



Article

DCFusion: Dual-Headed Fusion Strategy and Contextual Information Awareness for Infrared and Visible Remote Sensing Image

Qin Pu ¹, Abdellah Chehri ² , Gwanggil Jeon ^{3,*} , Lei Zhang ¹ and Xiaomin Yang ¹¹ College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China² Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON K7K 7B4, Canada³ Department of Embedded Systems Engineering, Incheon National University Academyro-119, Incheon 22012, Republic of Korea

* Correspondence: gjeon@inu.ac.kr

Abstract: In remote sensing, the fusion of infrared and visible images is one of the common means of data processing. Its aim is to synthesize one fused image with abundant common and differential information from the source images. At present, the fusion methods based on deep learning are widely employed in this work. However, the existing fusion network with deep learning fails to effectively integrate common and differential information for source images. To alleviate the problem, we propose a dual-head fusion strategy and contextual information awareness fusion network (DCFusion) to preserve more meaningful information from source images. Firstly, we extract multi-scale features for the source images with multiple convolution and pooling layers. Then, we propose a dual-headed fusion strategy (DHFS) to fuse different modal features from the encoder. The DHFS can effectively preserve common and differential information for different modal features. Finally, we propose a contextual information awareness module (CIAM) to reconstruct the fused image. The CIAM can adequately exchange information from different scale features and improve fusion performance. Furthermore, the whole network was tested on MSRS and TNO datasets. The results of extensive experiments prove that our proposed network achieves good performance in target maintenance and texture preservation for fusion images.

Keywords: image fusion; infrared image; visible image; target maintenance; texture preservation



Citation: Pu, Q.; Chehri, A.; Jeon, G.; Zhang, L.; Yang, X. DCFusion: Dual-Headed Fusion Strategy and Contextual Information Awareness for Infrared and Visible Remote Sensing Image. *Remote Sens.* **2023**, *15*, 144. <https://doi.org/10.3390/rs15010144>

Academic Editors: Jiaming Na, Hu Ding, Yingchun Fu and Fang Qiu

Received: 13 November 2022

Revised: 17 December 2022

Accepted: 20 December 2022

Published: 27 December 2022



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Due to certain limitations in the theory and technology of hardware devices [1], images obtained by a single sensor can not adequately display scene information. Therefore, image fusion technologies come into being and its purpose is to integrate the meaningful information for different modal images. Image fusion includes multi-modal image and digital photographic image fusion because of the existence of modal differences. Images from different sources of the same scene contain a wealth of complementary information. Moreover, image fusion has been widely applied in a wide range computer vision tasks, such as military operations, object detection [2], object tracking [3], pedestrian recognition [4] and semantic segmentation [5]. Due to the different imaging mechanisms of the sensors, the scene information for infrared and visible images has certain differences in contrast and texture. The infrared images can hold a certain level of sharpened thermal targets, but other targets that do not generate heat are easily overlooked. Conversely, visible images provide more texture detail and better visual performance. So, fusing the two type of images can obtain more all-round information compared with a single image, which is very meaningful for the previously mentioned tasks.

At present, many fusion methods already exist, both in traditional methods [6–11] and deep-learning-based methods. Through principles of mathematics, the traditional

fusion method mainly transforms the source image to the transform domain. Further, it measures the activity level and designs fusion strategies in the domain to integrate information. The fusion methods include five categories, i.e., multi-scale decomposition-based methods [12,13], subspace clustering-based methods [14], sparse representation-based methods [15], optimization-based methods [16] and hybrid methods [17]. However, the development of traditional fusion algorithms has entered a bottleneck period. Firstly, the transformation or representations methods used by traditional methods are becoming increasingly complex, which does not meet the needs of real-time computer applications [18]. Secondly, the measurement of hand-crafted activity level and fusion methods cannot adapt complex scenarios. Moreover, the heat radiation information of infrared images is characterized with pixel intensity and the detail information of visible images is described by edges and gradients. Since different source images have different image characteristics, the traditional fusion method may lead to loss of feature diversity and the fused image may produce certain artifacts. These drawbacks make the design of fusion rules in the traditional way more difficult and complex.

To alleviate the shortcomings of traditional methods, fusion algorithms with deep learning are rapidly emerging. The methods include the three key elements: feature extraction, fusion strategy and data-driven training. In deep learning, feature extraction is used to obtain deep representations of the source image by multiple convolution layers. Different fusion strategies also have significant influence on the fused image. In [15,19], the decision map of the source images are acquired by sparse representation and neural network. The fused images are obtained by using the decision map and suitable post-processing. Although these methods have achieved certain effects, the design of fusion strategies and post-processing is relatively difficult. Therefore, some end-to-end convergence frameworks are proposed, i.e., FusionGAN [20], FusionGANv2 [8] and DDcGAN [21]. Based on adversarial learning, the framework overcomes shortcomings such as artificial features and fusion strategies. However, these methods fail to adequately preserve image details. To better protect the detail information of visible images, the RFN-Nest [22] was proposed. However, the multi-scale fusion method is simple and fails to adequately integrate complementary information for source images.

To alleviate the problems mentioned earlier, we propose the DCFusion that includes three parts: encoder module, fusion strategy and decoder module. The encoder module is employed to extract different scale features of source image. A dual-headed fusion strategy (DHFS) as the fusion module is provided to preserve common and differential information from intermediate features adequately. Furthermore, we propose a contextual information awareness module (CIAM) as the decoder to stack different scale features.

In general, our contributions are the following:

- (1) We propose the DCFusion can preserve more significant information by a suitable dual-headed fusion strategy and reconstruction method.
- (2) The dual-headed fusion strategy (DHFS) is designed to integrate different modal features. In addition to preserving common information, the fusion strategy also allows for the effective integration of complementary information for source images by differential information compensation.
- (3) We propose a contextual information awareness module (CIAM) to merge different scale features and generate result. The module achieve more competitive reconstruction performance by exchanging information for different scale feature.

2. Related Works

In the following, we present some existing fusion methods that are the most relevant to our work, including traditional image fusion algorithms, AE-based image fusion algorithms, CNN-based image fusion algorithms and GAN-based image fusion algorithms.

2.1. Traditional Image Fusion Algorithms

The fusion methods includes three main elements: feature extraction, feature fusion and feature reconstruction. The feature extraction and feature fusion are more important compared with feature reconstruction. To achieve more effective feature representation, many feature extraction methods was proposed such as multi-scale transforms, sparse representation and subspace clustering. In these feature extraction methods, the multi-scale transforms divide image features into different scales and combine the features in separated levels. So far, there are already classical multi-scale transforms such as discrete wavelet [23], shearlet [24], latent low-rank representation [12] and nonsubsampling contourlet transform [25]. Beyond multi-scale transforms, sparse representation [15] is embedded in fusion framework to extract features, which represent images using sparse encoding in a complete dictionary. Moreover, subspace clustering, such as independent component analysis [26], principal component analysis [27] and non-negative component matrix factorization [28], can extract mutually independent subcomponents by mapping high-dimensional images to low-dimensional subspace.

Moreover, optimization-based fusion methods also achieved a certain of development. Further, gradient transfer fusion drives the development of CNN-based fusion methods and GAN-based fusion methods, which defines overall intensity fidelity and texture structure preservation [16] as objective functions of image fusion. Moreover, some researchers improve fusion performance combining the advantages of different frameworks. Further, Liu et al. introduced a novel fusion algorithm and achieved better fusion performance by employing multi-scale transform and sparse representation [29].

2.2. AE-Based Image Fusion Algorithms

Since its powerful nonlinear fitting capability, deep learning was adequately employed in many computer vision tasks. In image fusion, data-driven methods have been explored such as the approach based on an auto-encoder. Similar to traditional fusion methods, AE-based methods also include three categories: feature extraction, feature fusion and feature reconstruction. Prabhakar et al. first finished feature extraction with a small number of convolutional layers in image fusion [30]. The reconstruction was also achieved with a few convolutional layers. In addition, the fused features for different modal images was generated by the element-wise addition.

However, a few convolution layers fail to extract deep features with semantic information. By introducing dense connection and the multi-scale encoder–decoder network, Li et al. had extracted deeper and more comprehensive features and achieved features reuse [31]. However, the fusion performance was influenced to some extent because of the handcrafted fusions rules.

2.3. CNN-Based Image Fusion Algorithms

To achieve better fusion performance, CNN-based image fusion networks have developed by some researchers. Zhang et al. designed a novel CNN-based fusion network that the ratio of gradient and intensity are maintained by intensity and gradient path [32]. The model was trained through a general loss function. In addition, Ma et al. introduced a salient target mask to enable its network extract and fuse more meaningful features [33]. Moreover, Xu et al. proposed a fusion network with elastic weight consolidation since the cross-fertilization of different image fusion tasks [34]. However, the framework fails to fully demonstrate its potential performance because of the missing of ground truth to the fusion results.

2.4. GAN-Based Image Fusion Algorithms

The image fusion is an unsupervised task because a lack of ground truth. Because the adversarial loss restricts the networks at the distribution level, GAN-based methods are relatively sensible for fusion tasks. Ma et al. first introduced the generator and the discriminator in fusion networks. Discriminators facilitate generators capable of generating

images with more meaningful textures [20]. However, the use of a single discriminator is somewhat flawed. It may lead to an imbalance in the data distribution between different images. Further, Ma et al. also presented a new GAN-based image fusion network with a dual-discriminator to achieve the balance [21]. Moreover, Li et al. built the GAN-based fusion network through in on the attention regions with the dual-discriminator and the guidance of attention loss [35]. Subsequently, Ma et al. converted image fusion to multi-distribution simultaneous estimation. Moreover, they introduced a novel image fusion algorithm with multi-classification [36]. The fusion image have more significant contrast and plentiful texture with the multi-classifier.

3. Methodology

In the part, we illustrate implementation details of the overall proposed network. The fusion network is presented in Section 3.1 and the introduction of loss function of training is presented in Section 3.2.

3.1. Network Architecture

We propose an effective DCFusion network. The network includes three parts: encoder module, fusion strategy and decoder module. This overall network is described in Figure 1.

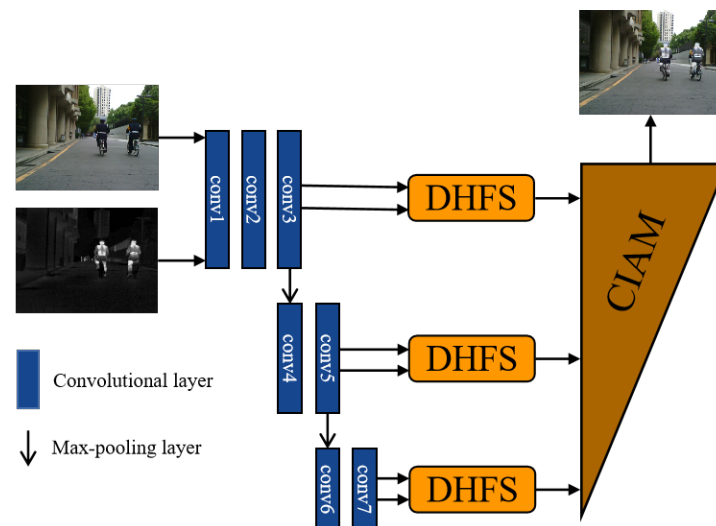


Figure 1. The proposed DCFusion network (S_{vi} , S_{ir} represents source images and O indicate output image).

3.1.1. Encoder Module

The purpose of the encoder module is to obtain multi-scale features of source images. As shown in Figure 1, the encoder includes multiple convolutional layers and max-pooling layers. Combining convolution and pooling operations, we can obtain deep features at different scales. The description of the modules is presented in Table 1.

Table 1. The setting details of the encoder module.

	Size	Stride	Channel (Input)	Channel (Output)	Activation
conv1	1	1	1	16	ReLU
conv2	3	1	16	8	ReLU
conv3	3	1	8	64	ReLU
conv4	3	1	64	32	ReLU
conv5	3	1	32	112	ReLU
conv6	3	1	112	56	ReLU
conv7	3	1	56	160	ReLU

3.1.2. Fusion Strategy

To compensate differential information, we design a dual-headed fusion strategy (DHFS) to merge two modal deep features from the encoder. Subsequently, the output of the module is fed into the reconstruction module to further integrate features. The fusion strategy is presented in Figure 2.

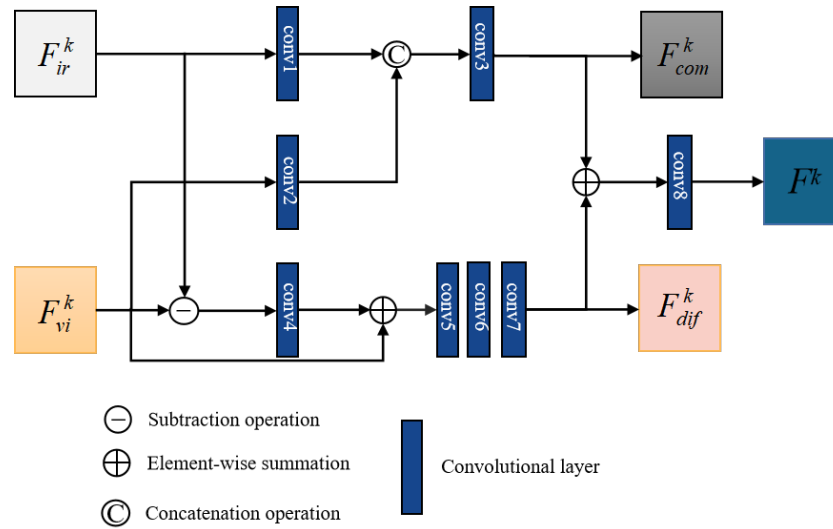


Figure 2. The proposed DHFS module.

In Figure 2, F_{vi}^k and F_{ir}^k , respectively, denote the deep features of infrared and visible image. k indicates the level of different scale deep feature maps. The fusion module is to adequately merge common and differential information and the process can be divided into two heads. In Figure 2, the first head is to fully integrate common information from different modal features. F_{com}^k are the output features of the first branch and contains rich common information. For example, some background information common to infrared images and visible images. Moreover, the F_{com}^k can be defined as:

$$F_{com}^k = \text{conv}(\text{concat}(\text{conv}(F_{vi}^k)), \text{conv}(F_{ir}^k)). \quad (1)$$

where concat indicates concatenation operation in channel dimension for different modal features and every conv indicates different convolutional layer. Furthermore, the second head is to compensate differential information of different modal features. F_{dif}^k are the output features for the second branch and contains abundant differential information. For example, thermal target information for infrared images and texture detail information for visible images. The F_{dif}^k can be defined as:

$$F_{dif}^k = \text{conv}(F_{vi}^k + \text{conv}(F_{ir}^k - F_{vi}^k)). \quad (2)$$

where $+$ indicates element-wise summation operation and $-$ indicates element-wise subtraction operation. conv indicates convolutional layer. Furthermore, the final fused features F^k can be calculated as follows:

$$F^k = \text{conv}(F_{com}^k + F_{dif}^k). \quad (3)$$

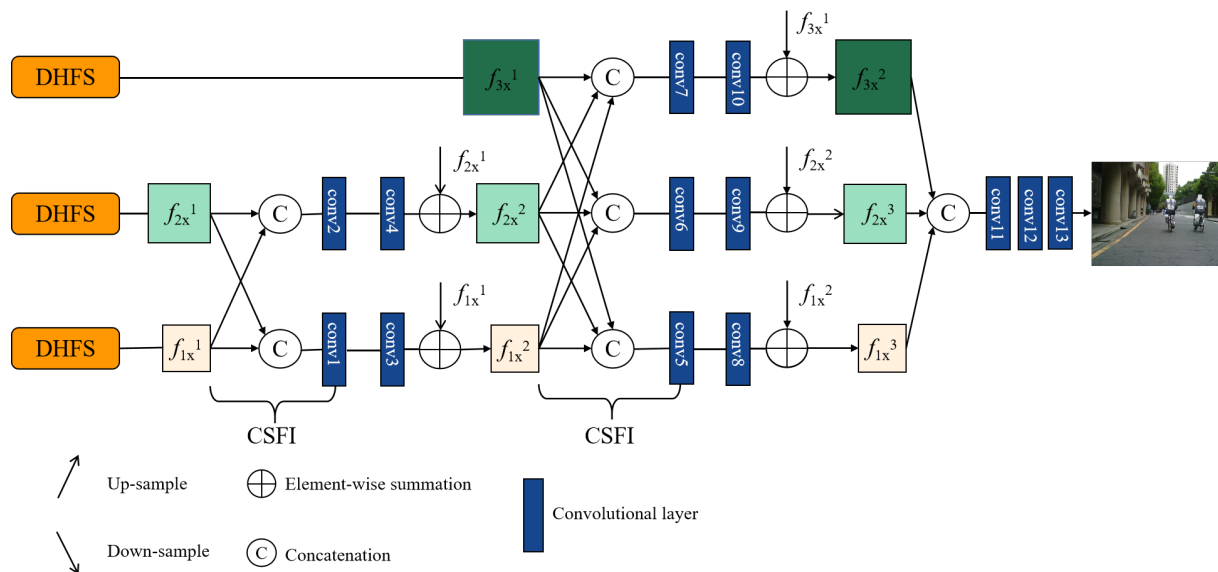
where $+$ indicates element-wise summation operation and conv indicates different convolutional layer. Moreover, the more setting details of the convolutional layers are shown in Table 2.

Table 2. The setting details of the DHFS module.

	Size	Stride	Channel (Input)	Channel (Output)	Activation
conv1-conv2	3	1	K_{input}	K_{input}	ReLU
conv3	3	1	$2 \times K_{input}$	K_{input}	ReLU
conv4-conv7	3	1	K_{input}	K_{input}	ReLU
conv8	1	1	K_{input}	64	ReLU

3.1.3. Decoder Module

The fusion module output the multi-scale feature consisting of three features (F_1 , F_2 , F_3) corresponding to three resolution scale ($\times 4$, $\times 2$, $\times 1$). To further stack the fused multi-scale features, we proposed a contextual information awareness module (CIAM) based on multiple cross-scale feature integration module (CSFI). The decoder is shown in Figure 3. The CIAM can adequately exchange information for different scale feature maps. The low-scale feature is up-sampled to higher-scale feature and the high-scale feature is down-sample to lower feature while the cross-scale feature integration module (CSFI) is employed each time. Inside the CSFI, each scale feature receives exchanged information from other scale features by the up-sample, down-sample and concatenation operations. The channel number of features remains unchanged by convolution operation. Eventually, the channel number of the feature is recovered to one to achieve final fusion performance by three convolutional layers. The CIAM can further improve our fusion performance by combing cross-scale integration module (CSFI) and the setting description of the decoder module is presented in Table 3.

**Figure 3.** The proposed CIAM.**Table 3.** The setting details of the CIAM network.

	Size	Stride	Channel (Input)	Channel (Output)	Activation
conv1-conv2	3	1	128	64	-
conv3-conv4	3	1	64	64	-
conv5-conv7	3	1	192	64	-
conv8-conv10	3	1	64	64	-
conv11	3	1	192	64	-
conv12	3	1	64	32	-
conv13	3	1	32	1	-

3.2. Loss Function

The proposed network architecture aims to obtain satisfactory fusion performance by training abundant visible and infrared images. To achieve the goal, we train the network with MSRS dataset [37] as the input images. In the training processing, a feature-preserving loss function is introduced to ensure that our fused result retains meaningfully common and complementary information. It is worth noting that our loss function mainly refer to proposed loss function of PIAFusion network [37]. The loss function includes three sub-loss function: illumination-aware loss, auxiliary intensity loss and texture loss. The illumination-aware loss is calculated as:

$$L_{illum} = w_{ir}l_{int}^{ir} + w_{vi}l_{int}^{vi}. \quad (4)$$

where l_{int}^{ir} and l_{int}^{vi} , respectively, refer to intensity loss function for infrared and visible image and be calculated as:

$$l_{int}^{ir} = \frac{1}{hw} \|I_o - I_{ir}\|_1. \quad (5)$$

$$l_{int}^{vi} = \frac{1}{hw} \|I_o - I_{vi}\|_1. \quad (6)$$

where h denotes the height and w denotes the width of the inputed images. $\|\cdot\|$ refers to the $-$ -norm. Moreover, I_o , I_{ir} and I_{vi} represent the output, the infrared and the visible images, respectively. w_{ir} and w_{vi} denote illumination-aware weights for infrared and visible images, respectively. The weights are defined as:

$$w_{ir} = \frac{p_n}{p_d + p_n}. \quad (7)$$

$$w_{vi} = \frac{p_d}{p_d + p_n}. \quad (8)$$

where p_d and p_n , respectively, represent the probability that the scene of the images belong to day or night. Moreover, p_d and p_n is calculated by a illumination-aware sub-network [37]. The illumination-aware loss can preserve intensity information but not keep the output in a best intensity distribution. Therefore, we introduce auxiliary intensity loss. The loss is illustrated as follows:

$$L_{aux} = \frac{1}{hw} \|I_o - \max(I_{ir}, I_{vi})\|_1. \quad (9)$$

where $\max(\cdot)$ stands of the element-wise maximum calculation. Furthermore, we introduce texture loss to retain optimal texture of the output and it is defined as follows:

$$L_{tex} = \frac{1}{hw} \|\nabla I_o - \max(|\nabla I_{ir}|, |\nabla I_{vi}|)\|_1 \quad (10)$$

where ∇ stands for the gradient operation and $|\cdot|$ indicates the absolute operation. Therefore, the total loss function can be defined as follows:

$$L_{total} = \gamma_1 L_{illum} + \gamma_2 L_{aux} + \gamma_3 L_{tex}. \quad (11)$$

where $\gamma_i (i = 1, 2, 3)$ represents the weight of different sub-loss function.

4. Experiments

4.1. Experimental Details

The network architecture is trained by the MSRS training dataset [37] and the dataset consists of 26112 pairs images of training set. Before training the encoder-decoder network, we transform visible image into YCbCr color space. At the end of the training, we used the opposite operation again to recover the color image. Moreover, all images of training set are normalized to interval in $[0, 1]$. In the test processing, we test the validity on testing dataset MSRS, which selects 361 pairs of images as the testing images. Moreover, we test

generalization performance for the method with 20 pairs of test images of the TNO [38] dataset.

To comprehensively test performance of the algorithm, we apply the mutual information (MI), the standard deviation (SD), the visual information fidelity (VIF), entropy (EN) and Q_{abf} . The MI can measure the degree of similarity between two images. The larger the MI, the more source image information is retained in the fused image and the quality of the fused image is better. The SD takes stock of the contrast and distribution of the fused images from a statistical point of view. The VIF reflects the fused information fidelity from the human visual perspective. The EN is mainly an objective evaluation metric to measure the amount of information contained in an image. The Q_{abf} evaluates the amount of fused edge information from the source images. Moreover, a fusion model with a better performance usually has more large MI, SD, VIF, EN and Q_{abf} .

To further verify the role of the DCFusion, we compare the method with seven advanced algorithms on MFNet and TNO dataset, namely, MDLatLRR [12], FusionGAN [20], GANMcC [36], IFCNN [39], RFN-Nest [22], SDnet [40] and PIAFusion [37]. The fusion models for all comparisons are evaluated with public code where the relevant settings of the experiment remain unchanged. In the proposed network, the hyper-parameter γ_1 , γ_2 and γ_3 are, respectively, set as 3, 7 and 50. The batch size and epoch are 80 and 30, respectively. The initial learning rate is 0.0001 and then decay after 15 epoch exponentially. In addition to comparative and generalization experiment, we verify the importance of our fusion strategy and decoder module by ablation experiments. Finally, we compared the running efficiency of our method with other methods. Related experiments for our method were conducted on the PyTorch platform. The MDLatLRR was run on a MATLAB R2017b with Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz. Then our DCFusion, RFN-Nest, IFCNN, SDNet and PIAFusion experiments were implemented by PyTorch with one GeForce GTX 1080 Ti GPU. The FusionGAN and GANMcC were implemented by TensorFlow with a GeForce GTX 1080 Ti GPU.

4.2. Comparative Experiment

4.2.1. Qualitative Analysis

Since the MSRS dataset consists of daytime scene and nighttime scene, we evaluate subjective performance with two daytime and two nighttime images. In daytime scenarios, thermal radiation information from infrared images is integrated into visible images as complementary information. A satisfactory fused result should include rich textural details and prominent targets. The daytime fused images are shown in Figures 4 and 5. To show the subjective effect of the fused image more visually, we enlarge the area in the green box to show the texture details better and mark the highlighted targets with red box. It can be seen that the FusionGAN fails to adequately preserve texture details from the visible images, and there is a blurring of the edges of highlighted targets. Although the MDLatLRR can sharpen the edge of the highlighted target, the texture of visible images cannot also be protected fittingly. Moreover, the IFCNN, the GANMcC, the RFN-Nest and the SDnet can integrate meaningful information, but the fusion process is inevitably suffering some degree of interference of useless information. Only our DCFusion and PIAFusion can adequately integrate meaningful information for source images and avoid other distractions at the same time. In the nighttime scenes, the provided scene information for source images is limiting, so it is not easy to obtain a fused result with a wealth of information. The nighttime fused images are shown in Figures 6 and 7. We find that all the algorithms can fuse the common and complementary information from source images to some degree. However, some differences are existing to the fused images of these methods. In addition to the PIAFusion and our DCFusion, other methods introduce contaminated detailed textures and weakened salient targets in the fused images. Moreover, our method achieve a better visual performance to some extent compared with PIAFusion. Therefore, our fusion network can better integrate meaningful information according to the fusion strategy and the proposed decoder.

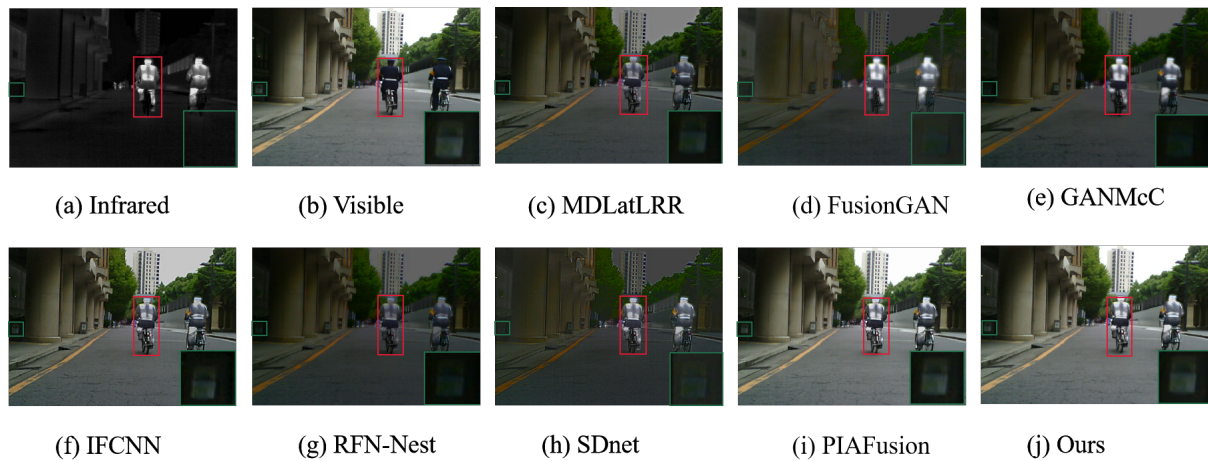


Figure 4. Comparative experiment for different fusion network of a typical daytime sample from the MSRS dataset.

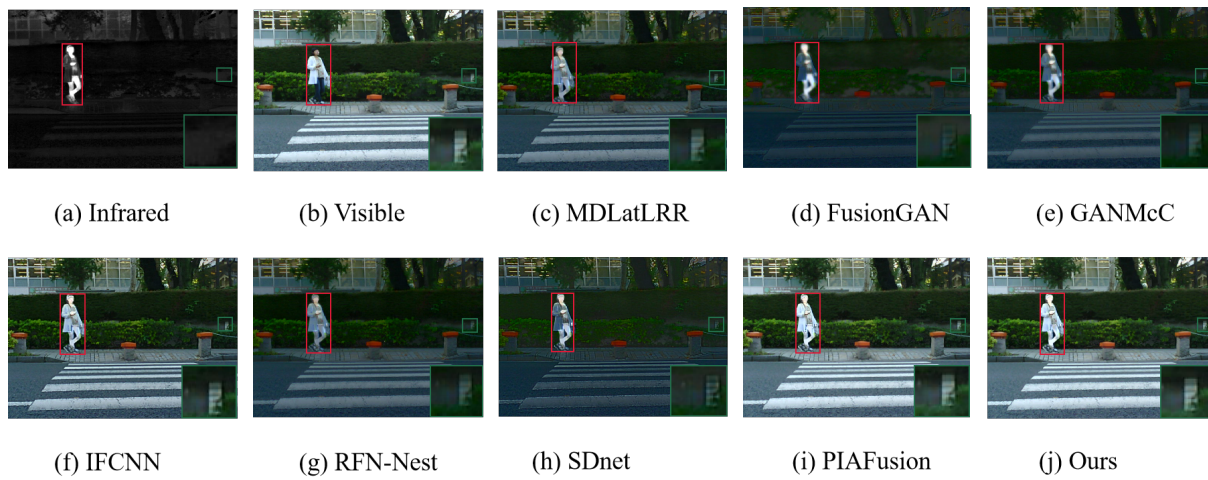


Figure 5. Comparative experiment for different fusion network of a typical daytime sample from the MSRS dataset.

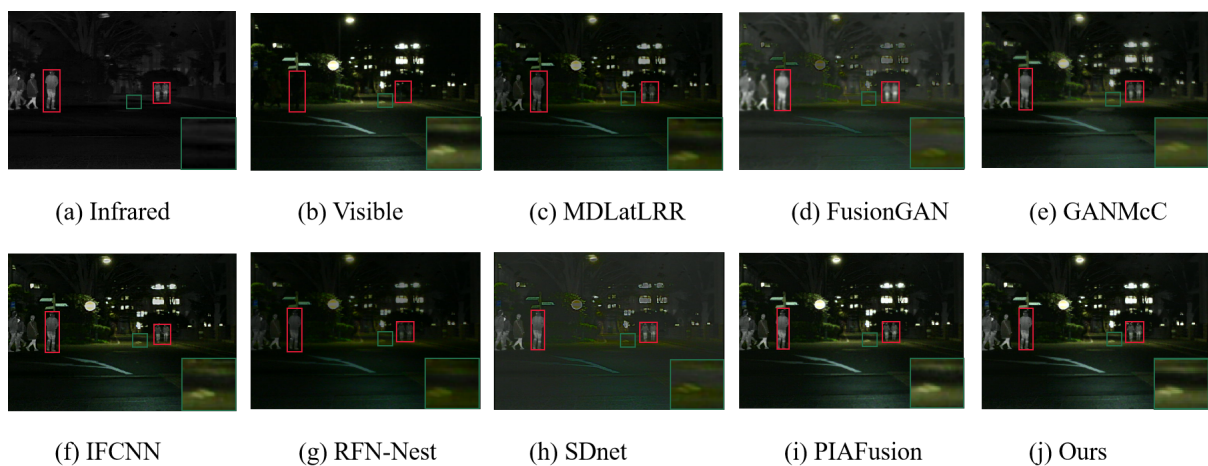


Figure 6. Comparative experiment for different fusion network of a typical nighttime sample from the MSRS dataset.

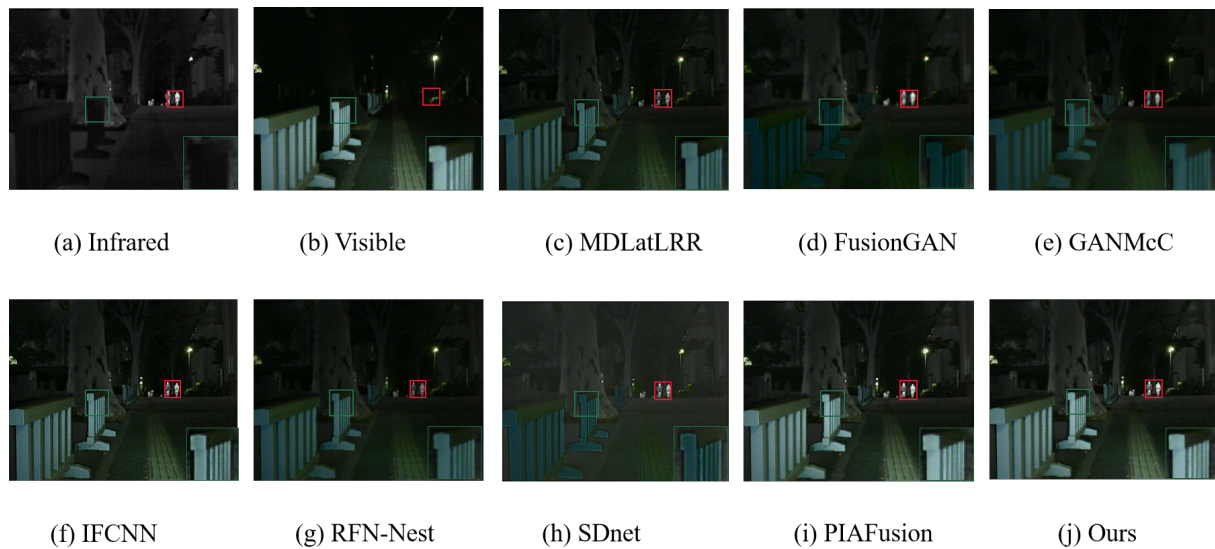


Figure 7. Comparative experiment for different fusion network of a typical nighttime sample from the MSRS dataset.

4.2.2. Quantitative Analysis

To further verify the validity of the method more accurately, we select MI, SD, VIF and Qabf as evaluation metrics. The values of five objective metrics for the eight methods are given in Table 4. The values in the table indicate mean values of 361 pairs of images from MSRS dataset. We observe that our method obtains the best values in 5 metrics. The best MI Indicates that the fused image retains more source image information by the proposed network. The best SD represents the highest contrast for our fused results. The best VIF implies that the fused results own the highest information fidelity. The best EN means that fused images include the most information by our method. The best Q_{abf} indicates that the edge information for fused results is effectively protected. Although our DCFusion is similar to the PIAFusion in subjective performance, our objective metrics are superior to the PIAFusion. Therefore, our fusion network can better integrate meaningful information such as thermal target information for infrared images and texture detail information for visible images.

Table 4. The results of different metrics for the eight algorithms on the 361 MSRS dataset. The red value indicates the best result and the blue value indicates the second best result.

	MI	SD	VIF	Q_{abf}	EN
MDLatLRR	2.5017	7.4546	0.7529	0.5319	6.0099
FusionGAN	1.8569	5.9602	0.4999	0.1396	5.4404
GANMcC	2.5211	8.3479	0.6567	0.2975	6.1232
IFCNN	2.8397	7.9868	0.8530	0.6015	6.4379
RFN-Nest	2.4728	7.0526	0.6514	0.2660	5.6984
SDNet	1.6601	5.7893	0.4338	0.3707	5.2535
PIAFusion	4.5837	8.3451	0.9476	0.6597	6.5710
Ours	4.7319	8.3783	0.9749	0.6624	6.6164

4.3. Generalization Experiment

As we all know, the generalization performance of a data-driven model is an important aspect to evaluate its goodness. Therefore, we select 20 pairs of images in the TNO dataset to evaluate generalizability of our DCFusion. The qualitative results for different models are provided in Figures 8 and 9. To show the subjective effect of the fused image more visually, we enlarge the area in the green box to show the texture details better, and mark the highlighted targets with red box. We observe that the RFN-Nest and the MDLatLRR can

weaken the salient targets. The FusionGAN and the GANMcC can blur the edge of thermal targets. All methods except our DCFusion and the PIAFusion suffer from some degree of spectral contamination. Moreover, we provide quantitative results for different methods in Table 5. We can notice that our DCFusion achieves the best results in SD and EN and obtains the second-best results in the rest of the metrics on the TNO dataset. Moreover, the other metrics are very close to the best values—unsurprising considering that our DCFusion is trained with the MSRS dataset and tested directly on the TNO dataset. Therefore, this proves that our DCFusion can preserve the texture details, sharpen the salient targets and maintain a comparatively good generalization performance effectively.

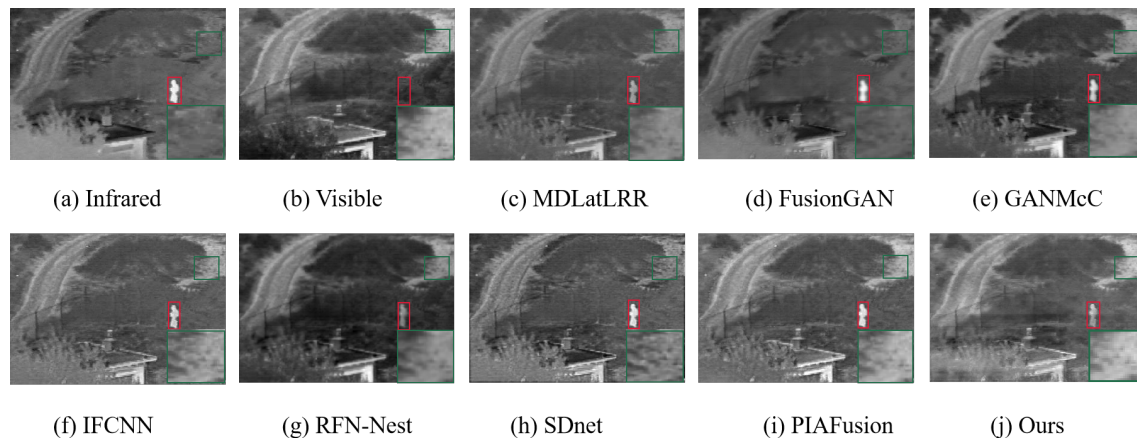


Figure 8. Generalization experiment for different fusion network of a typical sample from the TNO dataset.

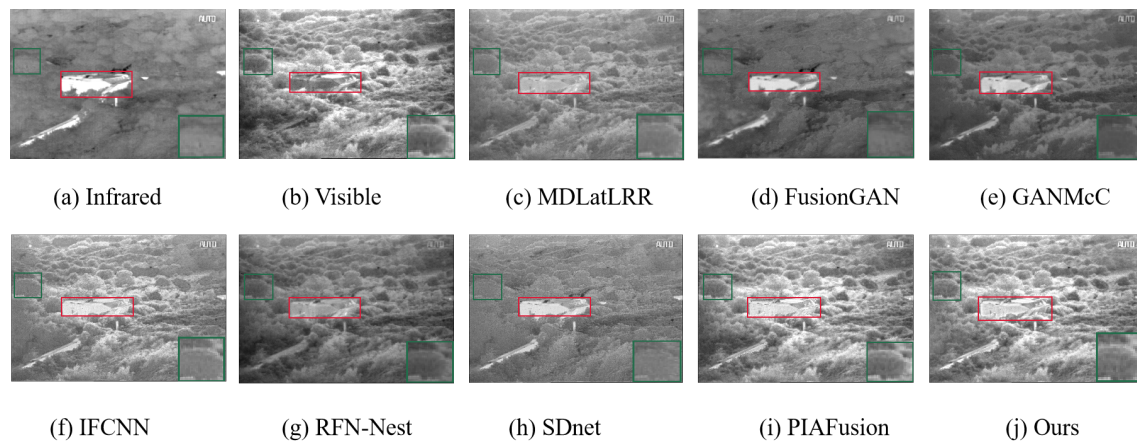


Figure 9. Generalization experiment for different fusion network of a typical sample from the TNO dataset.

Table 5. The results of different metrics for the eight algorithms on the 20 TNO dataset. The red value indicates the best result and the blue value indicates the second best result.

	MI	SD	VIF	Q_{abf}	EN
MDLatLRR	2.2154	8.8255	0.7115	0.4791	6.5917
FusionGAN	2.3543	8.4866	0.6560	0.2202	6.5932
GANMcC	2.5211	8.3479	0.6567	0.2975	6.7322
IFCNN	2.5156	9.1559	0.8026	0.5381	6.9264
RFN-Nest	2.1575	9.1453	0.7661	0.3352	6.9585
SDNet	2.1048	8.7633	0.7027	0.4627	6.6938
PIAFusion	3.5409	9.3599	0.9015	0.5937	7.0528
Ours	2.8913	9.4613	0.8509	0.5794	7.1252

4.4. Ablation Experiment

In our DCFusion network, the DHFS is applied to fully merge the multi-modal deep features, the CSFI is utilized to adequately exchange information for different scale feature maps and the CIAM is employed to reconstruct the fusion result as a decoder of the network. In this section, we further analyze the role of the proposed DHFS, CSFI and CIAM for fusion performance by ablation experiments.

4.4.1. Ablation Study for the DHFS

It is worth noting that we replace the DHFS with three classical fusion strategies including of “add”, “max” and “RFN”. The “add” indicates that the fused features are acquired by adding the multi-modal features and the “max” means an element-wise choose-max method. In addition, the RFN module is utilized in RFN-Nest [22] fusion network as a fusion strategy based on the residual block. Except for the DHFS, the other modules of the proposed network remain unchanged. We select 361 pairs of images from the MSRS dataset as testing images to evaluate fused performance for different fusion strategy. Moreover, the objective metrics (MI, SD, VIF, EN, Q_{abf}) of different fusion strategies are shown in Table 6. We observe that the DHFS achieves the best values in MI, SD, VIF and the second best values in Q_{abf} , EN. Moreover, the Q_{abf} and EN of the DHFS is very close to the best values. Therefore, it powerfully proves that the DHFS can more effectively fuse common and differential information compared with other fusion strategies.

Table 6. The results of ablation study for the DHFS on the 361 MSRS dataset. The red value indicates the best result and the blue value indicates the second-best result.

	MI	SD	VIF	Q_{abf}	EN
add	3.6191	8.3286	0.8640	0.6285	6.4944
max	3.9265	8.3660	0.9561	0.6720	6.6289
RFN	4.5917	8.3671	0.9421	0.6567	6.5379
DHFS	4.7319	8.3783	0.9749	0.6624	6.6164

4.4.2. Ablation Study for the CSFI

In the part, we analysis the role of the CSFI in the decoder module. The decoder module without the CSFI is shown in Figure 10. The modified network is trained by the same training strategy with the DCFusion. The quantitative results of the five metrics are shown in Table 7. The “No-CSFI” means that the decoder does not have the CSFI module. Compared with “No-CSFI”, the DCFusion obtains the best values for all metrics. It indicates that the CSFI models plays an positive role to exchange information of different scale features and boost the reconstruction capacity for the decoder model.

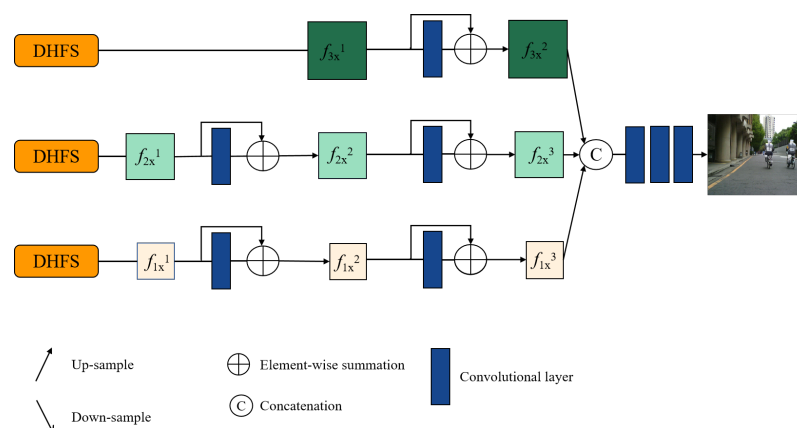


Figure 10. Ablation study for the decoder without CSFI.

Table 7. The results of ablation study for the CSFI on the 361 MSRS dataset. The red value indicates the best result and the blue value indicates the second-best result.

	MI	SD	VIF	Q_{abf}	EN
No-CSFI	4.6376	8.3574	0.9547	0.6444	6.6065
DCFusion	4.7319	8.3783	0.9749	0.6624	6.6164

4.4.3. Ablation Study for the CIAM

In the paper, we introduce a completely new decoder module to integrate different scale features and obtain the final images. To verify the validity of the CIAM, we replace the CIAM with the decoder network of RFN-Nest [22]. The difference is that the decoder contains only three scales of feature fusion in our experiment. Apart from the decoder, all other settings remain unchanged compared with original experiment. We tested the performance of 361 pair images from the MSRS dataset to validate the role of the CIAM. Further, the object metrics of the different decoder are shown in Table 8. Compared with the decoder of RFN-Nest, the proposed CIAM achieves best values in MI, SD, VIF and EN and the second-best Q_{abf} . Moreover, the Q_{abf} of the CIAM is very close to the best value. Therefore, it validates that the CIAM can more effectively reconstruct the fused images.

Table 8. The results of ablation study for the CSFI on the 361 MSRS dataset. The red value indicates the best result and the blue value indicates the second best result. The RFN-decoder indicates the decoder of the RFN-Nest.

	MI	SD	VIF	Q_{abf}	EN
RFN-Decoder	4.6560	8.3488	0.9663	0.6851	6.5884
CIAM	4.7319	8.3783	0.9749	0.6624	6.6164

4.5. Efficiency Comparison Experiment

To verify the execution efficiency of our method, we tested the processing time for each image pair on the 361 MSRS dataset and compared the time with other fusion methods. The final result is shown in Table 9. It can be seen that the efficiency of our method is higher compared to traditional methods. Because our method requires feature extraction at multiple scales, it is relatively time-consuming. Fortunately, the running efficiency of our method stays within an acceptable range.

Table 9. The average processing time for each image pair on the 361 MSRS dataset. The red value indicates the best result and the blue value indicates the second best result (unit: s).

Method	Time
MDLatLRR	123.4549
FusionGAN	0.0681
GANMcC	0.1333
IFCNN	0.0160
RFN-Nest	0.1924
SDNet	0.0154
PIAFusion	0.0895
Ours	0.1546

5. Conclusions

In the paper, we design a new image fusion network for infrared and visible images with a dual-headed fusion strategy and a contextual information awareness. First of all, the encoder is employed to extract different scale features of infrared and visible images. Secondly, the dual-headed fusion strategy (DHFS) is devised to sufficiently integrate common and differential information from different modal features. Thirdly, we introduce

a contextual information awareness module (CIAM) to reconstruct fused images based on the CSFI model. Combined with three loss function of L_{illum} , L_{aux} and L_{tex} , the proposed DCFusion obtains the best performance in qualitative and quantitative evaluation compared with seven progressive fusion methods. Moreover, we prove a good generalization ability of the DCFusion in TNO dataset with a generalization experiment. Finally, the expanded ablation experiments validate that the proposed DHFS, CSFI and CIAM are beneficial to boost the fused performance.

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

Funding: The research in our paper is sponsored by the funding from Sichuan University and Yibin Municipal People's Government University and City strategic cooperation special fund project (Grant No. 2020CDYB-29); Science and Technology plan transfer payment project of Sichuan province (2021ZYSF007); The Key Research and Development Program of Science and Technology Department of Sichuan Province (No. 2020YFS0575, No.2021KJT0012-2021YFS0067).

Data Availability Statement: Not applicable.

Acknowledgments: We thank all the editors and reviewers in advance for their valuable comments that will improve the presentation of this paper.

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

References

1. Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; Ma, J. Image fusion meets deep learning: A survey and perspective. *Inf. Fusion* **2021**, *76*, 323–336. [\[CrossRef\]](#)
2. Cao, Y.; Guan, D.; Huang, W.; Yang, J.; Cao, Y.; Qiao, Y. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *Inf. Fusion* **2019**, *46*, 206–217. [\[CrossRef\]](#)
3. Li, C.; Zhu, C.; Huang, Y.; Tang, J.; Wang, L. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 808–823.
4. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-modality person re-identification with shared-specific feature transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13379–13389.
5. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5108–5115.
6. 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. [\[CrossRef\]](#)
7. Li, H.; Qi, X.; Xie, W. Fast infrared and visible image fusion with structural decomposition. *Knowl.-Based Syst.* **2020**, *204*, 106182. [\[CrossRef\]](#)
8. Ma, J.; Zhou, Y. Infrared and visible image fusion via gradientlet filter. *Comput. Vis. Image Underst.* **2020**, *197*, 103016. [\[CrossRef\]](#)
9. Liu, F.; Chen, L.; Lu, L.; Ahmad, A.; Jeon, G.; Yang, X. Medical image fusion method by using Laplacian pyramid and convolutional sparse representation. *Concurr. Comput. Pract. Exp.* **2020**, *32*, e5632. [\[CrossRef\]](#)
10. Liu, F.; Chen, L.; Lu, L.; Jeon, G.; Yang, X. Infrared and visible image fusion via rolling guidance filter and convolutional sparse representation. *J. Intell. Fuzzy Syst.* **2021**, *40*, 10603–10616. [\[CrossRef\]](#)
11. Gao, C.; Liu, F.; Yan, H. Infrared and visible image fusion using dual-tree complex wavelet transform and convolutional sparse representation. *J. Intell. Fuzzy Syst.* **2020**, *39*, 4617–4629. [\[CrossRef\]](#)
12. 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. [\[CrossRef\]](#)
13. Chen, J.; Li, X.; Luo, L.; Mei, X.; Ma, J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf. Sci.* **2020**, *508*, 64–78. [\[CrossRef\]](#)
14. Bai, X.; Zhou, F.; Xue, B. Fusion of infrared and visual images through region extraction by using multi scale center-surround top-hat transform. *Opt. Express* **2011**, *19*, 8444–8457. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Liu, Y.; Chen, X.; Ward, R.K.; Wang, Z.J. Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* **2016**, *23*, 1882–1886. [\[CrossRef\]](#)

16. 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. [\[CrossRef\]](#)
17. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17. [\[CrossRef\]](#)
18. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [\[CrossRef\]](#)
19. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [\[CrossRef\]](#)
20. 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. [\[CrossRef\]](#)
21. 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. [\[CrossRef\]](#)
22. 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. [\[CrossRef\]](#)
23. Liu, Y.; Jin, J.; Wang, Q.; Shen, Y.; Dong, X. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Process.* **2014**, *97*, 9–30. [\[CrossRef\]](#)
24. Liu, X.; Mei, W.; Du, H. Structure tensor and nonsubsampling shearlet transform based algorithm for CT and MRI image fusion. *Neurocomputing* **2017**, *235*, 131–139. [\[CrossRef\]](#)
25. Zhang, Q.; Maldague, X. An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing. *Infrared Phys. Technol.* **2016**, *74*, 11–20. [\[CrossRef\]](#)
26. Cvejic, N.; Bull, D.; Canagarajah, N. Region-based multimodal image fusion using ICA bases. *IEEE Sens. J.* **2007**, *7*, 743–751. [\[CrossRef\]](#)
27. Fu, Z.; Wang, X.; Xu, J.; Zhou, N.; Zhao, Y. Infrared and visible images fusion based on RPCA and NSCT. *Infrared Phys. Technol.* **2016**, *77*, 114–123. [\[CrossRef\]](#)
28. Mou, J.; Gao, W.; Song, Z. Image fusion based on non-negative matrix factorization and infrared feature extraction. In Proceedings of the 2013 6th International Congress on Image and Signal Processing (CISP), Hangzhou, China, 16–18 December 2013; Volume 2, pp. 1046–1050.
29. Liu, Y.; Liu, S.; Wang, Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* **2015**, *24*, 147–164. [\[CrossRef\]](#)
30. Ram Prabhakar, K.; Sai Srikar, V.; Venkatesh Babu, R. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4714–4722.
31. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [\[CrossRef\]](#)
32. 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.
33. 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*, 1–13. [\[CrossRef\]](#)
34. 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. [\[CrossRef\]](#)
35. 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. [\[CrossRef\]](#)
36. Ma, J.; Zhang, H.; Shao, Z.; Liang, P.; Xu, H. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–14. [\[CrossRef\]](#)
37. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83*, 79–92. [\[CrossRef\]](#)
38. Toet, A. TNO Image fusion dataset. *Figshare Dataset* **2014**. [\[CrossRef\]](#)
39. 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. [\[CrossRef\]](#)
40. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [\[CrossRef\]](#)

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.