applied sciences

MDPI

Article

AFSFusion: An Adjacent Feature Shuffle Combination Network for Infrared and Visible Image Fusion

Yufeng Hu [1], Shaoping Xu [2,*], Xiaohui Cheng [2], Changfei Zhou [2] and Minghai Xiong [2]

[1] The School of Qianhu, Nanchang University, Nanchang 330031, China; 5903120029@email.ncu.edu.cn
[2] School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China; 406100210094@email.ncu.edu.cn (X.C.); 409100220054@email.ncu.edu.cn (C.Z.); 419100220136@email.ncu.edu.cn (M.X.)
* Correspondence: xushaoping@ncu.edu.cn

Abstract: To obtain fused images with excellent contrast, distinct target edges, and well-preserved details, we propose an adaptive image fusion network called the adjacent feature shuffle-fusion network (AFSFusion). The proposed network adopts a UNet-like architecture and incorporates key refinements to enhance network architecture and loss functions. Regarding the network architecture, the proposed two-branch adjacent feature fusion module, called AFSF, expands the number of channels to fuse the feature channels of several adjacent convolutional layers in the first half of the AFSFusion, enhancing its ability to extract, transmit, and modulate feature information. We replace the original rectified linear unit (ReLU) with leaky ReLU to alleviate the problem of gradient disappearance and add a channel shuffling operation at the end of AFSF to facilitate information interaction capability between features. Concerning loss functions, we propose an adaptive weight adjustment (AWA) strategy to assign weight values to the corresponding pixels of the infrared (IR) and visible images in the fused images, according to the VGG16 gradient feature response of the IR and visible images. This strategy efficiently handles different scene contents. After normalization, the weight values are used as weighting coefficients for the two sets of images. The weighting coefficients are applied to three loss items simultaneously: mean square error (MSE), structural similarity (SSIM), and total variation (TV), resulting in clearer objects and richer texture detail in the fused images. We conducted a series of experiments on several benchmark databases, and the results demonstrate the effectiveness of the proposed network architecture and the superiority of the proposed network compared to other state-of-the-art fusion methods. It ranks first in several objective metrics, showing the best performance and exhibiting sharper and richer edges of specific targets, which is more in line with human visual perception. The remarkable enhancement in performance is ascribed to the proposed AFSF module and AWA strategy, enabling balanced feature extraction, fusion, and modulation of image features throughout the process.

Keywords: infrared and visible image fusion; adjacent feature shuffle fusion; adaptive weight adjustment strategy; subjective and objective evaluation

check for updates

Citation: Hu, Y.; Xu, S.; Cheng, X.; Zhou, C.; Xiong, M. AFSFusion: An Adjacent Feature Shuffle Combination Network for Infrared and Visible Image Fusion. *Appl. Sci.* **2023**, *13*, 5640. https://doi.org/10.3390/app13095640

Received: 6 April 2023
Revised: 25 April 2023
Accepted: 27 April 2023
Published: 3 May 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

In recent years, with advancements in sensor-imaging technologies, cameras have gained the ability to capture both infrared (IR) and visible images. IR images are effective in capturing target objects under low-illumination conditions by leveraging the difference between the target object and environmental thermal radiation. However, these images often lack sufficient information on the edges and texture details of the target. On the other hand, visible images can capture high-resolution texture details. Therefore, researchers have become interested in the optimal integration of complementary information from IR and visible images through image fusion techniques to compensate for sensor limitations and enhance human and machine perception [1–5]. The fusion of IR and visible images

has a wide range of applications, including intelligent surveillance, target monitoring, and video analysis. This technology enables subsequent image-processing tasks and assists in more comprehensive and intuitive analysis or decision-making. Image fusion techniques have evolved into two main categories: traditional methods and deep learning-based methods for both IR and visible images.

Traditional methods for image fusion can generally be categorized into three stages, namely feature extraction, feature fusion, and image reconstruction [1–3]. Typically, the fusion method employed in this case involves converting the image from the spatial domain into a domain that enables the extraction of features, which are then integrated using specific principles. The features extracted from the source images are then fused using fusion techniques. Finally, the fused features are converted back to the spatial domain representation to achieve image fusion. For instance, in the study by Jin et al. [4], the discrete stationary wavelet transform (DSWT) was utilized to decompose significant features of source images into subgraph collections of varying levels and spatial resolutions. The significant details are then separated using the discrete cosine transform (DCT) into subgraph sets based on the energy of the individual frequencies. The local spatial frequency (LSF) is used to enhance the regional features of DCT coefficients. To combine the significant features in the source images, a segmentation strategy based on the LSF threshold is employed, resulting in DCT coefficient fusion. Finally, the fused image is reconstructed by performing the inverse transform of DCT and DSWT sequentially. More recently, Ren et al. [5] proposed a non-subsampled shearlet transform (NSST)-based multiscale entropy fusion method for IR and visible images. The approach initially employs an NSST-based multiscale method to decompose the source images to acquire both high and low-frequency information. The high and low-frequency data are then fused at different scales, and the weight values are determined using multiscale entropy. Finally, the fusion results at different scales are obtained via an inverse NSST transformation to generate the fused images. These aforementioned studies demonstrate that effectively extracting features from both IR and visible images and designing appropriate fusion rules are crucial for addressing the fusion problem through traditional fusion methods. The design of reasonable fusion rules is essential to maximizing the preservation of complementary content between the two images in the output image. However, these tasks rely on manual experience, and the fusion result has significant limitations and cannot adapt to complex scene changes.

The enhancement of deep learning techniques [6] over the last decade and their successful application to the field of low-level vision processing have resulted in the development of various types of network models for IR and visible image fusion by researchers in this field [7–19]. Based on the stage of IR and visible image feature fusion, this process can generally be classified into the following three types [20]: (1) early fusion [7–9], which involves merging the original images at a low-level feature extraction stage. For instance, an end-to-end fusion model using detail-preserving adversarial learning (DPAL) techniques was proposed by Ma et al. [7]. The middle section consisting of Conv+BN+PReLU+Conv+BN repeats five times, where BN represents batch normalization. The back section contains a 3×3 convolutional layer, a PReLU, and a 1×1 convolutional layer. DPAL uses an advanced adversarial learning technique that has proved to be highly effective across multiple fields. This approach can be implemented on diverse image data types and can rapidly train high-quality models in a short duration. However, this kind of method tends to produce fused images of a generic nature, as the network architecture lacks appropriate mechanisms to facilitate image feature fusion, ultimately failing to fully leverage the complementary nature of IR and visible images. This inadequate fusion of image features often leads to an excessive redundancy in input feature information, complicating image reconstruction tasks in later stages. (2) Medium fusion [10–13]. Li et al. [11] proposed a network model called dense fusion (DenseFuse) for the fusion of IR and visible images, which achieved image fusion by cascading feature maps of multiple images of the input. Compared to other image fusion methods, DenseFuse retains detailed information in the source images, while

facilitating superior robustness and scalability. This method reduces the noise and produces a visually appealing output image. However, it has also been observed to lead to a loss of image information and reduced spatial resolution. The applied network architecture, which uses medium-term feature fusion, initially extracts source image features before proceeding with the fusion. Lastly, the fused features are reconstructed into the final fused image using a relatively complex feature reconstruction module. Although the fusion framework in this architecture successfully prevents the introduction of redundant information compared to earlier feature fusion architectures, the feature fusion module is relatively simple and may not fuse features effectively. Consequently, image feature fusion is inadequate, leading to an inferior image reconstruction quality post-fusion. Overall, this approach is not optimal and can be improved. (3) Late fusion [14–19]. In late fusion, the input images are first processed separately to extract relevant features, and these features are then fused to create a final image. Tang et al. [14] proposed a novel light-awareness-based network designed to manage luminance distribution for salient target retention while preserving texture in the background. This approach fully incorporates complementary and public information within the inherent constraints of light awareness. The initial design involves a 1×1 convolutional layer to minimize modal discrepancies between the IR and visible images. Four convolutional layers with shared weights are then implemented to extract deeper IR and visible image features, using leaky rectified linear unit (LeakyReLu) activations within each layer of the feature extractor. These extracted features are then concatenated and supplied as input to the image reconstructor, composed of five convolutional layers. This module works towards integrating common and complementary information to produce a fused image. Each of the convolutional layers, except the last one, has a kernel size of 3×3, while the last layer has a kernel size of 1×1. LeakyReLu activations are used in most of the layers, with the last layer adopting the Tanh activation. Given that late image fusion merges multiple images, it can lead to distorted images if not carefully performed, and fusion occurs in the later part of the network, resulting in a relatively weak processing ability in the image reconstruction module. Thus, several limitations should be addressed in the final fusion effect.

Based on the aforementioned studies, IR and visible image fusion models can be readily developed using end-to-end training methods supported by a large amount of training data. Generally speaking, various complex image feature representations, extraction designs, and fusion rules that heavily rely on human expertise, are replaced by numerous network parameters, whose values are determined mainly by the training dataset without the need for human intervention. Therefore, fusion models based on deep learning are relatively easy to design and use. In these methods, the network architecture design and the combination of loss functions play a vital role in improving performance. The network architecture design determines the potential ability to extract features from IR and visible light images and the subsequent fusion, while the choice of loss function determines whether the performance of the fused image can be optimized. However, there is still room for improvement in both network architecture and loss functions in current methods.

In this work, we present an adjacent feature shuffle combination network (AFSFusion) as a solution to maximize the preservation of complementary information in both IR and visible images, as well as to obtain fused images that display the highest quality and resolution across various objective evaluation metrics. The primary contributions of our proposed network are as follows:

- Our AFSFusion network differs from the resolution transformation of the UNet [21], wherein the network is specifically designed to optimize image feature processing for image fusion. We establish a hybrid fusion network to achieve fusion by scaling up and down several feature channels based on the improved UNet-like architecture. The first half of the network uses the adjacent feature shuffle-fusion (AFSF) module. This module fuses the features of each adjacent convolutional layer several times using shuffle operations, resulting in the full interaction of the input feature information. Consequently, information loss during feature extraction and fusion is significantly

reduced, and the complementarity of feature information is enhanced. The proposed network architecture effectively integrates complementary features using the AFSF module, achieving a balanced and collaborative approach to feature extraction, fusion, and modulation. This method is typically not found in traditional methods, and it has resulted in improved fusion performance. These changes allow the proposed network to achieve excellent performance in terms of subjective and objective evaluation.

- Three loss functions, mean square error (MSE), structural similarity (SSIM), and total variation (TV), are used to construct the total loss function; to enable the total loss function to play a significant role in network training, an adaptive weight adjustment (AWA) strategy is designed to automatically set the weight values of the different loss functions according to extracted feature responses for the IR and visible light, which are based on the image content (the magnitude of the feature response values are extracted and calculated using VGG16 [22]). As such, the fused images produced by the fusion network can achieve the best possible fusion result that aligns with the visual perceptual characteristics of the human eye.

The remainder of this study is organized as follows: Section 2 offers a brief overview of previous research on the convolutional neural network backbone architecture, UNet. Section 3 provides a thorough explanation of the proposed AFSFusion, comprising the model architecture, backbone network, and loss functions. In Section 4, we present several experiments performed to demonstrate the superiority of the model and illustrate its excellent performance compared to other alternatives based on subjective and objective analyses. The main conclusions are outlined in Section 5.

2. Related Work

UNet

The UNet backbone network, proposed by Ronneberger et al. [23], is a classical end-to-end network that exploits deep networks, residual learning, and regularization techniques. The UNet architecture adopts a conventional encoder–decoder network architecture that has a remarkable learning ability to preserve high-level semantic information within the pre-trained backbone (i.e., the encoder), while leveraging its decoder to provide detailed, fine-grained information. As shown in Figure 1, the UNet architecture comprises two primary components: the contracting and the expansive paths. The contracting path is responsible for capturing image context information, whereas the expansive path is responsible for precise image reconstruction. The contracting path includes a sequence of convolutional layers followed by max-pooling layers that reduce the spatial resolution while simultaneously increasing the number of channels in the feature maps. Conversely, the expansive path comprises multiple deconvolutional layers that enhance spatial resolution while reducing the number of channels. The use of skip connections enables the contracting and expansive paths to connect their corresponding layers, permitting the network to reuse learned features from the contracting path during the reconstruction stage. The primary benefits of the UNet architecture include its ability to handle images of different resolutions and scales, capture both local and global feature representations of input images, and increase the effective size of the training set owing to its ability to manage samples with limited data. However, since the UNet architecture lacks simultaneous feature extraction and fusion mechanisms, it can only achieve early or late fusion when used for the fusion of IR and visible images, while a more effective medium fusion cannot be directly achieved. Therefore, the baseline network is selected for the proposed fusion network, and significant modifications are made to enable the fusion network architecture to successfully attain feature extraction and fusion. These modifications are introduced and elaborated on in the following sections.

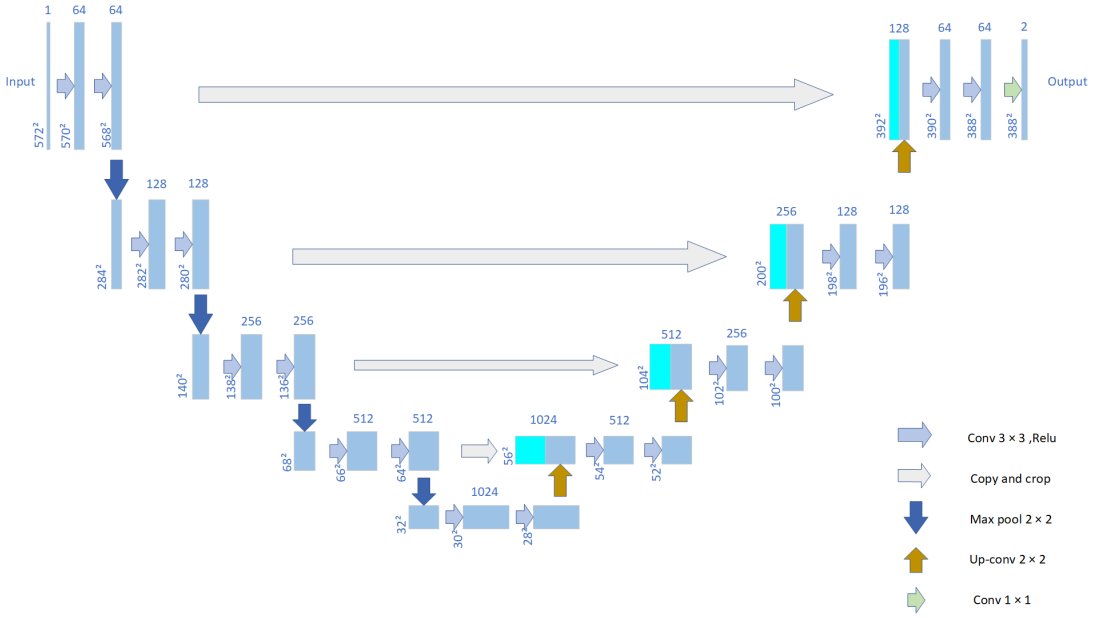


Figure 1. The standard U-Net framework.

3. AFSFusion Network Model

3.1. Basic Idea

The proposed AFSFusion network model, as depicted in Figure 2, is engineered to process both IR and visible images. It trains a considerable number of parameters in the deep convolutional neural network and generates the output image, commonly referred to as the fused image. To construct the total loss function, the model adopts three renowned types of loss functions, namely, MSE, SSIM, and TV. An AWA strategy is devised to enable each loss function to automatically set the weight values allocated to the IR and visible images. The strategy utilizes a VGG16 network model to extract features at various network levels for both the IR and visible images. The feature response values are then computed, normalized, and employed to generate the corresponding weight values, denoted by ω_{ir} for the IR images and ω_{vis} for the visible images. This facilitates the preservation of the most significant texture details in the IR or visible light in the final fused image, rendering the generated images more consistent with human eye perception. During the development process, the model constantly evaluates its performance on a validation set, and the architecture and parameters are adjusted as needed.

3.2. Backbone Network

In theory, state-of-the-art convolutional networks, such as DnCNN [24], DenseNet [25], and ResNet [26] can be employed to construct a backbone network for the AFSFusion model. As previously mentioned, current state-of-the-art image fusion models, comprising early, medium, and late fusion strategies, exhibit varying levels of deficiencies across the three stages of feature extraction, feature fusion, and fused image reconstruction. This results in suboptimal fusion outcomes. Consequently, this study introduces a novel fusion network that can simultaneously extract and fuse intricate image features in the first half of the network. Moreover, fully fused features are further modulated during complex reconstruction processes in the rear half section of the network. Thus, the fusion process is ubiquitous throughout the entire network architecture, ensuring that the ultimate fusion result is of high image quality and accords with human eye perception.

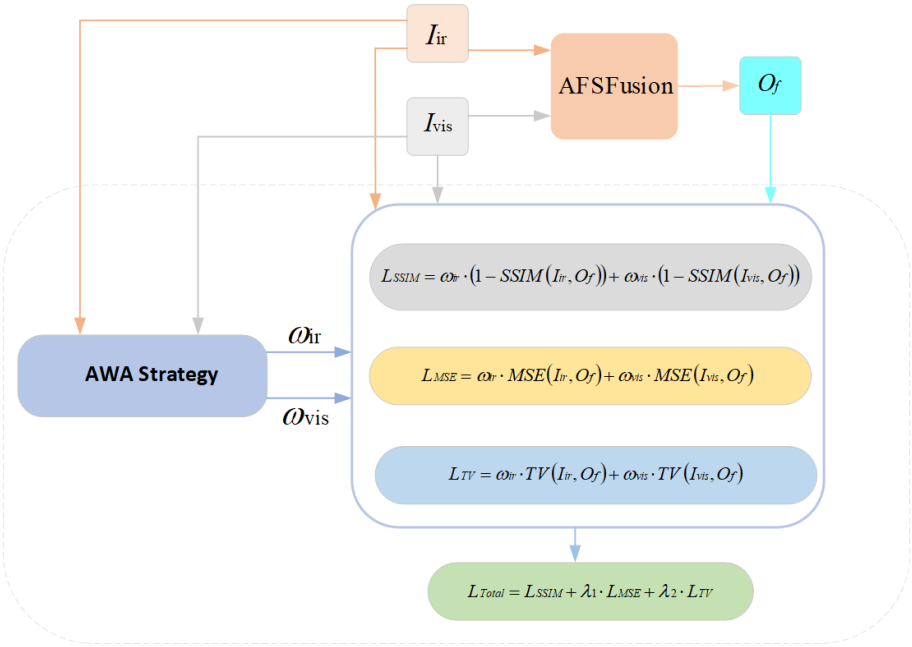


Figure 2. The AFSFusion network framework.

The proposed network utilizes channel expansion and contraction to achieve a balanced and collaborative approach to feature extraction, fusion, and modulation during the process of feature propagation. As shown in Figure 3, the AFSFusion network adopts a UNet-like network architecture. The major differences between the AFSFusion network and UNet are:

1. The inputs. It is modified in a dual-channel format to enable the system to receive and process both IR and visible images simultaneously.
2. The elimination of the skip connection. Unlike traditional U-Net, our network utilizes channel expansion and contraction instead of resolution expansion and contraction. The purpose of skip connections is to transfer information between the encoder and decoder, which helps to address issues of information loss and inaccurate semantic information during segmentation. As our network has a relatively high input image resolution and does not use downsampling, feature information is not lost during the fusion process.
3. The addition of the AFSF module enhances feature interaction and information transfer capability. In the first half of UNet, the feature channels are sequentially passed from front to back, which limits their information interaction ability. Therefore, the AFSF module proposed in this study fuses feature information from the previous convolutional layer with the output of the current convolutional layer, which is then passed to the next convolutional layer, and adds a shuffle operation to facilitate feature interaction. The details of the specific module are presented in Section 3.3.
4. The elimination of the maximum pool layer. Typically, the max pooling layer is employed to reduce dimensionality, compress features, lower network complexity, and decrease the number of required parameters. However, since the AFSFusion network primarily utilizes the AFSF module to enhance feature channels, the use of only the max pooling layer can result in ambiguity regarding the number of channels. The depth of the network architecture is relatively shallow, while the number of channels is relatively large. The max pooling layer is not utilized to alleviate computational overhead.
5. The replacement of the ReLU layer. To solve the problem in which the ReLU layer cannot handle negative values, it is replaced with a Leaky ReLU layer to address the gradient disappearance problem.

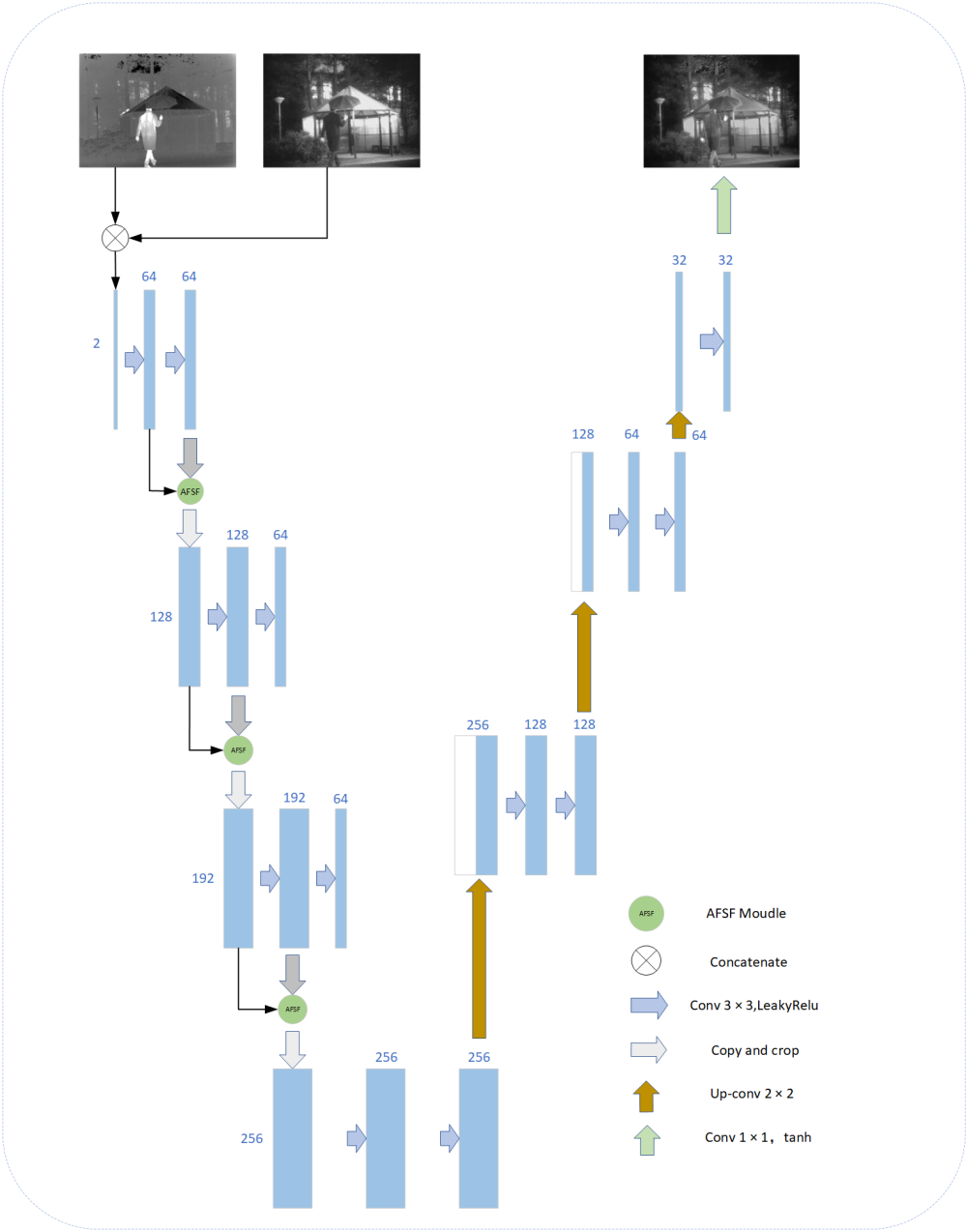


Figure 3. AFSFusion algorithm backbone network framework.

To ensure the efficacy of feature extraction and fusion, the AFSF module is introduced to operate prior to the middle section of the AFSFusion network, enabling the sequential introduction of feature information from the IR and visible images into the subsequent network layers, thereby preserving all feature information. The internal architecture of the AFSF module comprises a double-branching channel and a shuffling operation, facilitating the fusion and interaction of feature information with the double-branching structure, ensuring maximum fusion and transmission of feature information. The feature channels at the rear of the network are directly modified using the convolutional layer, which modulates the fused features to continually reduce the number of features to reconstruct the complete fused image. Thus, AFSFusion aims to achieve a balanced execution of the three sub-tasks of feature extraction, feature fusion, and image reconstruction, ensuring the production of a fused image of exceptional quality. Compared to state-of-the-art fusion networks, the topology and implementation complexity of AFSFusion are relatively low. However, with appropriate network architecture and loss function design, its fusion capability is

significantly superior to those of other state-of-the-art fusion methods in all objective evaluation metrics, as described in the comparative experimental data section in Section 4.

3.3. AFSF Module

The figures depicted in Figure 4 represent the number of feature channels utilized in the network model, commencing from 2 (representing the IR and visible images for fusion) and progressively increasing up to 256. Subsequently, there is a layer-by-layer reduction before eventually reverting back to a single channel that corresponds to the fused image. In the front and middle sections of the network, image features are sequentially extracted and fully fused before being passed on to the next layer. In the middle and rear sections of the network, the fused network features are reconstructed layer-by-layer before being passed on to the next layer to form the fused image. Thus, the proposed network model ensures a methodical and well-balanced execution of the three critical tasks, namely, feature extraction, feature fusion, and fused image reconstruction. Notably, the AFSFusion network is simpler to construct than previous fusion networks in terms of the overall backbone network. Our investigation results demonstrate that enhancing performance in image fusion networks does not necessitate a complex network architecture and can be achieved using a well-designed network architecture. Detailed results for our comparative experiments are presented in Section 4 for further reference.

As shown in Figure 4, the proposed AFSF module in this study employs two branches for feature extraction and fusion. The first branch uses two parallel convolutional layers to extract the feature information from the two inputs, and each convolutional layer consists of convolution and activation functions. Then, three consecutive convolutional layers are used to modulate the extracted feature information. In the second branch, the two inputs are directly concatenated and then fed into convolutional layers for feature extraction. Finally, the features extracted from the two branches are added together and shuffled to enable effective interaction between them. The shuffle operation is applied at the end of the module to enhance the network's mapping and feature processing capabilities, thereby improving the degree of feature information extraction and achieving better training results. In addition, the channels of the input data are shuffled by the shuffle operation, allowing each channel to receive information from different groups, providing a powerful mechanism for feature fusion, enhancing the diversity and representativeness of features, ultimately improving the generalization performance of the model.

3.4. Loss Function

3.4.1. Selection of the Loss Function

To obtain a loss function with superior guidance capability, this study adopts three loss functions: MSE, SSIM, and TV, and sets the corresponding weights to form the total loss function. The MSE loss function restricts the level of dissimilarity between the images at the pixel level, but it does not visually match human eye perception. Therefore, SSIM, which measures the local structure of the images in terms of luminance, contrast, and structure, is selected as a loss function to better visualize the fused images. Additionally, the TV loss function is introduced to constrain the gradient information of the image, which directly reflects the contour and edge information of the object in an image. This is important for evaluating the quality of the fused image in IR and visible image fusion, as the clarity of the object contour depicted in the image is crucial. Variable weight values are set in the different loss functions for IR and visible images according to the differences in the human eye perception characteristics of IR and visible images. Different weight values are utilized in the three types of loss functions, and the specific effects can be elucidated in the analysis of the ablation experiments in Section 4.

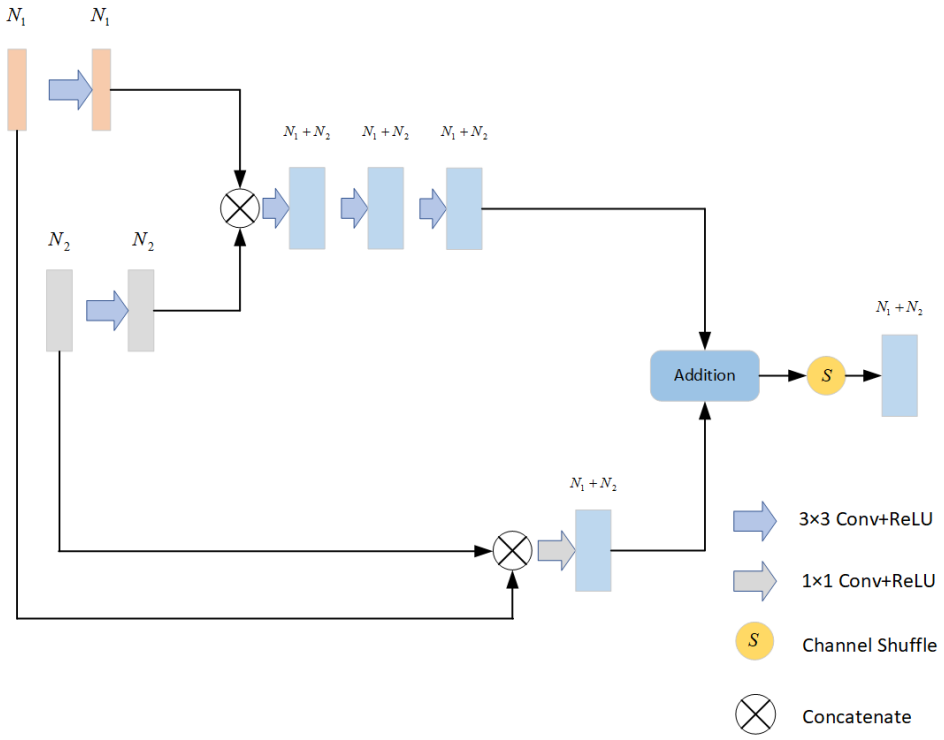


Figure 4. AFSF module architecture.

In particular, the MSE, SSIM, and TV functions are calculated, as shown using Equations (1)–(3), where $\|\cdot\|_2$ in Equation (1) represents the two-norm operation, and μ_x,μ_y,σ_x^2, and σ_y^2 in Equation (2) represent the mean, variance, and covariance of x and y, respectively. The parameters c_1 and c_2 are constants; ∂ and $\|\cdot\|_1$ in Equation (3) represent the gradient and one-norm operation, respectively.

$$MSE(x,y) = \|x-y\|_2 \tag{1}$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\sigma_x^2 + \sigma_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{2}$$

$$TV(x,y) = \|\partial x - \partial y\|_1 \tag{3}$$

3.4.2. AWA Strategy

As the AFSFusion network performs the image fusion task, it adjusts the weight values of the corresponding pixels of the IR and visible images in the final fused image. To extract the input IR and visible images separately, we propose using the VGG16 network model originally designed for image classification. Specifically, we extract feature responses that reflect the image content at each network layer and calculate weight coefficients adaptively for the MSE, SSIM, and TV loss functions. These weight coefficients are based on the response intensities of the two sets of images on the same feature layer. The weight coefficients for the IR and visible images of each loss function are determined based on the feature responses obtained from the five convolutional layers before the max-pooling layer of the VGG16 model. The feature response on the i-th convolutional layer is defined as $F^i(\cdot)$ and, hence, the feature responses for both the IR and visible images at the i-th convolutional layer can be denoted as $F^i(I_{ir})$ and $F^i(I_{vis})$, respectively. Based on this, the Laplace feature response values on the i-th convolutional layer are given by:

$$R^i(\cdot) = \frac{l}{H_i W_i C} \sum_{k=1}^{C} \|\nabla F_k^i(\cdot)\|_F^2 \tag{4}$$

where H_i and W_i denote the length and width of the feature in the i-th layer, respectively; C denotes the number of channels in the feature layer; $F_k^i(\cdot)$ is the feature map of the k-th channel in the i-th feature layer, $\|\cdot\|_F$ denotes the F-norm, and ∇ denotes the Laplacian operation.

The feature response values of each layer for the IR and visible images are then averaged using weights that correspond to the importance of each layer, and the final total feature response value of the image is obtained as follows:

$$R(\cdot) = \sum_{i=1}^{5} R^i(\cdot)/5 \tag{5}$$

Based on the preceding definition, the overall feature response values $R(I_{ir})$ and $R(I_{vis})$ of the IR and visible images are normalized to obtain the final weight coefficient values, ω_{ir} and ω_{vis}, respectively:

$$\begin{aligned} \omega_{ir} &= \frac{R(I_{ir})}{R(I_{ir}) + R(I_{vis})} \\ \omega_{vis} &= \frac{R(I_{vis})}{R(I_{ir}) + R(I_{vis})} \end{aligned} \tag{6}$$

where the normalized weight coefficients ω_{ir} and ω_{vis} have values between [0, 1] and add to 1. Based on the weight coefficient values ω_{ir} and ω_{vis}, the total loss function is defined as:

$$L_{Total} = L_{SSIM} + \lambda_1 \cdot L_{MSE} + \lambda_2 \cdot L_{TV} \tag{7}$$

where λ_1 and λ_2 are hyperparameters that are used to adjust the weights of the loss functions for different distances. The MSE, SSIM, and TV loss functions are expressed in Equations (8)–(10):

$$L_{MSE} = \omega_{ir} \cdot MSE(I_{ir}, O_f) + \omega_{vis} \cdot MSE(I_{vis}, O_f) \tag{8}$$

$$L_{SSIM} = \omega_{ir} \cdot (1 - SSIM(I_{ir}, O_f)) + \omega_{vis} \cdot (1 - SSIM(I_{vis}, O_f)), \tag{9}$$

$$L_{TV} = \omega_{ir} \cdot TV(I_{ir}, O_f) + \omega_{vis} \cdot TV(I_{vis}, O_f) \tag{10}$$

4. Experimental Results and Analysis

4.1. Experimental Setup

To evaluate the performance of the AFSFusion network model, we compared it with nine fusion methods reported in the literature, namely, GTF [27], SDNet [28], DCHWT [29], DenseFuse [11], PMGI [30], MUFusion [31], U2Fusion [32], SwinFusion [33], and RFN-Nest [10]. The test images were obtained from the TNO dataset [34] and VOT2020-RGBT dataset [35], and the test sets comprised 21 and 34 groups of images, respectively. The objective evaluation metrics used for the test results are entropy (En) [36], standard deviation (SD) [37], mutual information (MI) [38], multi-scale structural similarity (MS-SSIM) [39], visual quality fidelity (VIF) [40], and no-reference image quality metrics for contrast distortion (NIQMC) [41], with higher values for all metrics indicating better fusion. In addition, fused images are subjectively evaluated via manual visual inspection. The values in all tables represent the average values for each method for each dataset. For comparison, the best metrics obtained are presented in bold for each method. All the different methods were tested using the same hardware platform (Intel(R) Xeon(R) CPU E5-1603 v4 @ 2.80GHz RAM 16GB) and software environment (Windows 11 operating system).

4.2. Training Details

The AFSFusion network proposed in this study was trained on the modulated RoadScene dataset, with 40% of the images used for validation. The training data were shuffled

and divided into 18 batches. For each batch, training images were randomly selected from either the IR or visible light sources, and the images were randomly cropped into 64×64 patches and normalized to values between 0 and 1. The AFSFusion parameters were updated using the AWA strategy, which performed weight allocation operations. The learning rate was initialized as 1×10^{-4} and then exponentially decayed. Based on the ablation experiment, hyperparameters controlling the weights of each sub-loss term were set to $\lambda_1 = 1.5$ and $\lambda_2 = 1$, and the training process was iterated three times. After each batch of training, the performance of AFSFusion was evaluated using the validation set. Throughout the training process, the AWA strategy utilized the VGG16 to extract features from the IR and visible light images at different network layers. After calculating the corresponding feature values and normalizing them, the strategy generated corresponding weight values ω_{ir} and ω_{vis} for IR and visible light images, respectively. These weight values play an important regulatory role in constructing each loss function and can maximize the preservation of salient texture details in both the IR and visible light images in the fused image.

4.3. Ablation Studies

This section demonstrates the impact of the AWA strategy, the selection of hyperparameter values in the loss function, and evaluates the effectiveness of the AFSF module. All parameters and training results under various network architectures were obtained using the TNO dataset.

4.3.1. Ablation Study Involving the AWA Strategy

Table 1 provides an objective comparison of the ablated model without the AWA strategy and the complete AFSFusion network, highlighting the vital role of the AWA strategy in enhancing the effectiveness of the network. The symbol N denotes the case where the equivalent values of ω_{ir} and ω_{vis} in the loss function are set to 0.5 (no adaptive adjustment), and Y denotes the adaptive generation of the weight coefficients using the strategy. The experimental results suggest that the AWA strategy adapts by generating weight coefficients according to the content of the fused image, leading to optimization of the ability of the loss function to guide network parameters, ultimately resulting in high-quality fused images. It can be concluded that the AWA strategy plays a crucial role in ensuring that the fused images obtained are of excellent quality.

Table 1. Ablation study involving the use of the AWA strategy.

Strategy	En	SD	MI	MS-SSIM	VIF	NIQMC
N	6.8756	81.0549	13.7512	**0.9286**	0.7917	4.5564
Y	**7.1589**	**95.9510**	**14.3179**	0.9053	**0.8786**	**4.8369**

4.3.2. Ablation Study Involving Loss Items

Table 2 showcases the quantitative outcomes of the ablation analysis focused on the loss function. The weights of each sub-loss function involved in the combination were set to 1. From Table 2, we can observe that the MSE is relatively effective when a single type of loss function is used as the network loss function. The additional use of the TV function can lead to a substantial improvement in the fusion effect for the six metrics. Although further addition of the SSIM function resulted in a decrease in the effect for the MS-SSIM metric, improved results were obtained for the other five metrics. Therefore, the composition of the loss function of the AFSFusion network takes the form of a weighted combination of the aforementioned three types of loss functions.

Table 2. Ablation study involving the adoption of loss functions.

Selection	En	SD	MI	MS-SSIM	VIF	NIQMC
SSIM	6.7674	75.7708	13.5347	0.8763	0.8093	4.4303
MSE	6.8755	75.8543	13.7509	0.8893	0.7812	4.4854
TV	6.5573	59.3614	13.1146	0.8997	0.6567	3.9791
SSIM + MSE	6.9162	81.9746	13.8323	0.9121	0.8140	4.5599
SSIM + TV	6.9104	79.5288	13.8207	0.9052	0.7800	4.6015
MSE + TV	6.9942	83.7423	13.9884	**0.9236**	0.8143	4.6324
SSIM + MSE + TV	**7.0708**	**90.3205**	**14.1415**	0.9004	**0.8784**	**4.7150**

After determining the selection of the loss functions, for the best fusion result-based experiments, the optimal values of the hyperparameters λ_1 and λ_2 in the loss functions were determined using the grid search method, and the results are shown in Table 3. Table 3, shows that five of the six metrics performed optimally when λ_1 was 1.5 and λ_2 was taken as 1. MS-SSIM also works better for this metric; hence, we set the values of λ_1 and λ_2 to 1.5 and 1, respectively.

Table 3. Comparison of the results of different weight settings for λ_1 and λ_2.

λ_1 and λ_2 Values	En	SD	MI	MS-SSIM	VIF	NIQMC
(0.5, 0.5)	6.8913	78.2546	13.7825	0.9299	0.8018	4.4835
(0.5, 1)	6.8717	75.9035	13.7433	0.9043	0.7970	4.4580
(0.5, 1.5)	6.8782	76.0264	13.7563	0.9196	0.7929	4.4458
(1, 0.5)	6.7920	71.9125	13.5841	0.8869	0.7875	4.3727
(1, 1)	7.0708	90.3205	14.1415	0.9004	0.8784	4.7150
(1, 1.5)	6.9237	80.4903	13.8475	0.9292	0.8150	4.5627
(1.5, 0.5)	7.0926	93.318	14.1852	**0.9258**	0.8492	4.8003
(1.5, 1)	**7.1589**	**95.9510**	**14.3179**	0.9053	**0.8786**	**4.8369**
(1.5, 1.5)	6.8789	76.6077	13.7577	0.9122	0.8157	4.4652

4.3.3. Ablation Study Involving AFSF Module

The AFSF module is the key component of the proposed network, through which a feature extraction and fusion mechanism are facilitated, allowing for greater interaction of the feature information; moreover, it improves the quality of the fused image. Compared to AFSFusion-w/o-AFSF, the full model exhibited significantly improved results for all five metrics except for MS-SSIM. In addition, by removing the shuffle operation from the AFSFusion, the AFSFusion-w/o-shuffle decreases the interactivity of the feature information, which also reduces the quality of the fused image in terms of the evaluation metrics according to the evaluation in Table 4.

Table 4. Quantitative results of the ablation study involving the AFSF module.

Methods	En	SD	MI	MS-SSIM	VIF	NIQMC
w/o-AFSF	6.8926	81.0629	13.7852	**0.9320**	0.7992	4.5464
w/o-shuffle	6.8945	83.8112	13.7890	0.9248	0.7847	4.5976
AFSFusion	**7.1589**	**95.9510**	**14.3179**	0.9053	**0.8786**	**4.8369**

4.4. Analysis of Results

4.4.1. Objective Evaluation

To comprehensively and objectively compare the fusion performance of the AFSFusion network, each fusion method was tested on two benchmark test collections. The results are shown in Tables 5 and 6, whereas Figures 5 and 6 illustrate the performance of each fusion method on each of the evaluation metrics.

The AFSFusion model achieved 10 optimal values and 2 suboptimal values for each evaluation metric in both databases, while U2Fusion and SwinFusion methods attained

two optimal values. These results demonstrate the superior performance of the AFSFusion network over other fusion methods, highlighting its overall effectiveness. SwinFusion, SDNet, and MUFusion, despite being recently proposed methods, exhibited inferior fusion performance compared to the proposed method.

Table 5. Comparison of fusion results for each fusion method using the TNO dataset.

Models	SwinFusion	DenseFuse	PMGI	U2Fusion	MUFusion	SDNet	GTF	AFSFusion
En	6.83	6.38	6.97	6.90	6.71	6.57	6.72	**7.16**
SD	93.03	59.79	77.70	76.91	71.13	62.96	76.74	**95.95**
VIF	**0.90**	0.69	0.84	0.79	0.73	0.73	0.63	0.88
MI	13.66	12.76	13.94	13.79	13.43	13.16	13.44	**14.32**
MS-SSIM	0.90	0.88	0.88	**0.92**	0.90	0.86	0.82	0.91
NIQMC	4.52	3.77	4.59	4.48	4.23	4.07	4.29	**4.84**

Table 6. Comparison of the fusion results for each fusion method using the VOT2020 dataset.

Models	SwinFusion	DenseFuse	PMGI	U2Fusion	MUFusion	SDNet	GTF	AFSFusion
En	6.83	6.38	6.96	6.96	6.68	6.71	6.77	**7.11**
SD	89.19	55.32	76.69	76.78	65.47	70.79	83.04	**95.08**
VIF	**0.92**	0.68	0.85	0.80	0.72	0.75	0.62	0.89
MI	13.66	12.75	13.92	13.91	13.37	13.42	13.53	**14.22**
MS-SSIM	0.90	0.88	0.89	**0.93**	0.88	0.88	0.82	0.89
NIQMC	4.51	3.75	4.58	4.55	4.18	4.23	4.37	**4.80**

4.4.2. Subjective Evaluation

To visually evaluate the fusion results of the fused images obtained for each method, a subjective comparison of artificial vision was performed, and the results are shown in Figures 7 and 8.

The results presented in this study are based on the region defined by the red rectangular box in Figure 7. The target contours in the images fused using the GTF, PMGI, RFN-Nest, and SwinFusion methods appear blurred. While the DCHWT, DenseFuse, and U2Fusion methods avoid this problem, the fused images are not clear. Results are further analyzed in the region defined by the yellow box in Figure 7, where most compared methods are found to be strongly influenced by the IR image, resulting in dark fusion outcomes for the wheel part. Both U2Fusion and SDNet results are strongly influenced by visible light, resulting in the target and building appearing white, which does not match the visual perception characteristics of the human eye. The AFSFusion model produces the best visual results. A comprehensive comparison with state-of-the-art algorithms reveals that AFSFusion has better fusion effects, with fused image targets being clearer and more consistent with human visual characteristics. As shown in Figure 8, results are further observed in the area defined by the blue box of each fusion result. The GTF, RFN-Nest, and SDNet methods show a significant loss of contour information, while several methods, such as DenseFuse, SwinFusion, PMGI, U2Fusion, and MUFusion, achieve better fusion results but still suffer from IR image whitening. Additionally, the DCHWT fusion results contain many artifacts that severely degrade the visual effect. In conclusion, the AFSFusion network model outperforms state-of-the-art methods by producing fused images with clearer target contours and higher levels of texture detail.

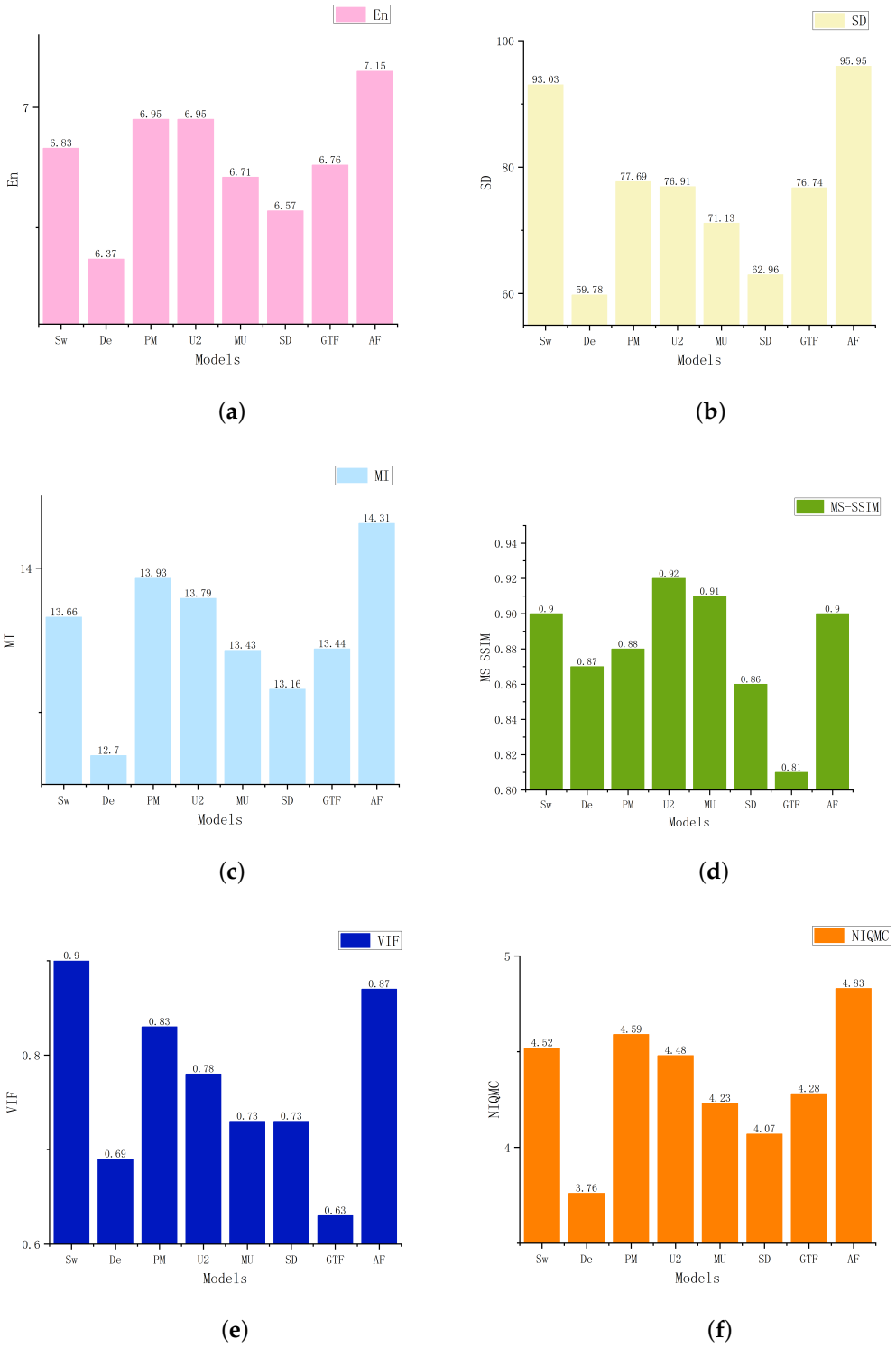


Figure 5. Comparison of fusion results of each fusion method for the TNO dataset. The following are acronyms for different method models: SwinFusion (Sw), DenseFuse (De), PMGI (PM), U2Fusion (U2), MUFusion (MU), SDNet (SD), GTF (GTF), and AFSFusion (AF). (**a**) En. (**b**) SD. (**c**) MI. (**d**) MS-SSIM. (**e**) VIF. (**f**) NIQMC.

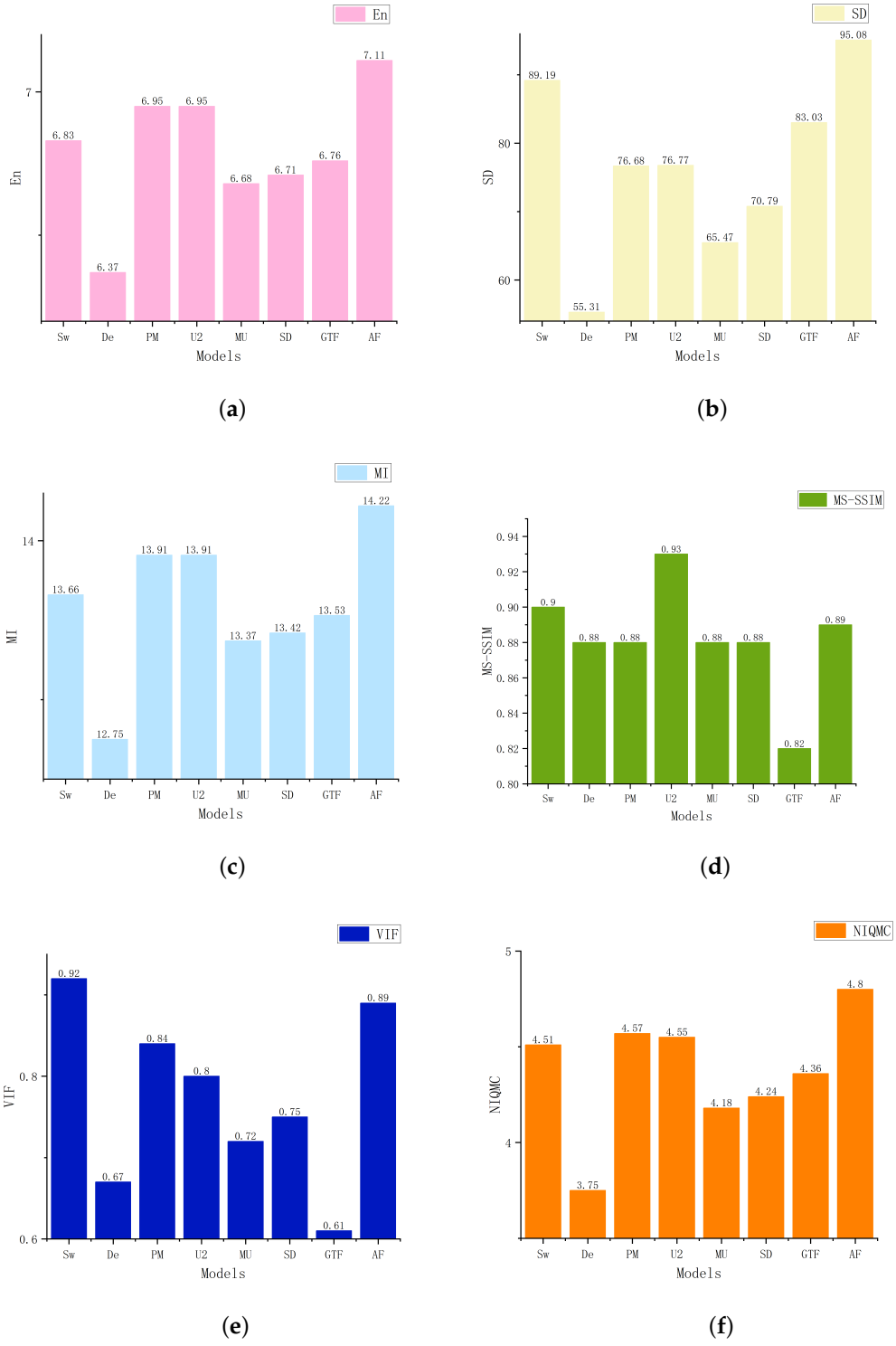


Figure 6. Comparison of the fusion results for each fusion method using the VOT2020 dataset. The following are acronyms for different method models: SwinFusion (Sw), DenseFuse (De), PMGI (PM), U2Fusion (U2), MUFusion (MU), SDNet (SD), GTF (GTF), and AFSFusion (AF). (**a**) En. (**b**) SD. (**c**) MI. (**d**) MS-SSIM. (**e**) VIF. (**f**) NIQMC.

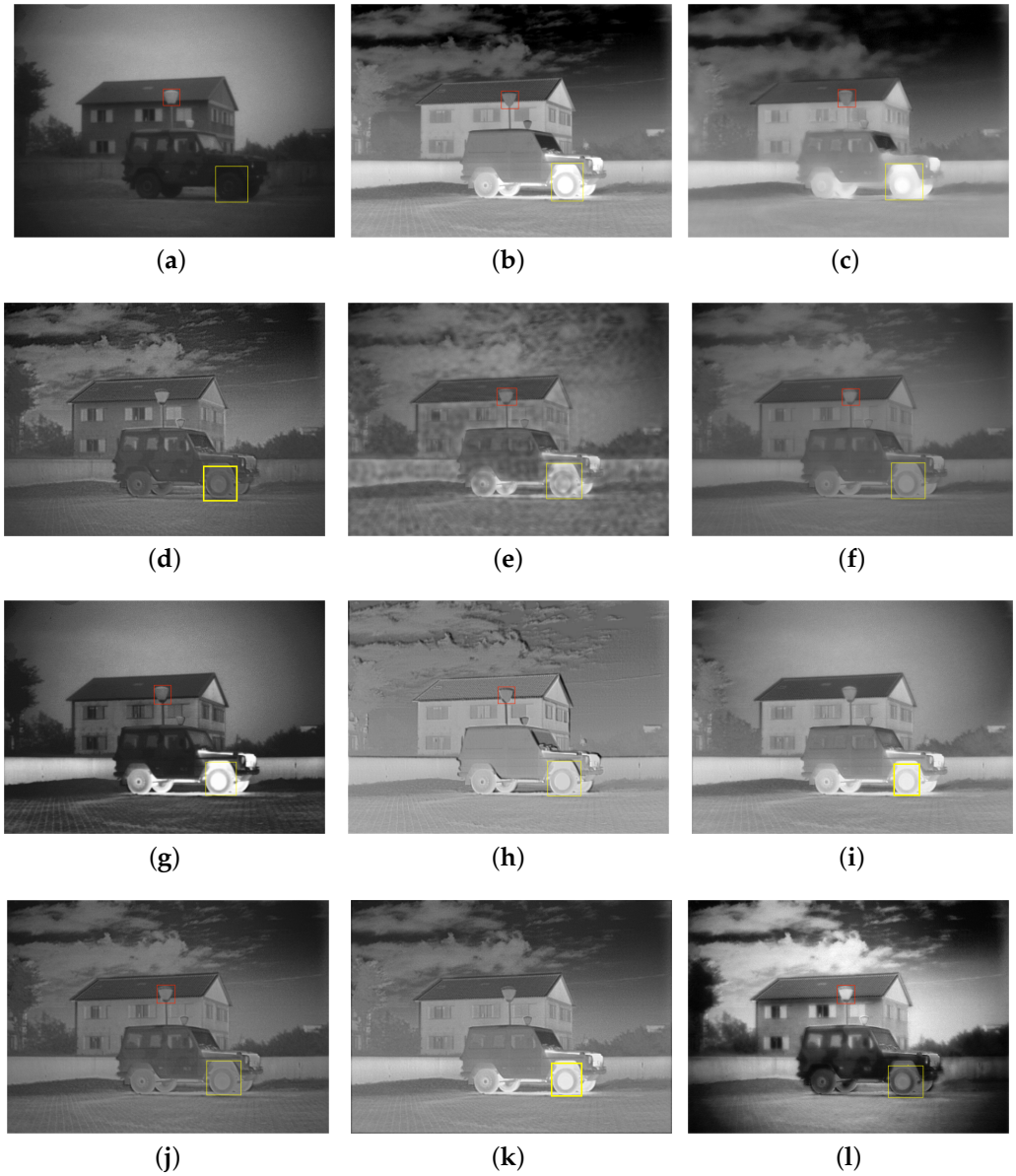


Figure 7. Comparison results of subjective evaluations using the car image from the TNO dataset. (**a**) Visible image. (**b**) IR image. (**c**) GTF. (**d**) MUFusion. (**e**) DCHWT. (**f**) DenseFuse. (**g**) RFN-Nest. (**h**) PMGI. (**i**) SwinFusion. (**j**) U2Fusion. (**k**) SDNet. (**l**) AFSFusion.

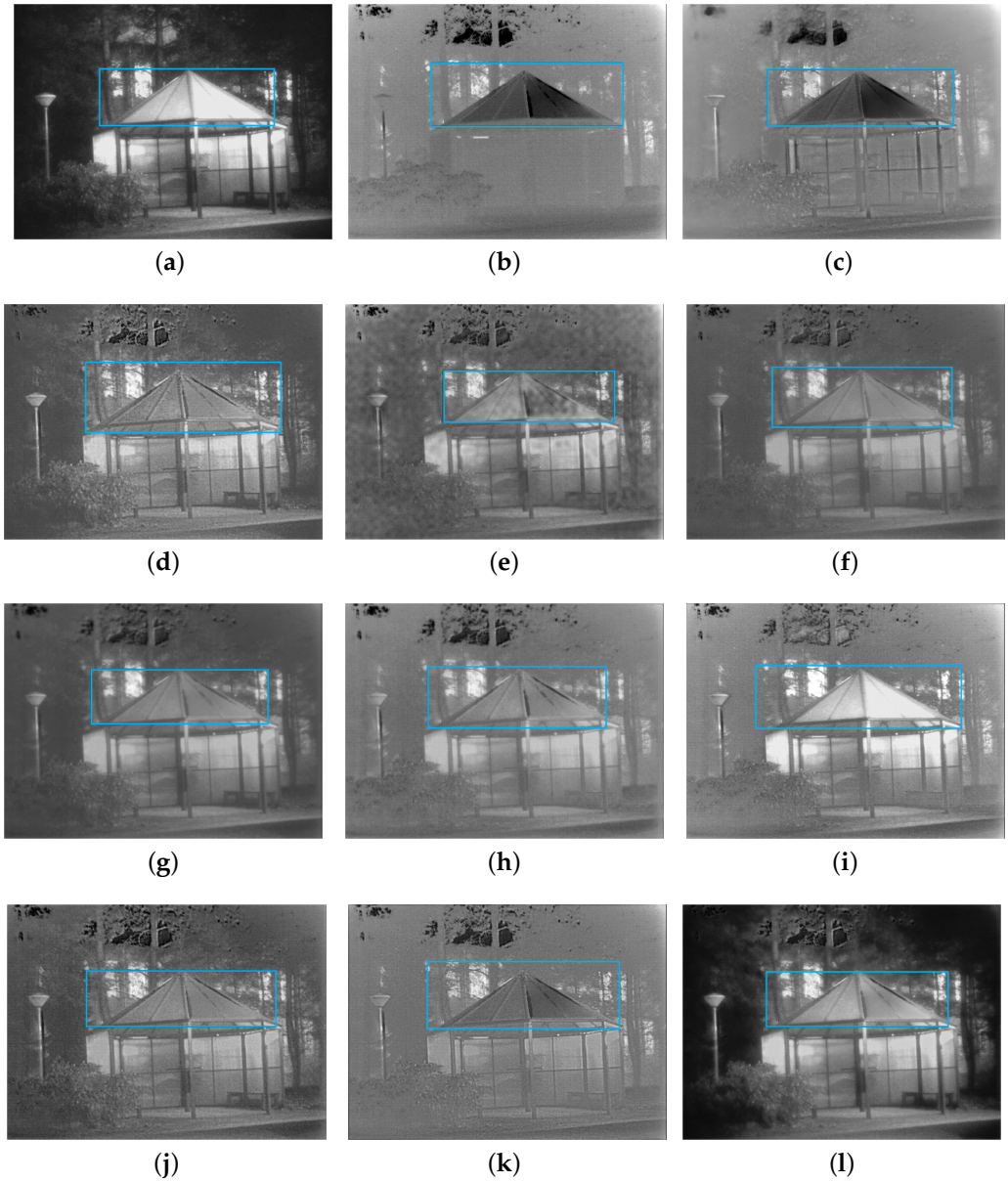


Figure 8. Comparison results of subjective evaluations using the house image from the VOT2020 dataset. (**a**) Visible image. (**b**) IR image. (**c**) GTF. (**d**) MUFusion. (**e**) DCHWT. (**f**) DenseFuse. (**g**) RFN-Nest. (**h**) PMGI. (**i**) SwinFusion. (**j**) U2Fusion. (**k**) SDNet. (**l**) AFSFusion.

5. Conclusions

In this study, we propose a novel adaptive fusion network model for IR and visible images via an adjacent feature fusion approach, called AFSFusion, which employs a UNet-like network architecture. The model uses a designed AFSF module multiple times in the front section of the network for sequential feature extraction, fusion, and shuffling in the module. This allows the input feature information to fully interact in the feature extraction to enhance the complementarity of feature information. To create fused images that closely approximate human visual perception, the fusion network uses three distinct loss functions: MSE, SSIM, and TV. Additionally, an AWA strategy is used to ensure that the corresponding pixels of IR and visible images have adaptive weighting values that optimize the fusion effect. Finally, by exploiting these improvements, the fusion network outputs high-quality images. Extensive experimental results demonstrate that the proposed AFSFusion network model produces better-fused images than other state-of-the-

art fusion methods in both subjective and objective metrics, indicating the effectiveness of the designed network architecture.

In future work, to enhance the overall versatility of the fusion network architecture, we plan to generalize the model to other fusion domains, including multi-exposure and multi-modal medical imaging, visible light and millimeter-wave image fusion, as well as visible light and ultrasound image fusion. Additionally, to improve network performance while reducing computational complexity, we will explore the Restormer structure [42], which introduces local receptive fields to limit the scope of the self-attention mechanism in each Transformer layer [43]. Specifically, This approach can reduce computational costs while preserving spatial positional information, allowing the model to handle high-resolution images. Next, it downsamples the spatial dimensions of the input image to varying degrees, allowing the model to simultaneously process multiscale information. Finally, group linear attention mechanisms and short-range connections between different layers can be used to facilitate gradient propagation and reduce the computational complexity of self-attention. With these changes, our AFSFusion will perform better in more metrics while reducing computational complexity.

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

Funding: This research was funded by the Natural Science Foundation of China, grant number 62162043.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data will be made available upon request.

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

References

1. Son, D.M.; Kwon, H.J.; Lee, S.H. Visible and near-infrared image synthesis using pca fusion of multiscale layers. *Appl. Sci.* **2020**, *10*, 8702. [CrossRef]
2. Zhang, Z.; Cao, Y.; Ding, M.; Tao, J. Fusion of infrared and visible images using multilayer convolution sparse representation. *J. Harbin Inst. Technol.* **2021**, *53*, 51–59.
3. Guan, Z.; Deng, Y.L.; Nie, R.C. Hyperspectral and panchromatic image fusion based on spectral reconstruction-constrained non-negative matrix factorization. *Comput. Sci.* **2021**, *48*, 153–159.
4. 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. [CrossRef]
5. Ren, Y.F.; Zhang, J.M. Fusion of infrared and visible images based on NSST multi-scale entropy. *J. Ordnance Equip. Eng.* **2022**, *43*, 278–285.
6. Aghamohammadi, A.; Ranjbarzadeh, R.; Naiemi, F.; Mogharrebi, M.; Dorosti, S.; Bendechache, M. TPCNN: Two-path convolutional neural network for tumor and liver segmentation in CT images using a novel encoding approach. *Expert Syst. Appl.* **2021**, *183*, 115406. [CrossRef]
7. Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J.; Jiang, J. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion* **2020**, *54*, 85–98. [CrossRef]
8. Poria, S.; Cambria, E.; Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015 ; pp. 2539–2544.
9. Martínez, H.P.; Yannakakis, G.N. Deep multimodal fusion: Combining discrete events and continuous signals. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; pp. 34–41.
10. 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]

11. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
12. Yi, S.; Jiang, G.; Liu, X.; Li, J.; Chen, L. TCPMFNet: An infrared and visible image fusion network with composite auto encoder and transformer–convolutional parallel mixed fusion strategy. *Infrared Phys. Technol.* **2022**, *127*, 104405. [CrossRef]
13. 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.* **2020**, *70*, 1–15. [CrossRef]
14. 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**, *83-84*, 79–92. [CrossRef]
15. 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. [CrossRef]
16. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14), Montreal, QC, Canada, 8–13 December 2014.
17. Wu, D.; Pigou, L.; Kindermans, P.J.; Le, N.D.H.; Shao, L.; Dambre, J.; Odobez, J.M. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1583–1597. [CrossRef]
18. Kahou, S.E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, Ç.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R.C.; et al. Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, Sydney, Australia, 9–13 December 2013; pp. 543–550.
19. Xu, H.; Wang, X.; Ma, J. DRF: Disentangled representation for visible and infrared image fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
20. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]
21. Saadi, S.B.; Ranjbarzadeh, R.; Amirabadi, A.; Ghoushchi, S.J.; Kazemi, O.; Azadikhah, S.; Bendechache, M. Osteolysis: A literature review of basic science and potential computer-based image processing detection methods. *Comput. Intell. Neurosci.* **2021**, *2021*, 4196241. [CrossRef] [PubMed]
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
24. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef]
25. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. 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. [CrossRef]
28. 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]
29. Shreyamsha Kumar, B. Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform. *Signal Image Video Process.* **2013**, *7*, 1125–1143. [CrossRef]
30. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; Ma, J. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12797–12804.
31. Cheng, C.; Xu, T.; Wu, X.J. MUFusion: A general unsupervised image fusion network based on memory unit. *Inf. Fusion* **2023**, *92*, 80–92. [CrossRef]
32. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef] [PubMed]
33. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
34. Toet, A. TNO Image Fusion Dataset. 2022. Available online: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029 (accessed on 23 April 2023).
35. Kristan, M.; Matas, J.; Leonardis, A.; Vojir, T.; Pflugfelder, R.; Fernandez, G.; Nebehay, G.; Porikli, F.; Čehovin, L. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2137–2155. [CrossRef]
36. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522.
37. Rao, Y.J. In-fibre Bragg grating sensors. *Meas. Sci. Technol.* **1997**, *8*, 355. [CrossRef]
38. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 1. [CrossRef]

39. Ma, K.; Zeng, K.; Wang, Z. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.* **2015**, *24*, 3345–3356. [CrossRef]
40. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef] [PubMed]
41. Gu, K.; Lin, W.; Zhai, G.; Yang, X.; Zhang, W.; Chen, C.W. No-reference quality metric of contrast-distorted images based on information maximization. *IEEE Trans. Cybern.* **2016**, *47*, 4559–4565. [CrossRef] [PubMed]
42. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

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.