. Multimodal sensor medical image fusion based on type-2 fuzzy logic in NSCT domain. *IEEE Sens. J.* **2016**, *16*, 3735–3745.
5. Jin, B.; Cruz, L.; Gonçalves, N. Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis. *IEEE Access* **2020**, *8*, 123649–123661. [CrossRef]
6. Zheng, Q.; Zhao, P.; Li, Y.; Wang, H.; Yang, Y. Spectrum interference-based two-level data augmentation method in deep learning for automatic modulation classification. *Neural Comput. Appl.* **2021**, *33*, 7723–7745. [CrossRef]
7. Li, B.; Li, Q.; Zeng, Y.; Rong, Y.; Zhang, R. 3D trajectory optimization for energy-efficient UAV communication: A control design perspective. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 4579–4593. [CrossRef]
8. Raza, A.; Liu, J.; Liu, Y.; Liu, J.; Li, Z.; Chen, X.; Huo, H.; Fang, T. IR-MSDNet: Infrared and visible image fusion based on infrared features and multiscale dense network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3426–3437. [CrossRef]
9. Mei, L.; Yu, Y.; Shen, H.; Weng, Y.; Liu, Y.; Wang, D.; Liu, S.; Zhou, F.; Lei, C. Adversarial multiscale feature learning framework for overlapping chromosome segmentation. *Entropy* **2022**, *24*, 522. [CrossRef] [PubMed]
10. 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]
11. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
12. Zhou, Z.; Wang, B.; Li, S.; Dong, M. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Inf. Fusion* **2016**, *30*, 15–26.
13. Zhang, Z.; Blum, R.S. A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proc. IEEE* **1999**, *87*, 1315–1326.
14. Cui, G.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.* **2015**, *341*, 199–209.
15. Zhao, J.; Chen, Y.; Feng, H.; Xu, Z.; Li, Q. Infrared image enhancement through saliency feature analysis based on multi-scale decomposition. *Infrared Phys. Technol.* **2014**, *62*, 86–93.
16. Mei, L.; Guo, X.; Huang, X.; Weng, Y.; Liu, S.; Lei, C. Dense contour-imbalance aware framework for colon gland instance segmentation. *Biomed. Signal Process. Control.* **2020**, *60*, 101988.
17. Yang, B.; Li, S. Multifocus image fusion and restoration with sparse representation. *IEEE Trans. Instrum. Meas.* **2009**, *59*, 884–892. [CrossRef]
18. Li, S.; Yin, H.; Fang, L. Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 3450–3459. [CrossRef]
19. Wang, J.; Peng, J.; Feng, X.; He, G.; Fan, J. Fusion method for infrared and visible images by using non-negative sparse representation. *Infrared Phys. Technol.* **2014**, *67*, 477–489.
20. Li, H.; He, X.; Tao, D.; Tang, Y.; Wang, R. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognit.* **2018**, *79*, 130–146. [CrossRef]
21. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17. [CrossRef]
22. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
23. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans. Multimed.* **2020**, *23*, 1383–1396.
24. Chen, J.; Wu, K.; Cheng, Z.; Luo, L. A saliency-based multiscale approach for infrared and visible image fusion. *Signal Process.* **2021**, *182*, 107936.
25. Zhao, J.; Zhou, Q.; Chen, Y.; Feng, H.; Xu, Z.; Li, Q. Fusion of visible and infrared images using saliency analysis and detail preserving based image decomposition. *Infrared Phys. Technol.* **2013**, *56*, 93–99. [CrossRef]
26. Guo, X.; Meng, L.; Mei, L.; Weng, Y.; Tong, H. Multi-focus image fusion with Siamese self-attention network. *IET Image Process.* **2020**, *14*, 1339–1346. [CrossRef]
27. Kumar, S.S.; Muttan, S. *PCA-Based Image Fusion, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII, 2006*; SPIE: Bellingham, WA, USA, 2006; pp. 658–665.
28. Li, H.; Ding, W.; Cao, X.; Liu, C. Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing. *Remote Sens.* **2017**, *9*, 441. [CrossRef]

29. Pu, Q.; Chehri, A.; Jeon, G.; Zhang, L.; Yang, X. DCFusion: Dual-Headed Fusion Strategy and Contextual Information Awareness for Infrared and Visible Remote sensing Image. *Remote Sens.* **2023**, *15*, 144. [CrossRef]
30. He, G.; Ji, J.; Dong, D.; Wang, J.; Fan, J. Infrared and visible image fusion method by using hybrid representation learning. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1796–1800. [CrossRef]
31. Zhu, D.; Zhan, W.; Jiang, Y.; Xu, X.; Guo, R. MIFFuse: A multi-level feature fusion network for infrared and visible images. *IEEE Access* **2021**, *9*, 130778–130792. [CrossRef]
32. Ma, Y.; Chen, J.; Chen, C.; Fan, F.; Ma, J. Infrared and visible image fusion using total variation model. *Neurocomputing* **2016**, *202*, 12–19. [CrossRef]
33. Li, S.; Yang, B.; Hu, J. Performance comparison of different multi-resolution transforms for image fusion. *Inf. Fusion* **2011**, *12*, 74–84. [CrossRef]
34. Kong, W.; Lei, Y.; Zhao, H. Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. *Infrared Phys. Technol* **2014**, *67*, 161–172.
35. Bavirisetti, D.P.; Dhuli, R. Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform. *IEEE Sens. J.* **2015**, *16*, 203–209. [CrossRef]
36. Yin, M.; Duan, P.; Liu, W.; Liang, X. A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation. *Neurocomputing* **2017**, *226*, 182–191. [CrossRef]
37. Jin, X.; Jiang, Q.; Yao, S.; Zhou, D.; Nie, R.; Lee, S.-J.; He, K. Infrared and visual image fusion method based on discrete cosine transform and local spatial frequency in discrete stationary wavelet transform domain. *Infrared Phys. Technol.* **2018**, *88*, 1–12.
38. 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]
39. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *IJCV* **2021**, *129*, 2761–2785. [CrossRef]
40. 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]
41. 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.
42. 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]
43. Li, B.; Zhang, M.; Rong, Y.; Han, Z. Transceiver optimization for wireless powered time-division duplex MU-MIMO systems: Non-robust and robust designs. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 4594–4607. [CrossRef]
44. Cheng, D.; Chen, L.; Lv, C.; Guo, L.; Kou, Q. Light-Guided and Cross-Fusion U-Net for Anti-Illumination Image Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8436–8449. [CrossRef]
45. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
46. Ma, J.; Chen, C.; Li, C.; Huang, J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion* **2016**, *31*, 100–109.
47. Yu, R.; Chen, W.; Zhou, D. Infrared and visible image fusion based on gradient transfer optimization model. *IEEE Access* **2020**, *8*, 50091–50106. [CrossRef]
48. 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]
49. 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]
50. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118.
51. 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]
52. Bein, B. Entropy. *Best Pract. Res. Clin. Anaesthesiol.* **2006**, *20*, 101–109.
53. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev.* **2004**, *69*, 066138. [CrossRef] [PubMed]
54. Sheikh, H.R.; Bovik, A.C. A Visual Information Fidelity Approach to Video Quality Assessment. In *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*; 2005; pp. 2117–2128. Available online: https://www.semanticscholar.org/paper/A-VISUAL-INFORMATION-FIDELITY-APPROACH-TO-VIDEO-Bovik/b70b6cf13b55b61a37133b921770dcf32ef0bcfd (accessed on 27 June 2023).
55. Shapley, R.; Lennie, P. Spatial frequency analysis in the visual system. *Annu. Rev. Neurosci.* **1985**, *8*, 547–581. [CrossRef] [PubMed]
56. Altman, D.G.; Bland, J.M. Standard deviations and standard errors. *BMJ* **2005**, *331*, 903. [CrossRef]
57. Hisham, M.; Yaakob, S.N.; Raof, R.; Nazren, A.A.; Wafi, N. Template matching using sum of squared difference and normalized cross correlation. In Proceedings of the 2015 IEEE Student Conference on Research and Development (SCOReD), Kuala Lumpur, Malaysia, 13–14 December 2015; pp. 100–104.

58. Schmidt, M.; Le Roux, N.; Bach, F. Minimizing finite sums with the stochastic average gradient. *Math. Program.* **2017**, *162*, 83–112. [CrossRef]
59. Xydeas, C.S.; Petrovic, V. Objective image fusion performance measure. *Electron. Lett.* **2000**, *36*, 308–309. [CrossRef]

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.