mathematics

Article

OMOFuse: An Optimized Dual-Attention Mechanism Model for Infrared and Visible Image Fusion

Jianye Yuan and Song Li *

Electronic Information School, Wuhan University, Wuhan 473072, China; yuan666@whu.edu.cn
* Correspondence: ls@whu.edu.cn

Abstract: Infrared and visible image fusion aims to fuse the thermal information of infrared images and the texture information of visible images into images that are more in compliance with people's visual perception characteristics. However, in the existing related work, the fused images have incomplete contextual information and poor fusion results. This paper presents a new image fusion algorithm—OMOFuse. At first, both the channel and spatial attention mechanisms are optimized by a DCA (dual-channel attention) mechanism and an ESA (enhanced spatial attention) mechanism. Then, an ODAM (optimized dual-attention mechanism) module is constructed to further improve the integration effect. Moreover, a MO module is used to improve the network's feature extraction capability for contextual information. Finally, there is the loss function $\mathcal{L}$ from the three parts of SSL (structural similarity loss), PL (perceptual loss), and GL (gap loss). Extensive experiments on three major datasets are performed to demonstrate that OMOFuse outperforms the existing image fusion methods in terms of quantitative determination, qualitative detection, and superior generalization capabilities. Further evidence of the effectiveness of our algorithm in this study are provided.

Keywords: image fusion; attention mechanism; loss function; quantitative and qualitative; infrared and visible images

MSC: 68T07

check for updates

Citation: Yuan, J.; Li, S. OMOFuse: An Optimized Dual-Attention Mechanism Model for Infrared and Visible Image Fusion. *Mathematics* **2023**, *11*, 4902. https://doi.org/10.3390/math11244902

Academic Editor: Teng Li

Received: 15 November 2023
Revised: 29 November 2023
Accepted: 29 November 2023
Published: 7 December 2023

1. Introduction

It is known that infrared sensors exhibit a superior capability to capture heat radiation targets, demonstrating robust anti-interference properties when detecting hidden or camouflaged objects. However, they display a less powerful capacity for capturing texture information and structural features [1]. Conversely, visible light sensors excel in capturing background information, texture details, and structural features of the target. However, they are incapable of highlighting the target and are readily susceptible to environmental (weather-related) changes [2]. Taking the strengths and weaknesses of infrared and visible light sensors into account, it is common to use image fusion algorithms to combine their advantages. This approach leverages their individual strengths, leading to better overall results. Images melded together often align more closely with human visual perception, thereby finding applications in related sectors such as image segmentation [3], biometric recognition [4], object detection [5], and object tracking [6]. Concurrently, the fusion of images can also provide increased data support for domains like video surveillance [7], military [8], and medical [9].

Many traditional image fusion methods, such as hybrid methods [10], sparse representation [11], and multi-scale transformations [12], are mainly conducted by designing corresponding fusion strategies and reconstructing the ultimate fusion results through inverse operations. For instance, a multi-scale method with Gaussian and bilateral filters effectively preserved the perception information of the visible light images [13]. Similarly, the guided filtering and weighted averaging techniques were used to incorporate spatial consistency into the base and detailed layers for enhancing the fusion effects [14]. The literature [15] employed guided image filtering and a weight map to integrate complementary information per pixel, which improved the rapid fusion effect with a multi-scale guide. Infrared image sensors show superior target detection performance under various weather conditions, whereas visible light image sensors excel at extracting strong texture features. However, traditional methods often overlook the modulation differentiation of images and the inherent characteristics of infrared and visible light. This particularity often leads them to use either similar representation models or uniform transformation for feature extraction from source images, which can result in a weaker fusion performance (noise, artifact). At the same time, the fusion models employed by traditional methods largely fall into the manual design [16], usually accompanied by very high complexity, diminishing their practical applications and research value.

The superior non-linearity fitting and feature extraction capabilities of deep learning have won the favor of many research scholars. In the task of image fusion, the existing deep learning methods mainly solve the issues of image feature extraction, feature fusion, and image reconstruction. In the literature [1], a novel non-linear image fusion variational model was proposed, which employed a minimal number of parameters to encode a priori information about the structure of the fused image, tackling the multi-modal fusion issues. The literature [17] addressed the problem of non-Gaussian noise image restoration degradation and proposed a novel CNFR image restoration algorithm. The experimental simulation results indicated that this method remarkably reduced the loss of contrast. In the literature [18], a higher-level image decomposition scheme and an advanced fusion rule were proposed. This method, utilizing a three-layer image decomposition and enhancement rule, performed best on grayscale computed tomography (CT) and magnetic resonance imaging (MRI) images. Although the above-mentioned work achieved appreciable results on image fusion through deep learning methods, there are still some issues to be resolved. Notably, the quality of the fused image depends not only on the texture features of local information but also on texture features of the global information of the entire image. These methods are conditioned by the convolution principle on the one hand, which prevents image information from being fully extracted by convolution features in all dimensions, like spatial dimensions and channel dimensions. On the other hand, the aforementioned methods are essentially based on local feature extraction. By extracting thermal infrared and texture features in a limited receptive field, they fail to extract long-distance information in the context.

To address these issues, we propose a method for image fusion, called OMOFuse (optimized dual-attention module; multi-head self-attention, optimized artificial neural network; optimized multi-layer perceptron). This approach introduces an innovative ODAM (optimized dual-attention mechanism) module, to accurately extract both infrared thermal radiation information and the texture information of visible light on spatial and channel dimensions. To further extract and establish long-distance dependencies, preserving global complementary contextual information, the MO (multi-head self-attention, optimized artificial neural network) module is introduced. Finally, a set of loss functions L from the perspectives of SSL (structural similarity loss) [19], PL (perceptual loss) [20], and GL (gradient loss) [21] is designed to further enhance the network fusion performance. As a result, the problem of high model complexity caused by manual design approaches is successfully avoided, significantly elevating the applicability and scientific value of this model.

To demonstrate the visual effects of our OMOFuse method, Figure 1 provides examples in comparison with the advanced image fusion algorithm SeAFusion [22] across three image fusion datasets. Upon manual visual inspection, it is observed that infrared light delivers a superior performance in detecting the thermal imagery of human bodies compared to visible light, especially during nighttime conditions. In identifying the "fire" scenario, the scope of recognition by SeAFusion is smaller compared to OMOFuse; in the detection of the "human body" image, SeAFusion performs on par with OMOFuse, although the degree of blurriness and transparency in the "human body, umbrella" portion is higher in the SeAFusion algorithm than in OMOFuse. In the third scenario, SeAFusion turns out more blurred with less sharpness than OMOFuse, resulting in less intensely captured image boundaries. Therefore, overall, OMOFuse surpasses SeAFusion with higher clarity, lower fuzziness, and a greater capacity to capture edge information, marking an overall enhanced performance.

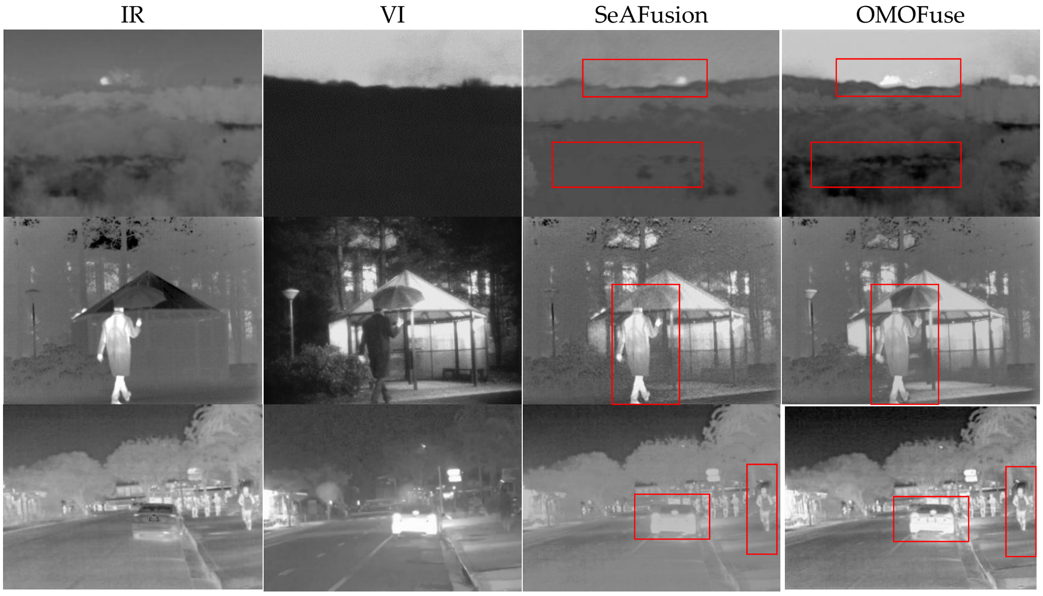

Figure 1. Comparison of effects between OMOFuse and SeAFusion.

Our contributions are listed as follows:

I. An end-to-end image fusion framework termed OMOFuse;

II. Two new mechanisms of attention are designed, that is, DCA (dual-channel attention) mechanism and ESA (enhanced spatial attention) mechanism. The DCA and ESA modules are combined to develop an ODAM module. This module efficiently gathers thermal radiation information from infrared channels and texture information from visible light, encompassing both the channel and spatial dimensions;

III. A MO module is constructed, which establishes long-distance dependency relationships and preserves comprehensive and global context information;

IV. SSL, PL, and GL perspectives are integrated to devise a new loss function, $\mathcal{L}$;

The rest of the paper is organized as follows. Section 2 summarizes some relevant research to the proposed method. Section 3 provides a detailed discussion of our OMOFuse. In Section 4, we present some qualitative and quantitative results on the image fusion tasks. In Section 5, we present the conducted extended experiments and reintroduce two public datasets to demonstrate the generalization ability of the algorithm in this paper. In Section 6, we present the performed ablation experiments on the OMOFuse algorithm to verify the effectiveness of each of our modules. Section 7 presents our concluding remarks.

2. Related Work

In recent years, rapid advancements in image processing technologies have intensified the need for improved fusion of images in infrared and visible light. Fusion methods can be classified according to the fused image level, which includes pixel-level, feature-level, and decision-level fusion; or the fusion strategy, encompassing traditional image fusion methods, basic deep learning-based approaches, methods based on the transformer [23] model, and others. Among these, deep learning-based methods have recently attracted significant attention. Deep learning-based image fusion frameworks can be further categorized into fusion methods based on AEs (auto-encoders) [24], fusion methods using CNN (convolutional neural networks), fusion methods relying on RNNs (recurrent neural networks) [25], and fusion methods utilizing GANs (generative adversarial networks) [26]. According to the application of deep learning methods, we will delve into the discussion of these particular image fusion methods.

2.1. Fusion Methods Based on AEs

The AE method of image fusion serves as a pioneer in contrast to other image fusion approaches by adopting AEs to extract features and reconstruct images from source images. Its primary focus is the manual design of fusion rules. The literature [27] provided a quintessential approach of employing AE methods for image fusion. This method incorporates a convolutional layer, fusion layer, and dense block, while also considering the limitations of AEs' feature extraction capability. In the same vein, the literature [28] introduced a novel retaining and feature enhancement loss function to guide the network for acquiring higher-dimensional detailed features. Moreover, the literature [29] proposed a NestFuse network model, further integrating nested connections in the network to allow the extraction of multi-scale feature information from the source images, thereby augmenting the fusion effectiveness. Despite these AE methods somewhat alleviating redundant information and resolution issues during the fusion process, there is still room for refinement. The literature [30] established a DRF fusion framework that decomposes source images into scene and attribute components, conducting their fusion separately to enhance the fusion effect. However, this method overlooks the principle of interpretability of fusion rules. Therefore, the literature [31] further proposed a novel learnable fusion rule, which amplifies the network's interpretability by assessing the significance of each pixel pair in feature images, thus improving the fusion effect.

2.2. Fusion Methods Based on CNN

In recent years, the rise of deep learning has become the mainstream in the task of image fusion, adroitly mitigating the shortcomings and complexity of manual rule-making. The literature [32] proposed the IFCNN framework, which makes use of two convolutional modules to extract salient features from source images in an end-to-end manner, manually adjusting fusion strategies to integrate more profound features. The literature [33] introduced a unified parameter and elastic weight framework, U2Fusion, which can be employed across various imaging tasks. The literature [34] incorporated a salient mask into the loss function to guide the network in identifying prominent targets in infrared images, simultaneously blending rich texture details into the visible light background, thus significantly addressing the issue of realism in fused images. The literature [32] put forth a symmetric feature-encoding network, SEDRFuse, harnessing feature compensation and attention mechanism fusion to amplify network performance. In summary, deep learning methods primarily rely on convolutional neural networks to extract local features of images and reconstruct information, mainly achieved by designing robust network module layers, such as residual blocks, multi-scale features, dense structures, attention mechanisms, etc. Compared to existing methods, we introduce a more powerful attention mechanism module to further enhance the algorithm's capability in global feature extraction and information reconstruction.

2.3. Fusion Methods Based on RNNs

RNNs hold a significant position within deep learning networks, demonstrating impressive performance in analyzing time series and sequences, especially with the emergence of ChatGPT, which has popularized image fusion as a commonly used combination tool [35]. RNNs have the ability to acquire information from previous inputs and integrate this information with input and output data, thus addressing issues related to sequence and time series. However, RNNs have their shortcomings; the standard RNN utilizes an invariant weight matrix that cannot be directly applied to images exhibiting spatial structural changes. Therefore, it has been proposed to use a weight map conditioned on the structure of the input image; the employed weight map and graphical representation are interrelated, and these visuals can effectively reveal essential internal structures of images [36]. The literature [37] proposed combining three CNNs and one RNN, aptly addressing the variability of spatial structure and a fuzzy process in RNNs. This framework, even with a relatively small network size, proposes a network that still possesses a large receptive field. The literature [38] suggested a trainable hybrid network to enhance similar degradation problems, assessed the global content of low-light inputs using an encoder–decoder network, and output a new spatially variational RNN as edge flow, exhibiting favorable image fusion performance. Among these methods, some researchers put forward a fusion framework based on content and detail branches. It uses content analysis for global information encoding while preserving most of the scene information. At the same time, it introduces a variational RNN to facilitate more accurate modeling of the internal structure of the source image. These two fusion strategies complement each other at the feature level.

2.4. Fusion Methods Based on GANs

Unsupervised methods excel in the evaluation of probability distribution tasks [39], prompting some scholars to apply GANs (generative adversarial networks) to image fusion. The FusionGAN framework, proposed for the first time in the literature [40], aptly mines texture information from images. However, its major drawback is a potential imbalance in the fusion process. To mitigate this issue, researchers in [36] utilized DDcGAN (dual discriminator conditioned GAN), eliminating the simple discernment of a visible light image and instead, incorporating two images for comparison simultaneously. Nonetheless, only the work of [41] has integrated multi-scale attention mechanics atop the DDcGAN framework. While preserving the foreground features of infrared images and background attributes of visible light images, this approach has introduced complexities in the training process. Consequently, [42] proposed a GANMcC (GAN framework with multi-class constraints). This strategy primarily addresses training difficulties, yet it also results in challenges of porting fusion results into subsequent visual tasks such as surveillance.

In summary, existing deep learning approaches in image fusion underscore the complementarity and feature balance of infrared and visible light information. Simultaneously, they neglect to a certain extent the extraction and transformation of information in various dimensions. Concerning long-distance dependency relationships and contextual information, including loss function features, these schemes have not achieved the desired results. Consequently, relevant higher-level visual tasks could be adversely affected. Thus, there is an urgent need for a method that can thoroughly extract features in various dimensions and network contexts.

3. Proposed Method

3.1. ODAM Module

Our ODAM algorithm module, as Figure 2 illustrates, predominantly comprises of a parallel connection between the DCA (dual-channel attention) module and the ESA (enhanced spatial attention) module. The input image, designated as $ODAM_f^{in}$, is initially divided into four branches $\{O^1, O^2, O^3, O^4\}$. Among these, O^2 undergoes processing through two CA (channel attention) mechanisms [43], linked in a series to form the aforementioned DCA. The main aim of the DCA process is to extract information efficiently from the channel dimensions of the O^2 feature maps, with the output being labeled as C^{mca} after completing the DCA processing. Meanwhile, another branch, O^4, proceeds sequentially through the SA (spatial attention) module [44], followed by an RC (residual connection) [45], and ultimately the MSF (multi-scale fusion) [46], culminating in an output of M^{esa}. The ESA module, composed of three sub-modules, namely, SA, RC, and MSF, further enhances the extraction of infrared and visible light information on the spatial dimension. All extracted infrared and visible light information are summed in terms of dimension to ultimately obtain O', which is then nonlinearized by the ReLU6 [47] activation function to yield the output value, denoted as $ODAM_f^{out}$. The crucial role of the ODAM module is to further enhance the network's infrared and visible light feature extraction capability on both the channel and spatial dimensions for the incoming image $ODAM_f^{in}$. Equation (1) shows the final output, represented as O^1, where $\oplus$ signifies the addition of channels. Equation (2) illustrates the final output of the ODAM module, denoted as $ODAM_f^{out}$.

$$O' = O^1 \oplus \left\{C^{mca} \oplus O^3\right\} \oplus \left\{M^{esa} \oplus O^3\right\} \tag{1}$$

$$ODAM_f^{out} = ReLU6\_O' \tag{2}$$

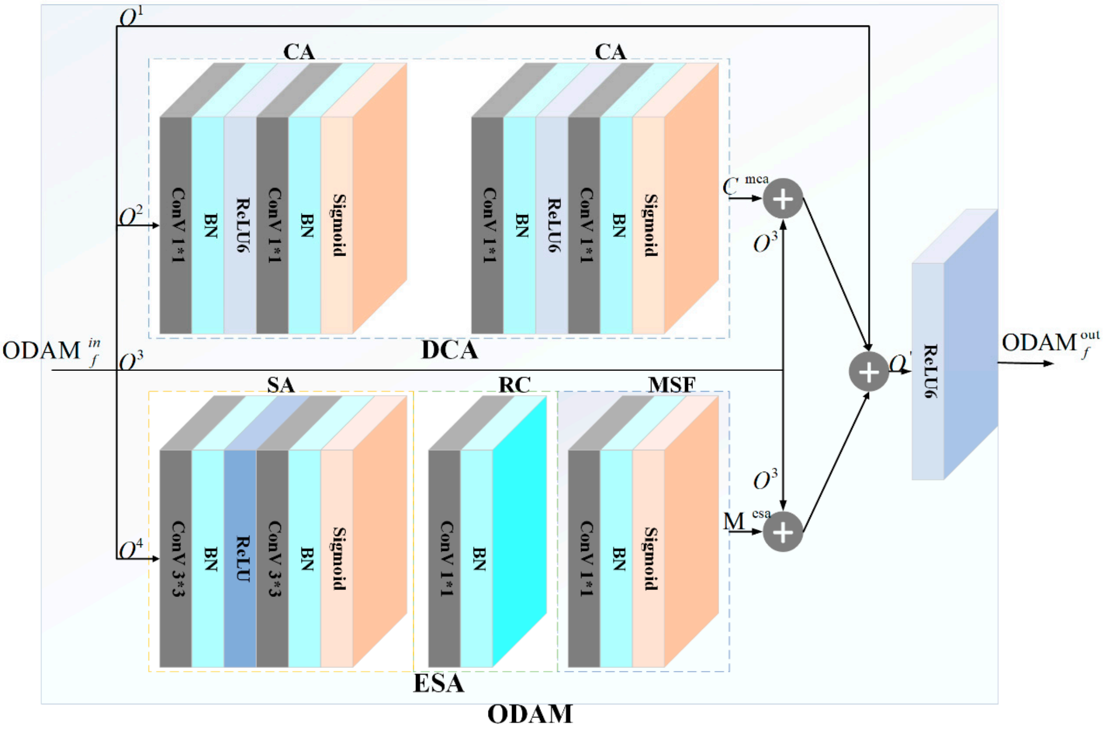

Figure 2. ODAM module structure.

3.1.1. DCA Mechanism

For the DCA module, the principle of its composition stems from the superior performance of CA in the channel dimension. To further enhance the effects in the channel dimension, we employ the DCA module, which comprises two serially connected CA modules. Simultaneously, to mitigate potential gradient explosion in some data, we chose the ReLU6 activation function. Equation (3) illustrates the formulation of the CA module.

$$CA^{out} = Sig\_BN\_ConV_{1,1}^{4,16}\_ReLU6\_BN\_ConV_{1,1}^{16,4}(O^2) \tag{3}$$

In the above equation, O^2 represents the input infrared and visible light images $\{I_{vi}, I_{ir}\}$. In $ConV_{\alpha^3,\alpha^4}^{\alpha^1,\alpha^2}$, α^1, α^2, α^3, and, α^4 stand for the number of input channels, the number of output channels, convolution kernel size, and stride in the convolution, respectively. *Sig* refers to the *Sigmoid* [48] activation function. The main role of the DCA module is to further enhance the effects of the network in the channel dimension on infrared and visible light images. We represent the DCA module in pseudo code, as Algorithm 1 shows below:

Algorithm 1: Dual-Channel Attention Mechanism Algorithm Principle

Input: $O^2(I_{vi}, I_{ir})$, feature map of a given layer in a deep learning model.
Output: $C^{mca}\left(I'_{vi}, I'_{ir}\right)$ feature map after applying DCA.
Parameter setting: Input channels: C. Reduces the channel dimension: R (default: 4).
The number of attention modules: N (default: 2).

Procedure:

```
- Initialize $DCA$ object which contains $N$ sequential attention modules.
- For each attention module:
  - Implement a sequential operation consisting of:
    - A $1 * 1$ convolution operation which reduces the channel number by a factor $R$, denoted as $ConV1$.
    - A BN operation applied to the output of $ConV1$, denored as $BN1$.
    - A $ReLU6$ activation function applied to the output of $BN1$.
    - A $1 * 1$ convolution operation which increases the channel number back to $C$, denoted as $ConV2$.
    - A BN operation applied to the output of $ConV2$, denoted as $BN2$.
    - A $Sigmoid$ activation function applied to the output of $BN2$ to obtain the attention map $A$.
  - Perform element-wise multiplication of the input feature map $O^2(I_{vi}, I_{ir})$ and the attention map $A$ to obtain the output feature map $C^{mca}\left(I'_{vi}, I'_{ir}\right)$ of this module.
- Repeat the above process for each attention module in the multi-channel attention object.
- Output the final feature map $C^{mca}\left(I'_{vi}, I'_{ir}\right)$ after passing through all attention modules.
```

The DCA mechanism module describes the multi-channel attention mechanism, where each channel in the input feature map is processed separately and then combined using attention maps. The attention modules are stacked together to form a deeper network, where each module generates a new feature map that is multiplied by its attention map. The final output is the feature map after passing through all the attention modules.

3.1.2. ESA Mechanism

In the context of ESA, the primary task is feature extraction from the input information, denoted as O^4, in the spatial dimension. Designed as an enhancement on the basic SA, ESA introduces the RC module, whose primary function is to process the input information through a convolutional layer and a BN layer. The processed input is then element-wise multiplied with the output from the SA module. This mechanism helps to enhance the extraction of important features from a spatial perspective, offering an increased provision of contextual information throughout the network and bolstering the focus on target areas. The addition of the MSF module serves to concatenate the output from the attention mechanism with the original infrared visible light input features, followed by the combination of these features through a convolutional layer and a BN layer. This step leverages multi-scale information holistically, further enhancing fusion effectiveness. Finally, the output from the MSF module is element-wise multiplied with the original input to yield the final output, denoted as M^{esa}. This ensures the comprehensive extraction of contextual information from the network. Equations (4)–(6) present the outputs of the SA, RC, and MSF modules, respectively.

$$SA^{out} = Sig\_BN\_ConV^{4,1}_{3,1,1}\_ReLU\_BN\_ConV^{16,4}_{3,1,1}\left(O^4\right) \tag{4}$$

$$RC^{out} = BN\_ConV^{16,16}_{1,1} \tag{5}$$

$$MSF^{out} = ConV^{17,1}_{1,1} \tag{6}$$

In the above formula, O^4 represents the input of infrared and visible light images $\{I_{vi}, I_{ir}\}$. In $ConV^{\alpha^1,\alpha^2}_{\alpha^3,\alpha^4,\alpha^5}$, α^1, α^2, α^3, α^4, and, α^5 represent the number of input channels of the convolution, the number of output channels, the size of the convolution kernel, the stride, and padding, respectively. Formula (7) expresses the output of the ESA module, M^{esa}.

$$M^{esa} = MSF^{out}\_RC^{out}\_SA^{out} \tag{7}$$

The ESA mechanism module describes the enhanced spatial attention mechanism. It enhances the standard spatial attention by adding a residual connection and a multi-scale fusion model, which can capture both local and global context information. The main role of the ESA mechanism module is to capture infrared and visible light information on different scales to better combine contextual information with spatial dimensions for extraction, thus enhancing the effectiveness of the spatial attention mechanism. Finally, for a clearer representation of the ESA module, we describe it in pseudo code as follows (Algorithm 2).

Algorithm 2: Enhanced Spatial Attention Mechanism Algorithm Principle

Input: $O^4(I_{vi}, I_{ir})$, feature map of a given layer in a deep learning model.
Output: $M^{esa}(I'_{vi}, I'_{ir})$, feature map after applying *ESA*.
Parameter setting: Input channels: *C*. Reduces the channel dimension: *R* (default: 4).

Procedure:

```
- Initialize the ESA object which consists of three essential parts: the main body, the RC, and the MSF.
- For the main body:
    - Implement a sequential operation consisting of:
        - A $3 * 3$ convolution operation that reduces the channel number by a factor of $R$, denoted as $ConV1$. This is padded to keep the same spatial dimension.
        - A BN operation applied to the output of $ConV1$, denored as $BN1$.
        - A $ReLU$ activation function applied to the output of $BN1$.
        - A $3 * 3$ convolution operation that reduces the channel number to 1, denoted as $ConV2$.
        - A BN operation applied to the output of $ConV2$, denoted as $BN2$.
        - A $Sigmoid$ activation function applied to the output of $BN2$ to obtain the attention map $A1$.
- For the RC:
    - Implement a sequential operation consisting of:
        - A $1 * 1$ convolution operation that keeps the channel number unchanged, denoted as $ConV3$.
        - A BN operation applied to the output of $ConV3$ to obtain the residual map $A2$.
- The attention output is the element-wise multiplication of $A1$ and $A2$.
- Concatenate $A1$ and the attention output in the channel dimension to form a multi-scale input.
- For the MSF:
    - Implement a sequential operation consisting of:
        - A $1 * 1$ convolution operation that reduces the channel number to 1, denoted as $ConV4$.
        - A BN operation applied to the output of $ConV4$.
        - A $Sigmoid$ activation function applied to the output to obtain the MSF map $A3$.
- The final output $M^{esa}(I'_{vi}, I'_{ir})$ is the element-wise multiplication of $A3$ and the input feature map $O^4(I_{vi}, I_{ir})$.
```

3.2. MO Module

Figure 3 shows the structure of the MO (multi-head self-attention, optimized MLP) module; through the figure, we can know that the MO module mainly consists of OMLP (optimized multi-layer perceptron) and MSA (multi-head self-attention) [49]. Equation (8) expresses its formula:

$$MO_f^{out} = OMLP\_LN\_MSA\_LN\left(MO_f^{in}\right) \tag{8}$$

where, MO_f^{in} represents infrared and visible image $\{I_{vi}, I_{ir}\}$ information; MSA is one of the modules we use.

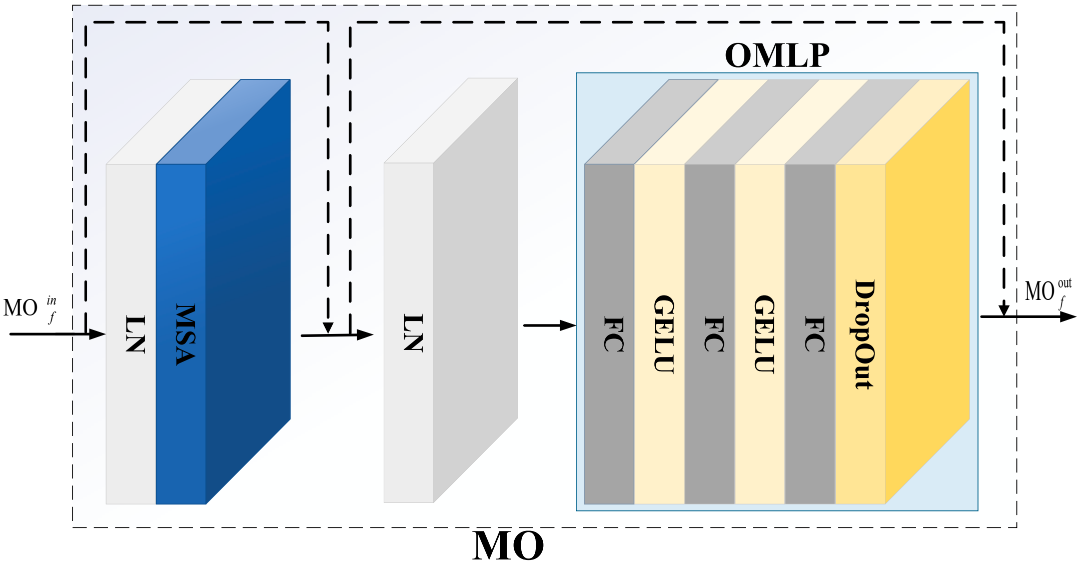

Figure 3. MO module structure.

The MSA module divides the input information MO_f^{in} into non-overlapping $N \times N$ windows, thus generating $HW/N^2 \times N^2 \times C$ features. The number of windows is HW/N^2, and each window automatically calculates its attention mechanism features. Initially, when the input of infrared and visible light information is $I = \{I_{vi}(i_1, i_2, i_3, \ldots, i_n),\ I_{ir}(i_1, i_2, i_3, \ldots, i_n)\}$, three vectors q, k, v can be obtained through linear transformation. The following equations represent the formula:

$$q = w_q I \tag{9}$$

$$k = w_k I \tag{10}$$

$$v = w_v I \tag{11}$$

where, $w_q \in \mathbb{R}^{d \times d_q}$, $w_k \in \mathbb{R}^{d \times d_k}$, and $w_v \in \mathbb{R}^{d \times d_v}$ represent the spatial transformation matrices; $d \times d_q$ represents the dimension of the output matrix. The calculation formula for the MSA module that we used is as follows:

$$MSA(q, k, v) = Sm\left(\frac{qk^t}{\sqrt{d_k}}\right)v \tag{12}$$

where, $\sqrt{d_k}$ represents a scaling factor, and $Sm(\blacksquare)$ represents the Softmax operation. Following this, the output from the MSA module, $MSA(q, k, v)$, is channel-wise added with the original input value MO_f^{in}, and then fed into the overhauled OMLP module through the LN (layer normalization) layer. The original MLP (multi-layer perceptron) only has two FC (fully connected) layers. In order to better extract infrared and visible light information, we fine-tune the MLP and set it to have three FC layers, replacing the original Sigmoid activation function with GELU. The OMLP module can be represented as Equation (13) shows.

$$OMLP(I_{ir}, I_{vi}) = DropOut\_FC\_GELU\_FC\_GELU\_FC(\blacksquare) \tag{13}$$

Lastly, the output value of the MO module is represented as MO_f^{out}, which can be denoted with the formula as Equation (14) illustrates.

$$MO_f^{out} = OMLP(I_{ir}, I_{vi}) \oplus \left\{ MSA(q, k, v) \oplus MO_f^{in} \right\} \tag{14}$$

The MO module represents the OMLP class in a simplified pseudocode format. The three-layer structure is common in MLP (multi-layer perceptron), but the specifics (such as the number of hidden features and the type of activation function) can be customized. Dropout is used to prevent overfitting by randomly setting a fraction of input units to 0 at each update during training time. We describe it in pseudo code as follows (Algorithm 3).

Algorithm 3: Optimized Multi-Layer Perceptron Algorithm Principle

```
Input: $MO_f^{in}(I_{vi}, I_{ir})$ hidden features: $HF$, output features: $OF$, active layer: $ReLU$, drop: 0.2.
Output: $MO_f^{out}(I'_{vi}, I'_{ir})$.
Parameter setting:
- Define a class $OMLP$ that inherits from module.
- In the initialize, $MO_f^{in}(I_{vi}, I_{ir})$, $HF$ (default: $MO_f^{in}(I_{vi}, I_{ir})$), $OF$ (default: $MO_f^{in}(I_{vi}, I_{ir})$), active layer (default: $ReLU$), and drop (default: 0.2).
Procedure:
- Initialize three fully connected layers: $fc1$, $fc2$ and $fc3$. $fc1$ and $fc2$ have the same number of $HF$, and $fc3$ has $OF$.
- Initialize $act1$ and $act2$ as instances of the activation function specified by active layer.
- Initialize a dropout layer with the specified dropout rate.
The forward function:
- For $P$ steps do:
    - Pass $MO_f^{in}(I_{vi}, I_{ir})$ through $fc1$, apply the activation function $act1$, and apply dropout to get intermediate output $Out_1$.
    - Pass $Out_1$ through $fc2$, apply the activation function $act2$, and apply dropout to get intermediate output $Out_2$.
    - Pass $Out_2$ through $fc3$ to get the final output $MO_f^{out}(I'_{vi}, I'_{ir})$.
- Return $MO_f^{out}(I'_{vi}, I'_{ir})$.
```

3.3. Loss Function

Owing to the intrinsic disparities in their imaging principles, infrared light sensors and visible light sensors exhibit different levels of sensitivity toward various environmental targets. Employing a loss function, specific parameters of the model can be optimized, thereby enhancing the performance of the fused image I_f. Adjusting to the current environment, we design the loss function from three aspects: SSL, PL, and GL, constituting the loss function of our OMOFuse network. Its definition is as follows:

$$\mathcal{L} = \alpha \mathcal{L}_s + \beta \mathcal{L}_p + \gamma \mathcal{L}_g \tag{15}$$

where, $\mathcal{L}_s$ represents SSL; $\mathcal{L}_p$ represents PL; $\mathcal{L}_g$ represents GL; α, β, and γ represent weight factors for each loss. In the context of OMOFuse, the weight factors α, β, and γ are assigned as 0.8, 5, and 100, respectively. Hence, the given $\mathcal{L}$ can also be elucidated as:

$$\mathcal{L} = 0.8\mathcal{L}_s + 5\mathcal{L}_p + 100\mathcal{L}_g \tag{16}$$

The magnitudes of the weighting factors α, β, and γ are adjustable based on experimental environments and data, thereby potentially enhancing fusion efficiency incrementally. Among these, the calculation formula for $\mathcal{L}_s$ can be represented as:

$$\mathcal{L}_s = \varphi_1\left(1 - I_f, I_{ir}\right) + \varphi_2\left(1 - I_f, I_{vi}\right) \tag{17}$$

In the above equation, φ_1 and φ_2 represent two data-driven weights, their magnitude determined by the image itself, rather than artificially set. The computational formula for $\mathcal{L}_p$ can be expressed as:

$$\mathcal{L}_p(x,y) = \sum_i \frac{1}{H_i W_i C_i} \|\phi_i(x) - \phi_i(y)\|_2^2 \tag{18}$$

In the above equation, $\phi_i(x)$ represents the feature map extracted by the convolutional layer before the i-th max-pooling, where the size of the i-th feature map is $H_i * W_i * C_i$. $\|\blacksquare\|_2^2$ denotes the *Frobenius* norm of a matrix. The PL of the i-th feature map can be expressed as:

$$\mathcal{L}_{p^i} = \varphi_{i_{ir}}\mathcal{L}_{p^i}\left(I_f, I_{ir}^i\right) + \varphi_{i_{vi}}\mathcal{L}_{p^i}\left(I_f, I_{vi}^i\right) \tag{19}$$

In the above equation, I_{ir}^i and I_{vi}^i represent the feature map information of the i-th infrared and visible light, respectively. To enhance the network's performance in terms of content and spatial structure, we introduced the $\mathcal{L}_g$ loss. The calculation formula for $\mathcal{L}_g$ can be expressed as:

$$\mathcal{L}_g(x,y) = \frac{1}{HW}\|\nabla x - \nabla y\|_f^2 \tag{20}$$

In the above formula, $\nabla(\blacksquare)$ composes the gradient change between image x and y. Equation (21) presents the gradient loss of image j.

$$\mathcal{L}_{g^j} = \varphi_{j_{ir}}\mathcal{L}_{p^j}\left(I_f, I_{ir}^j\right) + \varphi_{j_{vi}}\mathcal{L}_{p^j}\left(I_f, I_{vi}^j\right) \tag{21}$$

3.4. Network Structure of this Paper

Figure 4 illustrates the structure of the OMOFuse network. The inputs are infrared ($I_{ir} \in \mathbb{R}^{H\times W\times 3}$) and visible light ($I_{vi} \in \mathbb{R}^{H\times W\times 1}$) images. These inputs sequentially pass through a convolutional layer with a $5 * 5$ kernel, a BN layer, a ReLU6 layer, a convolutional layer with a $1 * 1$ kernel, a Sigmoid layer, a convolutional layer with a $3 * 3$ kernel, another BN layer, and another ReLU layer. Subsequently, the data passes through an ODAM module and two concatenated MO modules, finishing with a ReLU6 layer. The output is represented as $I_f \in \mathbb{R}^{H\times W\times 1}$. Through the simple configuration of the OMOFuse network, we achieved an efficient fusion of infrared and visible light of high-class significance.

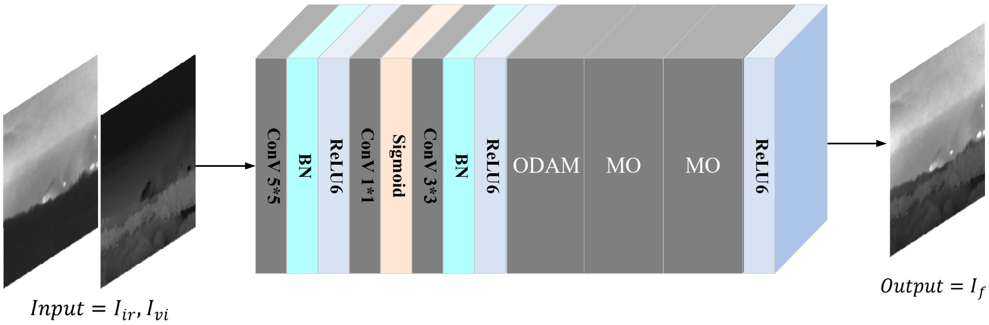

Figure 4. OMOFuse module structure.

4. Experimental Results and Analysis

In this section, we initiate with an elucidation of the experimental setting and performance metrics. Subsequently, we perform both a qualitative and quantitative comparison between the DHV datasets [50] and our selected evaluation indices.

4.1. Experimental Settings

To ensure the objectivity and fairness of the experimental data, all our tests are carried out on an NVIDIA GeForce RTX 3080TI GPU using the PyTorch framework. Concurrently, the code for the experiment utilizes the author's publicly disclosed parameter code, maintaining the original hyperparameters. Within the OMOFuse network, we have set the number of iterations to 100, the batch size to 24, and the initial learning rate to 1×10^{-2}. For the data augmentation phase, the size is randomly center-cropped to 224 × 224, and mosaic data augmentation [51] has been applied. As the optimizer, we selected AdamW [52].

4.2. Evaluation Metrics

While numerous evaluation metrics exist for image fusion, they can generally be categorized into four broad types [53]: distortion measurement metrics, similarity metrics, visual quality metrics, and blur measurement metrics. Each category provides a distinct emphasis in their assessment of the fused image I_f and takes into account respective information classes during design. Nevertheless, it has been observed that many researchers use a rather narrow selection of evaluation metrics in their holistic critique of the fused image I_f, resulting in a less comprehensive evaluation of algorithmic performance. Hence, this study has singled out ten evaluation metrics from the four categories to use in our paper. They include EN (entropy), SF (spatial frequency), SD (standard deviation), PSNR (peak signal-to-noise ratio), MSE (mean-square error), MI (mutual information), SCD (sum of correlation differences), $\mathrm{N}^{\mathrm{abf}}$ (noise adaptive fuzzy equalization), and MS_SSIM (multiscale structural similarity index measure). The computational methods for each evaluation metric are presented in the following:

The EN is predominantly used to assess the information entropy of fused images, evaluating the complexity of an image. A smaller value is considered more desirable, and Equation (22) shows the computation method.

$$EN = -\sum_{l=0}^{L-1} P_l log_2 P_l \tag{22}$$

In the formula above, L stands for the grayscale level of the image, while P_l represents the normalized histogram of the grayscale level corresponding to the fused image. SF serves to assess the spatial frequency response after image fusion, which gauges the detailed information of the image—the larger, the better. Equation (23) exhibits the calculation method.

$$SF = \sqrt{RF^2 + CF^2} \tag{23}$$

In the formula above, $RF = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (F(i,j) - F(i,j-1))^2}$ represents row frequency and $RF = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (F(i,j) - F(i-1,j))^2}$ represents column frequency. SD is utilized for evaluating the pixel value distribution of the fused images, serving as a measure of image contrast. A higher value suggests better contrast. Equation (24) illustrates the computation method.

$$SD = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (F(i,j) - \mu)^2} \tag{24}$$

In the formula above, μ represents the mean of the fused images. The PSNR is used to evaluate the extent of distortion between the fused images and the original ones, with a higher value indicating better performance. Equation (25) illustrates the calculation method.

$$PSNR = 10lg\frac{r^2}{MSE} \tag{25}$$

In the formula above, r represents the peak of the fused image, and MSE is employed to evaluate the level of distortion between the fused and original images. A lower MSE value is desirable, indicating minimal distortion. Equation (26) demonstrates the computation method of the MSE.

$$MSE = \frac{MSE^{A,F} + MSE^{B,F}}{2} \tag{26}$$

In the formula above, $MSE^{A,F}$ and $MSE^{B,F} = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(A/B(i,j) - F(i,j))^2$. MI is employed to appraise the information content of fused images, serving as a gauge for image similarity. Higher values connote better similarity. Equation (27) depicts the calculation method.

$$MI = MI^{A,F} + MI^{B,F} \tag{27}$$

In the formula above, $\mathrm{MI}^{\mathrm{A,F}}$ represents the information transferred from the original image A to the fused image F, and $\mathrm{MI}^{\mathrm{B,F}}$ represents the information transferred from the original image B to the fused image F. The calculation formulas for $\mathrm{MI}^{\mathrm{A,F}}$ and $\mathrm{MI}^{\mathrm{B,F}}$ are as follows: $MI^{A,F} = \sum_{a,f} P_{A,F}(a,f)log\frac{P_{A,F}(a,f)}{P_A(a)P_F(f)}$, $MI^{B,F} = \sum_{b,f} P_{B,F}(b,f)log\frac{P_{B,F}(b,f)}{P_B(b)P_F(f)}$. Here, $P_A(a)$, $P_B(b)$, and $P_F(f)$ represent the marginal histograms of the original images A and B, and the fused image F, respectively. $P_{A,F}(a,f)$ and $P_{B,F}(b,f)$ denote the joint histograms of original images A and B with the fused image F. SCD is utilized to evaluate the similarity between the fused image and the original images. A higher SCD value indicates a better match. Equation (28) shows the calculation method.

$$SCD = r(A, D_{A,F}) + r(B, D_{B,F}) \tag{28}$$

In the formula above, $\mathrm{D_{A,F}}$ and $\mathrm{D_{B,F}}$ represents the difference images between the original images A and B, and the fused image F, respectively. The quality of the fused image is measured by $\mathrm{N^{abf}}$, with a lower value indicating better fusion quality, as Equation (29) details.

$$N^{abf} = \frac{\sum_{i-1}^{M}\sum_{j=1}^{N} AM(i,j)\left[\left(1 - Q^{A,F}(i,j)\right)w^A(i,j) + \left(1 - Q^{B,F}(i,j)\right)w^B(i,j)\right]}{\sum_{i=1}^{M}\sum_{j=1}^{N}(w^A(i,j) + w^B(i,j))} \tag{29}$$

$$MS_{SSIM(x,f)} = [l_m(x,f)]^{\alpha_{M'}} \times \prod_{j=1}^{M'}[c_j(x,f)]^{\beta_j} \times \left[s_j(x,f)\right]^{\gamma_j} \tag{30}$$

In the formula above, $l_{M'}(x,f)$ represents the luminance similarity value in the M′th dimension, while $c_j(x,f)$ and $s_j(x,f)$ represent the image contrast and structural similarity in the jth dimension, respectively. The hyperparameters α, β, and γ primarily function to balance various image components.

In this section, the OMOFuse algorithm is compared to the latest or classic fusion algorithms, including CNN [54], Hybrid_MSD [3], IFEVIP [55], MDLatLRR [56], PPT Fusion [57], SDNet [58], SeAFusion [23], SuperFusion [59], SwinFuse [60], and TIF [61]. In our study, we utilized an array of comparative algorithms, which virtually encompassed the most recent advances in visible and infrared light fusion methodologies. The majority of these state-of-the-art (SOTA) algorithms have surfaced within the past three years, providing a representation of the developmental trajectory in the domain of infrared and visible light fusion to some extent.

4.3. DHV Datasets

We have selected the fire sequence of the DHV datasets, which includes a combination of 212 images—106 pairs of visible light and 106 pairs of infrared light fire imagery. Furthermore, all data are characterized by an identical shooting angle confirmed through image rectification. To illustrate the fusion algorithm, we choose to use a qualitative comparison and a quantitative comparison, respectively.

4.3.1. Qualitative Comparison

In our study, we conducted a qualitative comparison between the OMOFuse algorithm and ten classic algorithms, as well as SOTA algorithms. We randomly selected five images from the fire sequence of the DHV datasets for visualization. In these images, the first row represents infrared images, while the second row shows visible light images. As Figure 5 evidences, the OMOFuse algorithm exhibits a superior performance in terms of realism, especially in areas featuring "grasslands" and "trees". Its clarity and realistic rendering outperform other algorithms. Notably, we did not observe the loss of image detail that was common in the SDNet and SwinFuse algorithms, where their output tended to exhibit a darkened effect.

In the "fire point" combustion segment, infrared light sensors present several black spots, which do not perform as well as the visible light images. However, the merged "fire point" combustion section demonstrates superior image quality. The OMOFuse algorithm, particularly on the right-hand side of the "fire point" combustion segment, visually offers more enriched edge information and a higher degree of sharpness compared to other algorithms. In contrast, the Hybrid_MSD, MDLatLRR, PPT Fusion, and TIF algorithms introduce blurriness, with a reduction in clarity, while OMOFuse appears to behave normally. Concerning the fourth column of the "fire point" section, the image clarity is significantly improved after OMOFuse fusion. We also note that the fused results produced by the PPT Fusion algorithm are relatively blurred, indicating unsatisfactory fusion performance.

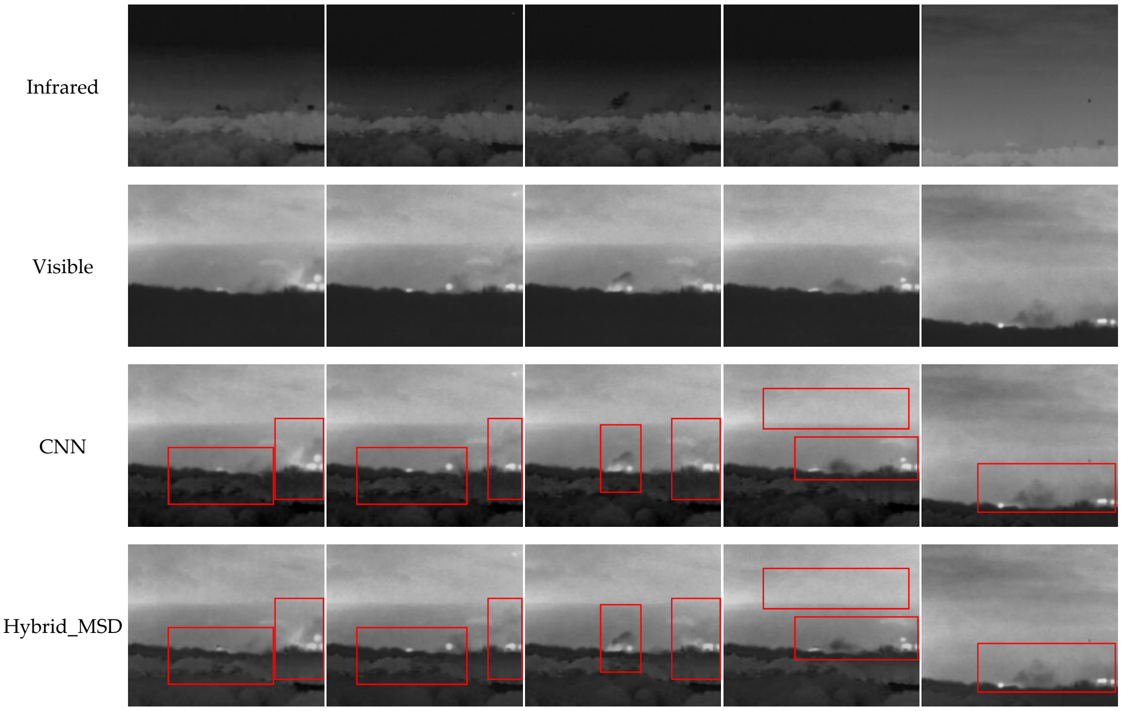

Figure 5. *Cont.*

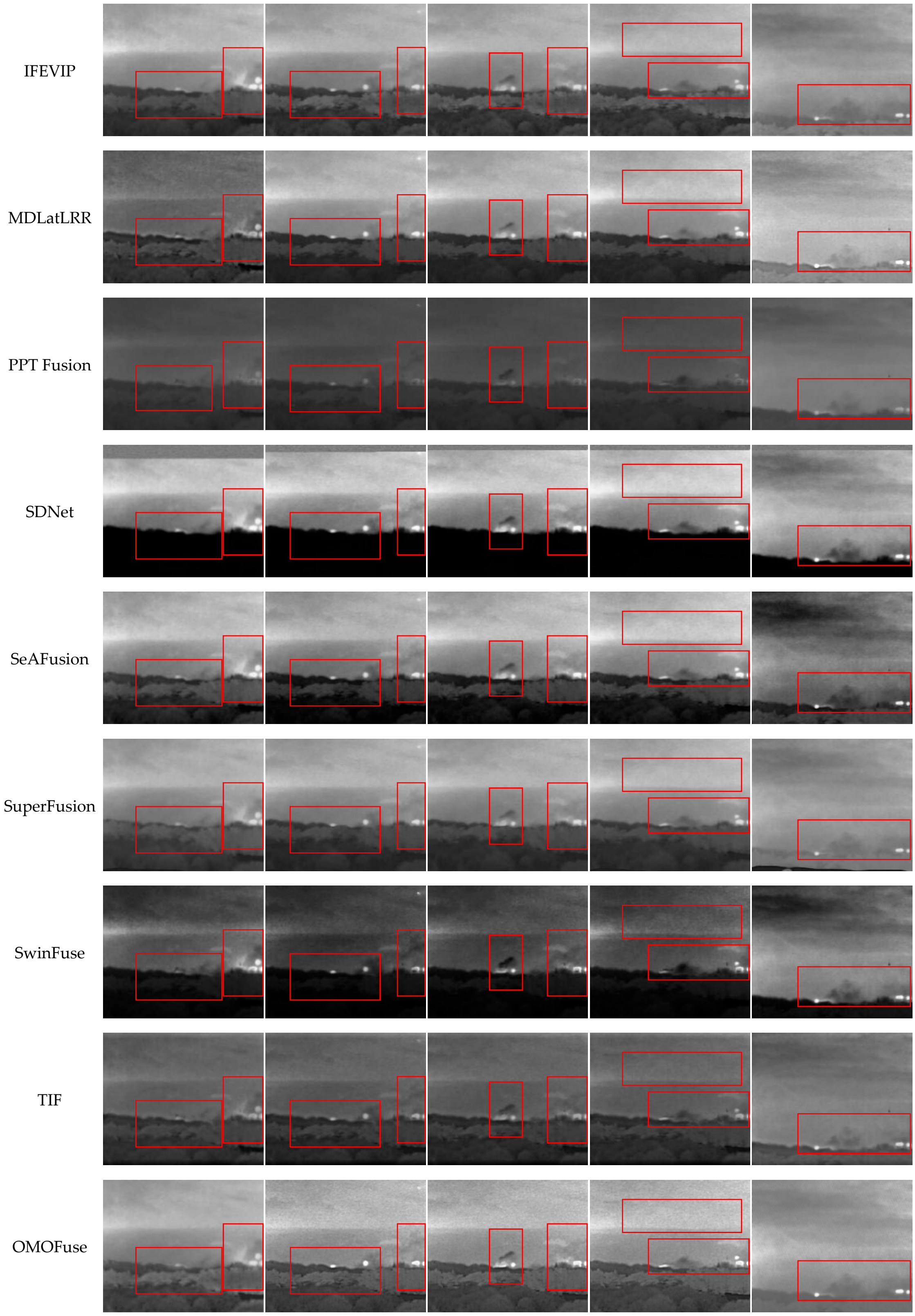

Figure 5. Comparison chart of partial visualization effects of DHV images datasets.

4.3.2. Quantitative Comparison

We subsequently carried out a quantitative comparison of the experimental results, as Table 1 illustrates. To ensure the validity of the data, the data in Table 1 were all averaged using 106 pairs of infrared and visible light fusion results. From Table 1, we discern that out of nine evaluation criteria, the OMOFuse algorithm clinches first place in four instances, second place in two, and third place in one. Despite not being ranked amongst the top three for the evaluation metrics MSE and N^{abf}, its performance is still notably above average amongst the ten compared algorithms. Hence, we can conclude that the OMOFuse algorithm holds undeniable superiority across all comparison metrics (distortion measurement metrics, similarity metrics, visual quality metrics, blur measurement metrics) in the algorithmic evaluation.

Table 1. Objective evaluation of classic and latest fusion algorithms. These are the results of the DHV images datasets. ↑/↓ for a metric represents that a larger/smaller value is better. The best three values in each metric are denoted in red, green, and blue, respectively. The first ranked effect is marked in red, the second ranked in green, and the third ranked in blue.

Method	**EN** ↓	**SF** ↑	**SD** ↑	**PSNR** ↑	**MSE** ↓	**MI** ↑	**SCD** ↑	$\mathbf{N^{abf}}$ ↓	**MS_SSIM** ↑
CNN	6.7051	0.0105	11.0482	60.7441	0.0598	5.4084	1.2782	0.1164	0.9569
Hybrid_MSD	6.6069	0.0108	10.5830	61.3735	0.0535	4.4112	1.3995	0.1540	0.9681
IFEVIP	6.8600	0.0118	10.3508	60.2827	0.0644	5.2314	1.5940	0.1885	0.9457
TIF	6.2582	0.0101	9.8712	63.1943	0.0311	3.1823	1.6139	0.2001	0.9551
MDLatLRR	6.2940	0.0153	9.8651	63.0112	0.0315	2.8957	1.6325	0.4287	0.9475
SwinFuse	6.5653	0.0125	10.2428	61.4893	0.0468	4.2972	1.5877	0.2616	0.9533
SuperFusion	6.7837	0.0102	10.3393	60.9628	0.0574	4.5202	0.5757	0.1612	0.9156
SeAFusion	7.2099	0.0138	10.6302	60.7421	0.0646	4.8932	1.3032	0.4054	0.9640
PPT Fusion	5.5468	0.0056	8.4609	63.3165	0.0330	3.7637	1.5796	0.0128	0.9114
SDNet	5.8295	0.0245	10.9580	58.6278	0.0933	4.6873	0.5985	0.2754	0.8512
OMOFuse	5.8332	0.0253	10.9961	63.4045	0.0413	5.3462	1.6638	0.1716	0.9713

Due to the possibility of substantial randomness in the data, we have prepared line graphs for the evaluation indicators of each of the 106 pairs of infrared and visible light images, as Figure 6 shows. This approach enables us to obtain results that are as accurate as possible. In the graphs, the horizontal axis (image pairs) spans the range of [0, 106], and the range of the vertical axis corresponds to the values of the metric. The table above shows that the OMOFuse algorithm performs best in terms of the mean values of SF, PSNR, SCD, and MS_SSIM. With regard to the SF evaluation indicator, Figure 6 shows that the overall performance of the OMOFuse algorithm considerably outperforms other algorithms. At the same time, we found that the performance of the SDNet algorithm fluctuates, but its overall performance is inferior to that of the OMOFuse algorithm. Therefore, our method surpasses the SDNet algorithm in terms of mean values. In reference to the evaluation index PSNR, Figure 6 reveals that the performance of the OMOFuse algorithm superseded other algorithms for most images, corroborating the data presented in the table above. Considering the SCD evaluation index, although the preliminary performance of the OMOFuse algorithm was not optimal, its latter half of impact culminated in an improved average, thereby the OMOFuse algorithm exhibits better average performance on the SCD index. As for the MS_SSIM evaluation index, Figure 6 indicates that the OMOFuse algorithm displays overall superior performance compared to other algorithms. Hence, we can validate that the OMOFuse algorithm exhibits the best average performance on SF, PSNR, SCD, and MS_SSIM. Additionally, in the evaluation indices SD and MI, the performance of the OMOFuse algorithm remains consistent with the data demonstrated in the table above, thus confirming the accuracy of the given data.

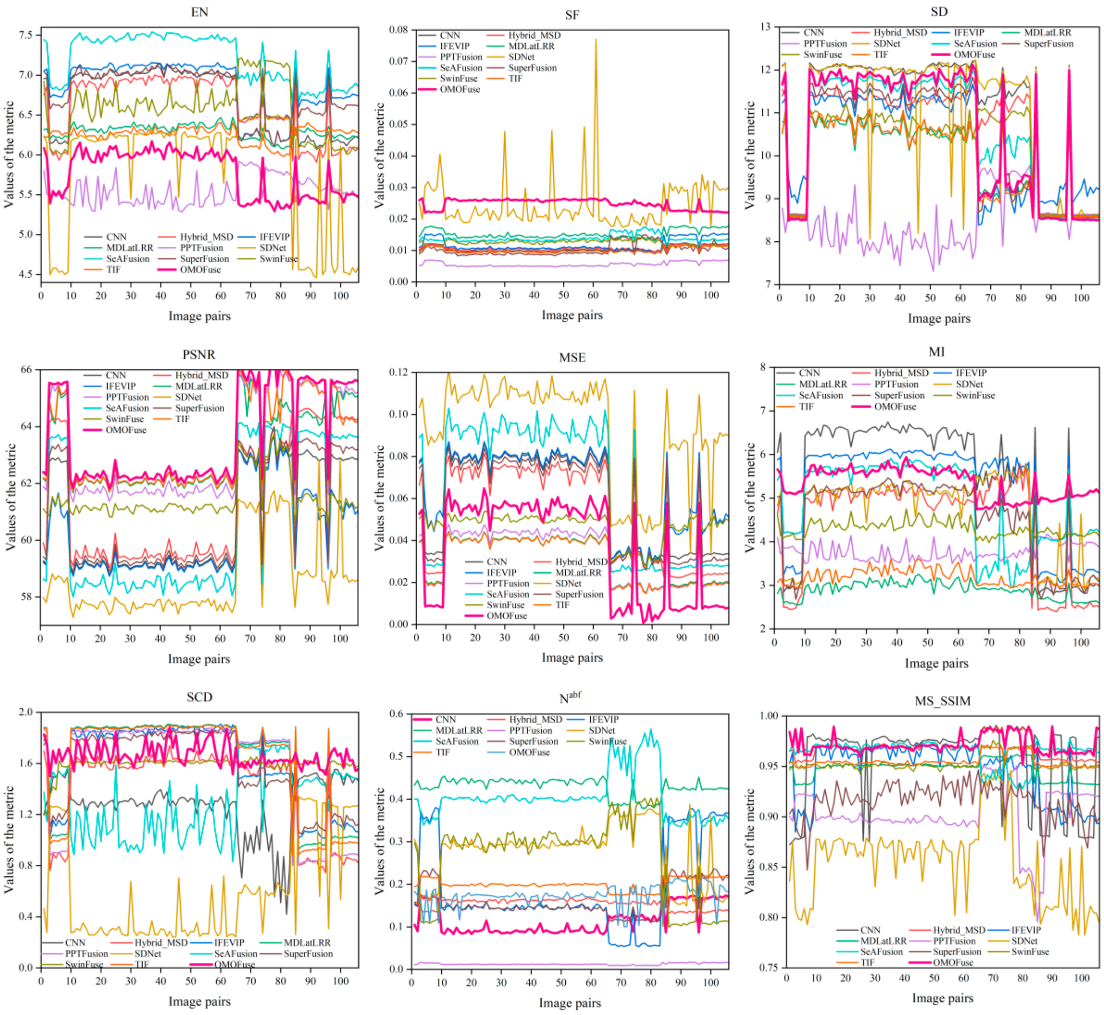

Figure 6. Variation curve graphs of various evaluation metrics for DHV datasets, where the OMOFuse algorithm is highlighted and marked in bold red. A quantifiable comparison is conducted with 9 evaluation metrics and 10 different methods.

5. Extended Study

In the following, we conduct both qualitative and quantitative comparisons using the M3FD TNO datasets [62] and Roadscene datasets [34]. More specifically, the M3FD TNO datasets consist of 37 pairs, in total 74, of infrared and visible light images that have been corrected and registered. The pictorial contents within these images comprise various scenarios including “tanks”, “personnel”, and “buildings”. The Roadscene datasets, which are constructed based on FLIR, constitute a fusion dataset between infrared and visible light images. The dataset contains image pairs that depict highly repetitive scenarios, primarily involving “roads”, “pedestrians”, and “vehicles”. We utilized 42 pairs or 84 images in total from this dataset.

5.1. M3FD TNO Datasets

5.1.1. Qualitative Comparison

Figure 7 visualizes the partial fusion effects of infrared and visible light from the M3FD TNO datasets. As Figure 7 displays, overexposure appears in the second and third columns for MDLatLRR, specifically in the “wheel” and “road” areas, resulting in excessive brightness and subpar performance. PPT Fusion, SDNet, and SwinFuse show a background darkening effect in the second and fourth columns, leading to a decrease in the clarity of some targets. SeAFusion and SuperFusion demonstrate higher levels of blurriness in the third column, suggesting inadequate image clarity. In a comprehensive perspective, the algorithms that perform relatively well on the M3FD TNO datasets include CNN, Hybrid_MSD, IFEVIP, TIF, and OMOFuse. A detailed observation reveals a relative blur in the "ground" section of the fourth column image by CNN, and TIF shows higher levels of blurriness in the "door" section of the fourth column image. Finally, through careful comparison, we determined that OMOFuse excels in terms of texture detail and realism compared to Hybrid_MSD and IFEVIP, thus demonstrating superior performance and making OMOFuse the optimal choice.

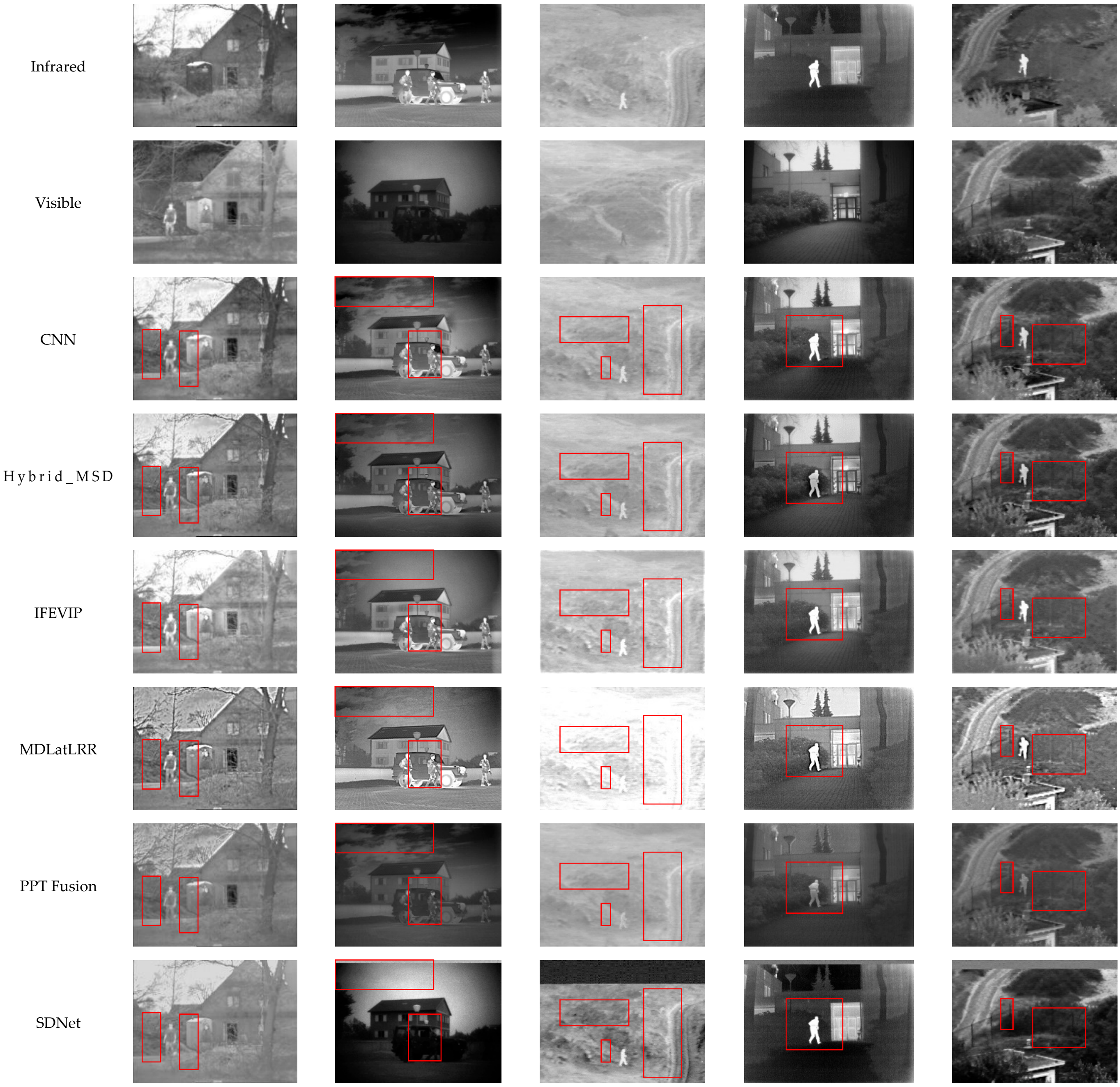

Figure 7. *Cont.*

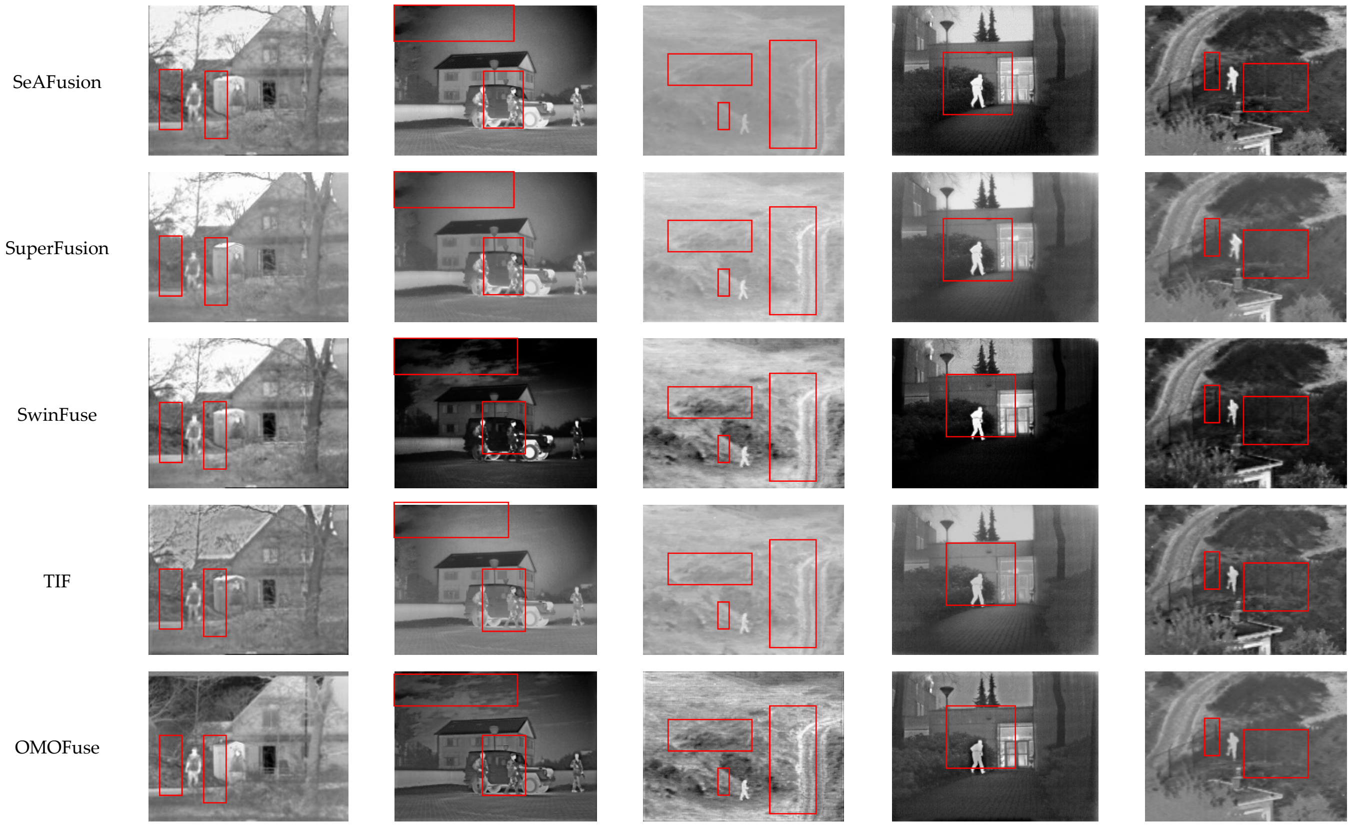

Figure 7. Comparison chart of partial visualization effects of M3FD TNO datasets.

5.1.2. Quantitative Comparison

Table 2 presents the experimental results of the M3FD TNO datasets. Each piece of data represents the average of 37 pairs of fused images. As Table 2 shows, OMOFuse secured three first-place positions, four second-place positions, and one third-place position in the M3FD TNO datasets, outperforming other classic algorithms and SOTA algorithms. This further establishes that our algorithm, OMOFuse, possesses superior generalization and fusion capabilities. The mean values in the above table to some extent reflect the performance of OMOFuse and other algorithms. For further inspection of each algorithm's performance, we have illustrated the fusion effects of 37 pairs of infrared and visible-light images as a line graph, as Figure 8 shows. Through Figure 8, we can more clearly observe the changing trajectory and magnitude of the nine evaluation indices for each image. The red bold line represents the OMOFuse algorithm. Our findings reveal that the overall performance of the OMOFuse algorithm is largely consistent with the data in Table 2, confirming the validity of Table 2's data.

Table 2. Objective evaluation of classic and latest fusion algorithms. These are the results of the M3FD TNO images datasets. ↑/↓ for a metric represents that a larger/smaller value is better. The best three values in each metric are denoted in red, green, and blue, respectively. The first ranked effect is marked in red, the second ranked in green, and the third ranked in blue.

Method	**EN** ↓	**SF** ↑	**SD** ↑	**PSNR** ↑	**MSE** ↓	**MI** ↑	**SCD** ↑	$\mathbf{N^{abf}}$ ↓	**MS_SSIM** ↑
CNN	7.1797	0.0509	9.6380	62.6111	0.0445	2.3819	1.6533	0.1348	0.9424
Hybrid_MSD	7.0103	0.0537	9.2939	63.3682	0.0367	2.2593	1.5827	0.1666	0.9340
IFEVIP	6.8565	0.0447	9.2252	61.4735	0.0548	3.5519	1.5320	0.1158	0.8630
MDLatLRR	6.9375	0.0706	9.2717	63.7603	0.0325	1.5358	1.6063	0.3734	0.9204
PPT Fusion	6.5203	0.0313	8.8465	64.4007	0.0295	2.2453	1.5328	0.0203	0.8628

Table 2. *Cont.*

Method	EN ↓	SF ↑	SD ↑	PSNR ↑	MSE ↓	MI ↑	SCD ↑	N^{abf} ↓	MS_SSIM ↑
SDNet	6.7433	0.0474	9.7989	60.2740	0.0694	6.9044	0.5985	0.1995	0.6846
SeAFusion	7.1796	0.0516	9.7298	61.9096	0.0511	2.7547	1.6917	0.2709	0.8877
SuperFusion	6.7751	0.0370	9.0967	61.9471	0.0512	2.8732	1.3693	0.1157	0.7310
SwinFuse	7.1295	0.0569	9.5963	61.7316	0.0492	2.5453	1.8313	0.2124	0.9110
TIF	6.7425	0.0453	9.1220	64.3530	0.0292	1.7497	1.5964	0.1257	0.9401
OMOFuse	6.5414	0.0594	10.2509	64.5596	0.0288	3.4156	1.7671	0.0528	0.7920

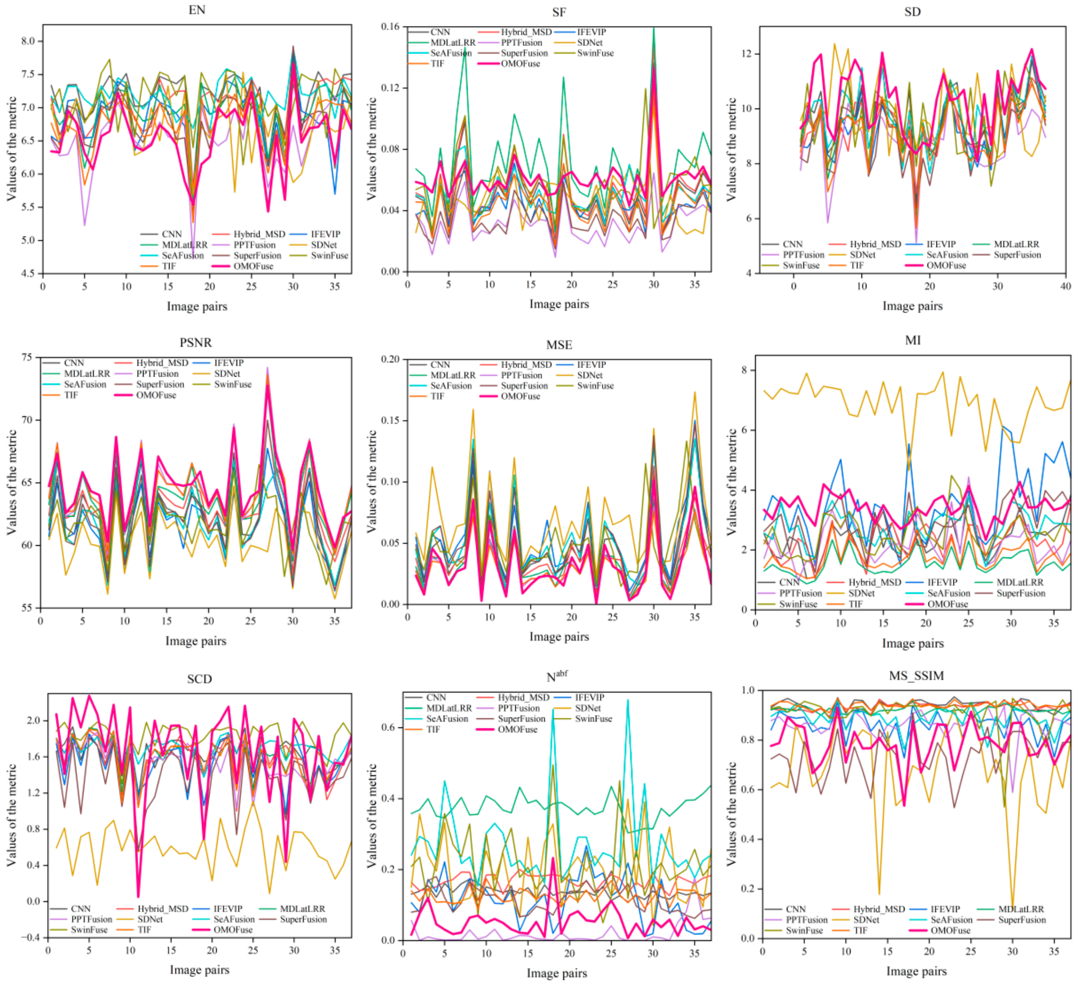

Figure 8. Variation curve graphs of various evaluation metrics for M3FD TNO datasets, where the OMOFuse algorithm is highlighted and marked in bold red. A quantifiable comparison is conducted with 9 evaluation metrics and 10 different methods.

5.2. Roadscene Fusion Datasets

5.2.1. Qualitative Comparison

In this study, we randomly selected five images from the Roadscene datasets for a visualization demonstration. As Figure 9 shows, the image clarity of the first column is relatively lower for the Hybrid_MSD and SuperFusion algorithms, with higher degrees of blur. Regarding the IFEVIP algorithm, its performance in fusing the “license plate” in the first and second columns was subpar. Additionally, the fusion performances of the IFEVIP, MDLatLRR, SeAFusion, and SwinFuse algorithms concerning second, third, and fifth column images were negatively impacted by excessive exposure in brighter areas of the photographs, compromising recognition effects. Furthermore, the realism of the fused “trees” in the fourth column for CNN, Hybrid_MSD, PPT Fusion, SuperFusion, and

SwinFuse was poor, as the deformed fusion effects of the trees were quite perceptible. TIF's performance regarding image clarity on the "license plate" in the second column was also unsatisfactory. Upon a comprehensive review, the OMOFuse algorithm exemplary demonstrated superior performance without the issues manifested in the other algorithms. Therefore, we ascertain from a visualization perspective that the OMOFuse algorithm possesses commendable fusion capabilities.

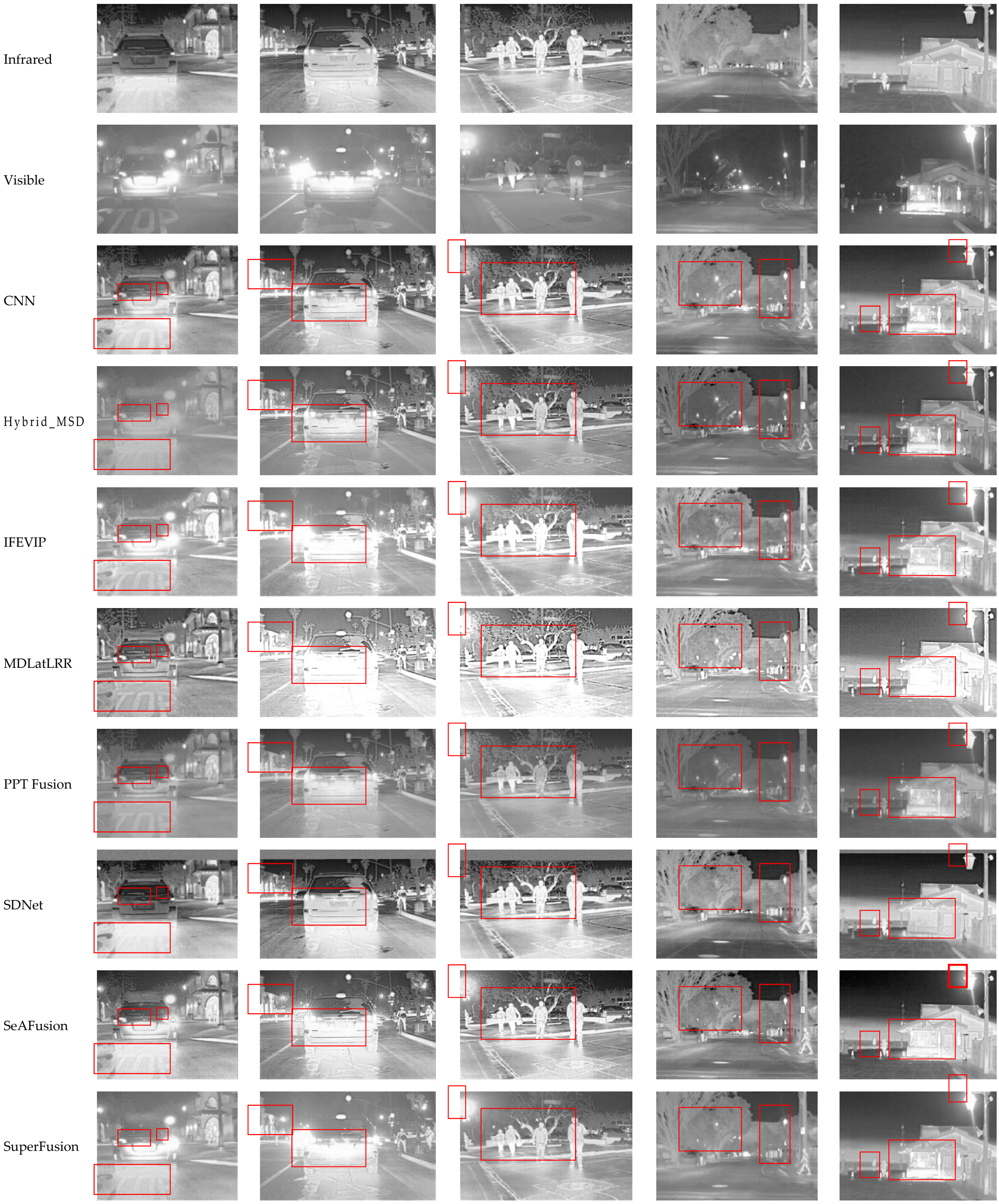

Figure 9. *Cont.*

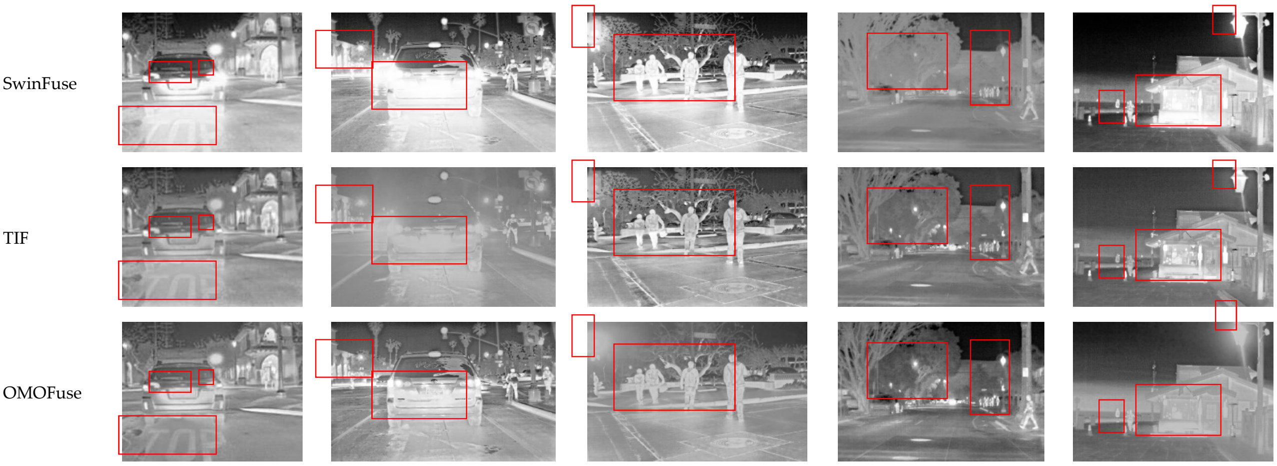

Figure 9. Comparison chart of partial visualization effects of Roadscene datasets.

5.2.2. Quantitative Comparison

Assuming you are now crafting a scientific paper and have presented all experimental data in Table 3 which, upon careful evaluation, shows the following results: OMOFuse ranks first in two categories, second in three instances, and third in another three. It is also worth noting that PPT Fusion exhibits reasonably good performance on Roadscene datasets, achieving first place in three instances and second in one. Furthermore, TIF stands at second place in one case and third in three. However, in light of overall performance, we opine that PPT Fusion lacks stability. For example, the SF of PPT Fusion is merely 0.0333, which is the worst performance in the entire table, whereas OMOFuse delivers a relatively stable performance. Figure 10 presents the location-wise fusion results of 42 pairs of infrared and visible light images of Roadscene datasets. The bold red line represents the OMOFuse algorithm. However, it is important to highlight that high scores of image fusion indexes may not universally indicate superior fusion performance, as an algorithm performing well on one image may not guarantee similar outcomes on others. This illustrates the fact that the performance of image fusion techniques varies with the nature of the images used. Hence, Figure 10 offers a more intuitive way to exhibit the performance of each pair of images based on evaluation indexes and, when contrasted with Table 3, aids in further validating the data within the table.

Table 3. Objective evaluation of classic and latest fusion algorithms. These are the results of the Roadscene Fusion images datasets. ↑/↓ for a metric represents that a larger/smaller value is better. The best three values in each metric are denoted in red, green, and blue, respectively. The first ranked effect is marked in red, the second ranked in green, and the third ranked in blue.

Method	**EN** ↓	**SF** ↑	**SD** ↑	**PSNR** ↑	**MSE** ↓	**MI** ↑	**SCD** ↑	**N^{abf}** ↓	**MS_SSIM** ↑
CNN	7.3992	0.0521	10.2391	63.8009	0.0337	3.4795	1.4021	0.1338	0.9483
Hybrid_MSD	7.1189	0.0562	9.4978	64.9477	0.0255	2.5874	1.2136	0.1590	0.9366
IFEVIP	7.0018	0.0494	9.9892	62.1615	0.0461	3.9051	1.2489	0.1950	0.8624
MDLatLRR	7.2945	0.0768	9.8720	64.4147	0.0277	2.3123	1.3084	0.3924	0.9055
PPT Fusion	6.9991	0.0333	9.6836	65.6863	0.0221	3.1211	1.2299	0.0147	0.8895
SDNet	7.3484	0.0586	10.4042	61.7559	0.0500	7.4707	0.7660	0.2606	0.7678
SeAFusion	7.5498	0.0644	10.6673	62.9487	0.0390	3.2919	1.6734	0.3270	0.8987
SuperFusion	7.0396	0.0390	10.0081	63.2152	0.0390	3.1650	1.1354	0.1493	0.7413
SwinFuse	7.6055	0.0556	10.7691	62.4752	0.0403	3.3302	0.8475	0.1994	0.9186
TIF	7.1699	0.0498	9.7740	65.3894	0.0229	2.6484	1.2979	0.1330	0.9454
OMOFuse	6.6558	0.0726	10.5918	65.4070	0.0233	4.1534	1.2386	0.0194	0.9570

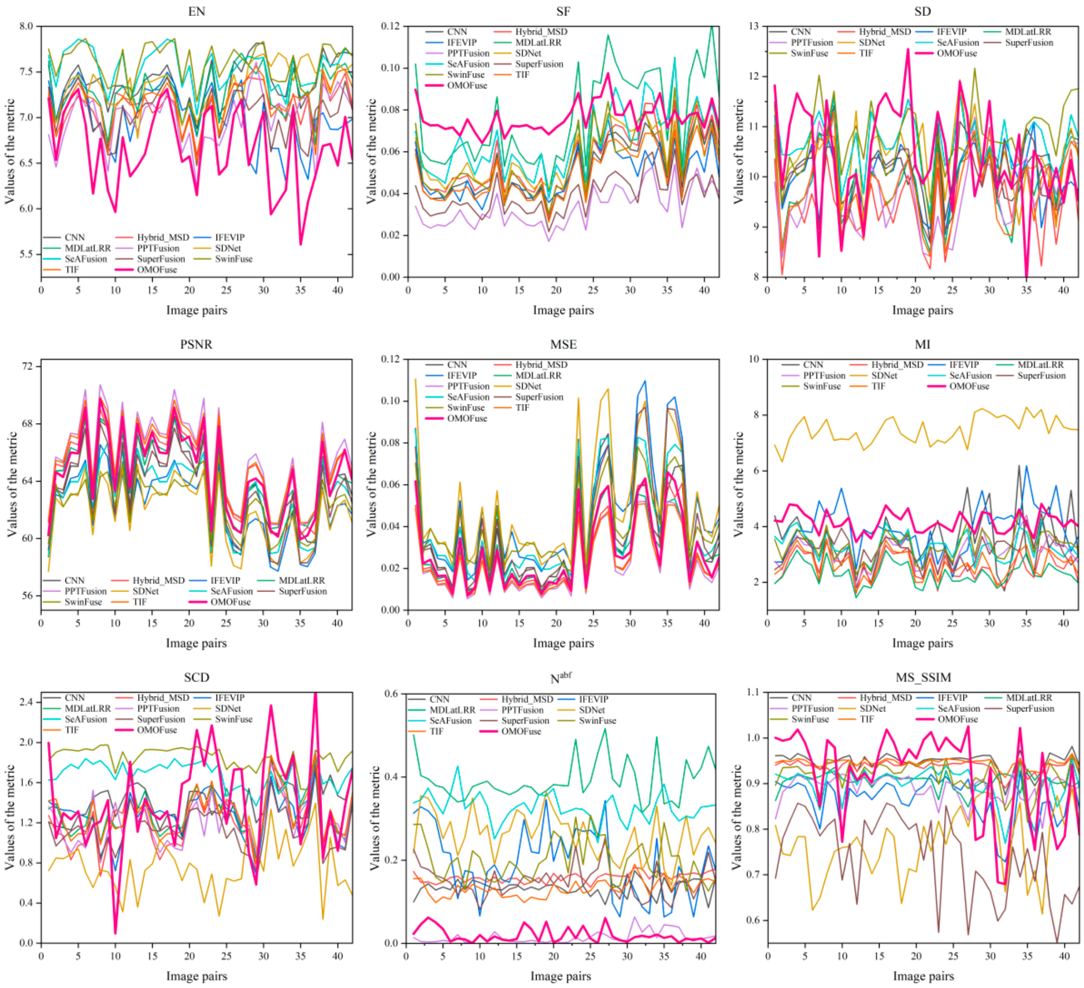

Figure 10. Variation curve graphs of various evaluation metrics for Roadscene datasets, where the OMOFuse algorithm is highlighted and marked in bold red. A quantifiable comparison is conducted with 9 evaluation metrics and 10 different methods.

6. Ablation Experiments

The following text is organized into three main sections: Section 6.1 carries out an ablation study of the ODAM module; Section 6.2 initiates an ablation study of the MO module; and Section 6.3 explores an ablation study of the loss function. Each section undergoes experimental verification, applying both qualitative and quantitative comparisons for a comprehensive analysis.

6.1. Ablation Experiment of ODAM

6.1.1. Qualitative Comparison

Figure 11 showcases the qualitative comparison from ablation studies conducted on each part of the ODAM module. It can be gleaned from Figure 11 that the OMOFuse algorithm demonstrates higher realism on “grassland” than NO_ODAM, NO_DCA, and NO_ESA. NO_ODAM, compared to NO_DCA and NO_ESA, shows a higher degree of blur, thus underscoring the crucial role of the attention mechanism in visually enhancing the sharpness of the fused image. In all images, within the “fire” section on the right, we can readily observe higher clarity with OMOFuse, and the “flame” section is noticeably brighter. This suggests that the OMOFuse algorithm outperforms the other algorithms. Furthermore, the NO_ODAM image is evidently darker than both NO_DCA and NO_ESA, with a lower level of edge information retrieval. Overall, the attention mechanism module plays a significant role in enhancing the algorithm’s clarity, reducing blurriness, and aiding in edge information retrieval. Moreover, the OMOFuse algorithm exhibits superior overall performance.

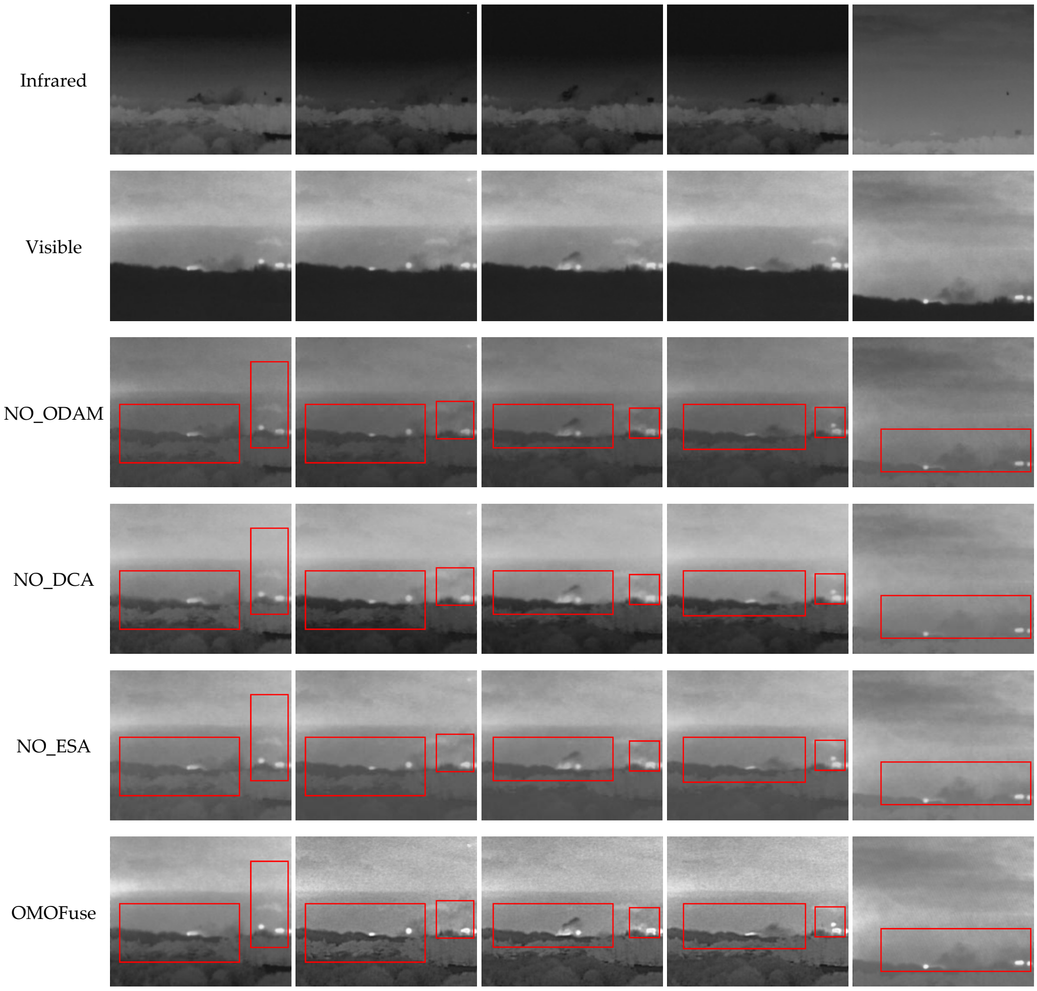

Figure 11. Comparison of visualization effects of the ODAM module ablation experiment. NO_ODAM represents the removal of the ODAM module; NO_DCA represents the removal of the DCA module; NO_ESA represents the removal of the ESA module.

6.1.2. Quantitative Comparison

As Table 4 indicates, it can be observed that the OMOFuse algorithm has a distinct advantage compared to NO_ODAM, NO_DCA, and NO_ESA, precisely aligning with the qualitative comparison shown in the previous figures. NO_ODAM only ranks first in terms of MSE, with the OMOFuse algorithm taking the lead in all the other rankings. Additionally, it can be noted that NO_DCA and NO_ESA enhance the effects of NO_ODAM in different dimensions, further emphasizing that the attention mechanism can improve the fusion performance.

Table 4. ODAM module data comparison.

Method	**EN** ↓	**SF** ↑	**SD** ↑	**PSNR** ↑	**MSE** ↓	**MI** ↑	**SCD** ↑	**N^{abf}** ↓	**MS_SSIM** ↑
NO_ODAM	6.1944	0.0151	9.7224	62.9970	0.0354	3.9826	1.5242	0.5302	0.9252
NO_DCA	6.0814	0.0160	10.6346	62.7961	0.0629	4.7839	1.4794	0.4224	0.9533
NO_ESA	6.3968	0.0094	9.9619	62.3531	0.0639	5.2015	1.5564	0.2036	0.9514
OMOFuse	5.8332	0.0253	10.9961	63.4045	0.0413	5.3462	1.6638	0.1716	0.9713

6.2. Ablation Experiment of MO

6.2.1. Qualitative Comparison

Figure 12 represents the qualitative comparison of the ablation study performed on each section of the ODAM module. As we can discern from Figure 12, despite that NO_MO performs better in terms of clarity of "trees", it still lags behind the OMOFuse algorithm. Furthermore, the clarity of all "fire points" is consistently better with the OMOFuse algorithm than with NO_MO, with sharper "flames" visible on the right side. This further demonstrates that the MO module can enhance the visual performance of the algorithm to a certain extent.

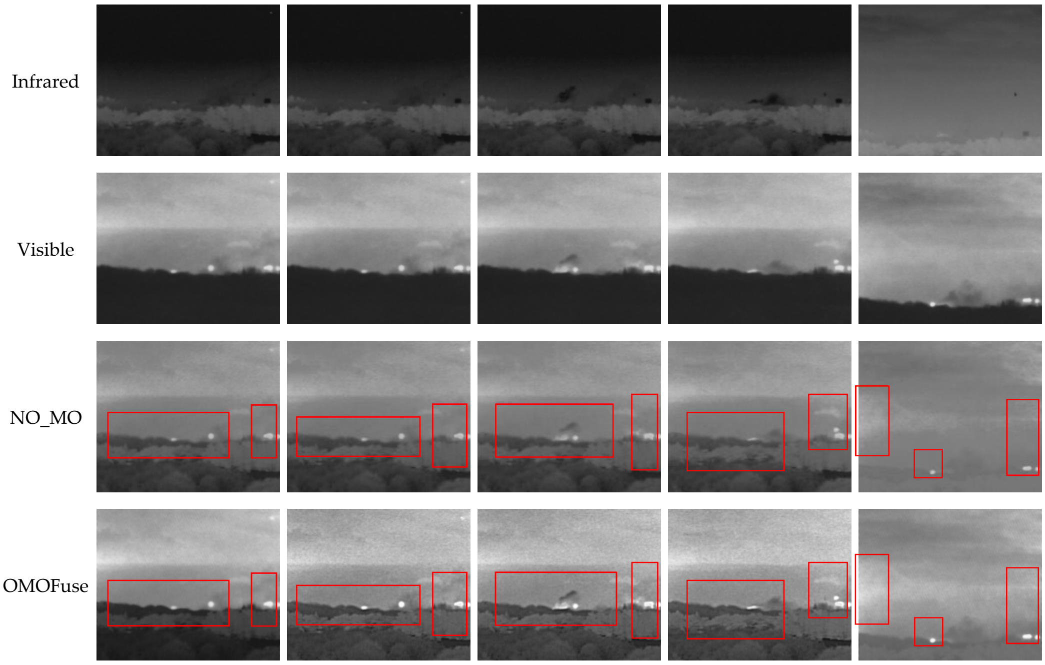

Figure 12. Comparison of visualization effects of the MO module ablation experiment. NO_MO represents the removal of the MO module.

6.2.2. Quantitative Comparison

Table 5 presents the evaluation metrics for both NO_MO and OMOFuse algorithms. Except for the MI and N^{abf} metrics where OMOFuse did not outperform, it leads in the remaining seven evaluation metrics. The overall performance of the OMOFuse algorithm is commendable, mirroring the previous figure, which substantiates that MO can enhance the visual appeal of the network.

Table 5. MO module data comparison.

Method	EN ↓	SF ↑	SD ↑	PSNR ↑	MSE ↓	MI ↑	SCD ↑	N^{abf} ↓	MS_SSIM ↑
NO_MO	6.3368	0.0085	9.5200	61.7127	0.0470	5.4111	1.5337	0.1228	0.9473
OMOFuse	5.8332	0.0253	10.9961	63.4045	0.0413	5.3462	1.6638	0.1716	0.9713

6.3. Ablation Experiment of Loss Function

6.3.1. Qualitative Comparison

Figure 13 presents a qualitative comparison of ablation studies conducted on each part of the ODAM module. It can be seen from Figure 13 that the NO_Loss manifests as black on the organized "fire points", indicating an inferior detection effect. In addition, the result on "tree" texture information-related aspects is also sub-optimal. Thus, the loss function $\mathcal{L}$, introduced by us, could improve the visual effects of the fused images to a certain extent.

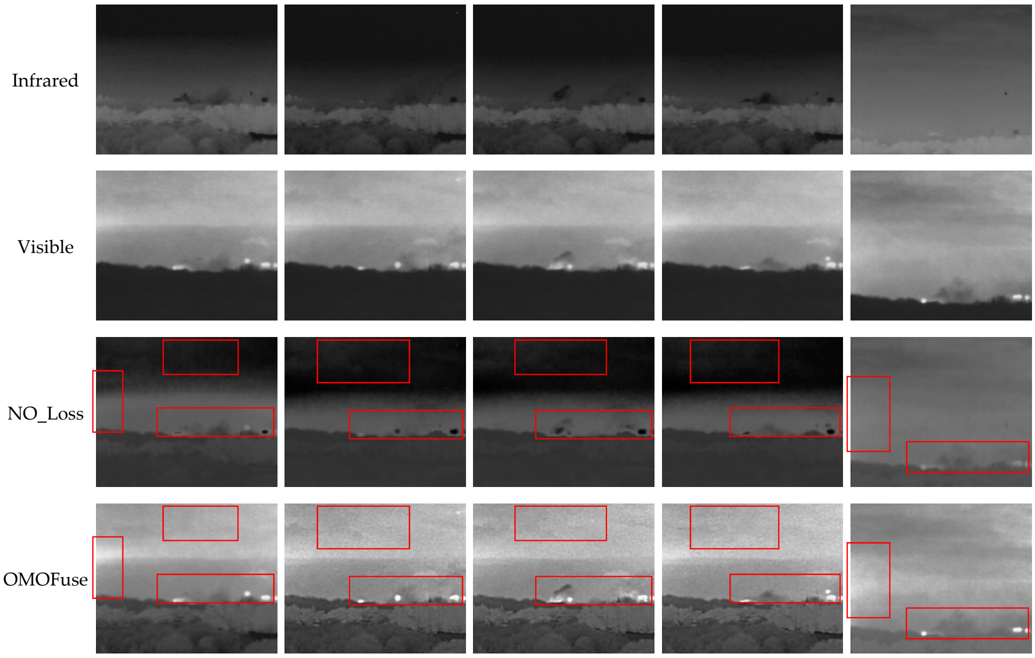

Figure 13. Comparison of visualization effects of the loss function ablation experiment. NO_Loss represents the removal of the loss function.

6.3.2. Quantitative Comparison

Table 6 compiles the evaluation metrics for NO_Loss and OMOFuse. It can be gleaned from Table 6 that, apart from EN and N^{abf} where OMOFuse did not perform optimally, it otherwise does robustly surpass NO_Loss in the remaining rankings. The overall performance is quite satisfactory, further evidencing the effectiveness of the loss function $\mathcal{L}$, as corroborated by the above-mentioned figures.

Table 6. Loss function data comparison.

Method	EN ↓	SF ↑	SD ↑	PSNR ↑	MSE ↓	MI ↑	SCD ↑	N^{abf} ↓	MS_SSIM ↑
NO_Loss	5.6625	0.0053	8.1691	61.8850	0.0507	4.1810	1.1303	0.0595	0.7864
OMOFuse	5.8332	0.0253	10.9961	63.4045	0.0413	5.3462	1.6638	0.1716	0.9713

7. Conclusions

Image fusion can ameliorate image quality to a certain extent, enhance visual effects, and facilitate subsequent processing. Originating from the pixel level, a novel image fusion algorithm called OMOFuse is proposed to avoid the complexity and suboptimal performance issues associated with manually designed algorithms. In detail, a DCA and ESA mechanism are first proposed, which were combined to form the ODAM module for enhancing the feature extraction capability of the network. Subsequently, a MO module was introduced to further improve the network's acquisition of contextual information. Finally, we have constructed a novel loss function L within three dimensions (SSL, PL, and GL). In a qualitative and quantitative comparison across three commonly employed public datasets, five classical algorithms alongside five SOTA ones were contrasted. Experimental outcomes indicate that OMOFuse surpasses both classical and SOTA algorithms in terms of clarity, texture, and visual features.

Author Contributions: J.Y.: theoretical development, system and experimental design, prototype development. S.L.: thesis format checking, experimental design. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the China Postdoctoral Science Foundation (Grant No. 2021M691684).

Data Availability Statement: The source code is available at https://github.com/jyyuan666/OMOFuse (accessed on 13 September 2023).

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

References

1. Parisotto, S.; Calatroni, L.; Bugeau, A.; Papadakis, N.; Schönlieb, C.-B. Variational Osmosis for Non-Linear Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 5507–5516. [CrossRef]
2. Sato, T.; Shimada, S.; Murakami, H.; Watanabe, H.; Hashizume, H.; Sugimoto, M. ALiSA: A Visible-Light Positioning System Using the Ambient Light Sensor Assembly in a Smartphone. *IEEE Sens. J.* **2022**, *22*, 4989–5000. [CrossRef]
3. Hoang, C.M.; Kang, B. Pixel-level clustering network for unsupervised image segmentation. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107327. [CrossRef]
4. Jin, Y.; Dong, Y.; Zhang, Y.; Hu, X. SSMD: Dimensionality Reduction and Classification of Hyperspectral Images Based on Spatial–Spectral Manifold Distance Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5538916. [CrossRef]
5. Su, N.; Chen, X.; Guan, J.; Huang, Y. Maritime Target Detection Based on Radar Graph Data and Graph Convolutional Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4019705. [CrossRef]
6. Chen, T.; Yang, P.; Peng, H.; Qian, Z. Multi-target tracking algorithm based on PHD filter against multi-range-false-target jamming. *J. Syst. Eng. Electron.* **2020**, *31*, 859–870. [CrossRef]
7. Meghdadi, A.H.; Irani, P. Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2119–2128. [CrossRef] [PubMed]
8. Ouyang, Y.; Wang, X.; Hu, R.; Xu, H.; Shao, F. Military Vehicle Object Detection Based on Hierarchical Feature Representation and Refined Localization. *IEEE Access* **2022**, *10*, 99897–99908. [CrossRef]
9. Shen, L.; Rangayyan, R.M. A segmentation-based lossless image coding method for high-resolution medical image compression. *IEEE Trans. Med. Imaging* **1997**, *16*, 301–307. [CrossRef]
10. Sotiras, A.; Davatzikos, C.; Paragios, N. Deformable Medical Image Registration: A Survey. *IEEE Trans. Med. Imaging* **2013**, *32*, 1153–1190. [CrossRef]
11. Bai, Y.; Tang, M. Object Tracking via Robust Multitask Sparse Representation. *IEEE Signal Process. Lett.* **2014**, *21*, 909–913.
12. Shi, Y.; Li, J.; Zheng, Y.; Xi, B.; Li, Y. Hyperspectral Target Detection with RoI Feature Transformation and Multiscale Spectral Attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5071–5084. [CrossRef]
13. 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. [CrossRef]
14. Li, S.; Kang, X.; Hu, J. Image Fusion with Guided Filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875. [PubMed]
15. Bavirisetti, D.P.; Xiao, G.; Zhao, J.; Dhuli, R.; Liu, G. Multi-scale Guided Image and Video Fusion: A Fast and Efficient Approach. *Circuits Syst. Signal. Process* **2019**, *38*, 5576–5605. [CrossRef]
16. Butakoff, C.; Frangi, A.F. A Framework for Weighted Fusion of Multiple Statistical Models of Shape and Appearance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1847–1857. [CrossRef] [PubMed]
17. Xia, Y.; Kamel, M.S. Novel Cooperative Neural Fusion Algorithms for Image Restoration and Image Fusion. *IEEE Trans. Image Process.* **2007**, *16*, 367–381. [CrossRef] [PubMed]
18. Du, J.; Li, W.; Tan, H. Three-layer image representation by an enhanced illumination-based image fusion method. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 1169–1179. [CrossRef]
19. Huang, Y.; Song, R.; Xu, K.; Ye, X.; Li, C.; Chen, X. Deep Learning-Based Inverse Scattering with Structural Similarity Loss Functions. *IEEE Sens. J.* **2021**, *21*, 4900–4907. [CrossRef]
20. Li, M.; Hsu, W.; Xie, X.; Cong, J.; Gao, W. SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising With Self-Supervised Perceptual Loss Network. *IEEE Trans. Med. Imaging* **2020**, *39*, 2289–2301. [CrossRef]
21. Balamurali, A.; Feng, G.; Lai, C.; Tjong, J.; Kar, N.C. Maximum Efficiency Control of PMSM Drives Considering System Losses Using Gradient Descent Algorithm Based on DC Power Measurement. *IEEE Trans. Energy Convers.* **2018**, *33*, 2240–2249. [CrossRef]
22. 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]
23. Vs, V.; Valanarasu, J.M.J.; Oza, P.; Patel, V.M. Image Fusion Transformer. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3566–3570.
24. Li, X.; Wen, J.-M.; Chen, A.-L.; Chen, B. A Method for Face Fusion Based on Variational Auto-Encoder. In Proceedings of the 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 14–16 December 2018; pp. 77–80.
25. Namhoon, L.; Wongun, C.; Paul, V.; Christopher Bongsoo, C.; Philip, H.S.T.; Manmohan Krishna, C. Desire: Distant Future Prediction in Dynamic Scenes with Interacting Agents. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2165–2174.

26. Jin, X.; Hu, Y.; Zhang, C.-Y. Image restoration method based on GAN and multi-scale feature fusion. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 2305–2310.
27. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef] [PubMed]
28. 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]
29. Li, H.; Wu, X.-J.; Durrani, T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [CrossRef]
30. Xu, H.; Wang, X.; Ma, J. DRF: Disentangled representation for visible and infrared image fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5006713. [CrossRef]
31. Jian, L.; Yang, X.; Liu, Z.; Jeon, G.; Gao, M.; Chisholm, D. Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5002215. [CrossRef]
32. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [CrossRef]
33. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518. [CrossRef]
34. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An Infrared and Visible Image Fusion Network Based on Salient Target Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5009513. [CrossRef]
35. Liu, S.; Pan, J.; Yang, M.-H. Learning recursive filters for low-level vision via a hybrid neural network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 560–576.
36. Zhang, J.; Pan, J.; Ren, J.; Song, Y.; Lau, R.W. Dynamic scene deblurring using spatially variant recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2521–2529.
37. Ren, W.; Liu, S.; Ma, L.; Xu, Q.; Xu, X.; Cao, X.; Du, J.; Yang, M.-H. Low-light image enhancement via a deep hybrid network. *IEEE Trans. Image Process.* **2019**, *28*, 4364–4375. [CrossRef]
38. Xu, M.; Tang, L.; Zhang, H.; Ma, J. Infrared and visible image fusion via parallel scene and texture learning. *Pattern Recognit.* **2022**, *132*, 108929. [CrossRef]
39. 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]
40. 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] [PubMed]
41. 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* **2021**, *23*, 1383–1396. [CrossRef]
42. 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.* **2021**, *70*, 5005014. [CrossRef]
43. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
44. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9355–9366.
45. Liu, X.; Suganuma, M.; Sun, Z.; Okatani, T. Dual residual networks leveraging the potential of paired operations for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7007–7016.
46. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 14141–14152.
47. Kim, H.; Park, J.; Lee, C.; Kim, J.J. Improving accuracy of binary neural networks using unbalanced activation distribution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7862–7871.
48. Hanna, M.H.; Kaiser, A.M. Update on the management of sigmoid diverticulitis. *World J. Gastroenterol.* **2021**, *27*, 760. [CrossRef]
49. Wang, E.; Yu, Q.; Chen, Y.; Slamu, W.; Luo, X. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Inf. Fusion* **2022**, *88*, 78–85. [CrossRef]
50. Toet, A. The TNO Multiband Image Data Collection. *Data Brief* **2017**, *15*, 249–251. [CrossRef]
51. Hou, L.; Chen, C.; Wang, S.; Wu, Y.; Chen, X. Multi-Object Detection Method in Construction Machinery Swarm Operations Based on the Improved YOLOv4 Model. *Sensors* **2022**, *22*, 7294. [CrossRef] [PubMed]
52. Ascencio-Cabral, A.; Reyes-Aldasoro, C.C. Comparison of Convolutional Neural Networks and Transformers for the Classification of Images of COVID-19. Pneumonia and Healthy Individuals as Observed with Computed Tomography. *J. Imaging* **2022**, *8*, 237. [CrossRef] [PubMed]
53. Zhang, X.; Ye, P.; Xiao, G. VIFB: A visible and infrared image fusion benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 104–105.

54. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolution Inf. Process.* **2018**, *16*, 1850018. [CrossRef]
55. Zhang, Y.; Zhang, L.; Bai, X.; Zhang, L. Infrared and visual image fusion through infrared feature extraction and visual information preservation. *Infrared Phys. Technol.* **2017**, *83*, 227–237. [CrossRef]
56. Li, H.; Wu, X.-J.; Kittler, J. MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef] [PubMed]
57. Fu, Y.; Xu, T.; Wu, X.; Kittler, J. PPT Fusion: Pyramid Patch Transformerfor a Case Study in Image Fusion. *arXiv* **2022**, arXiv:2107.13967.
58. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [CrossRef]
59. Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; Ma, J. SuperFusion: A Versatile Image Registration and Fusion Network with Semantic Awareness. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 2121–2137. [CrossRef]
60. Wang, Z.; Chen, Y.; Shao, W.; Li, H.; Zhang, L. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5016412. [CrossRef]
61. Bavirisetti, D.P.; Dhuli, R. Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Phys. Technol.* **2016**, *76*, 52–64. [CrossRef]
62. Alexander, T. TNO Image Fusion Dataset. *Data Brief* **2017**, *15*, 249–251. [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.