Journal of
Marine Science and Engineering

Article

Fusion2Fusion: An Infrared–Visible Image Fusion Algorithm for Surface Water Environments

Cheng Lu [1], Hongde Qin [2,3], Zhongchao Deng [2,3] and Zhongben Zhu [2,3,*]

[1] Yantai Research Institute, Harbin Engineering University, Yantai 264000, China
[2] Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao 266000, China
[3] Science and Technology on Underwater Vehicle Technology Laboratory, Harbin Engineering University, Harbin 150001, China
[*] Correspondence: zhuzhongben@hrbeu.edu.cn

Abstract: Infrared images can rely on the thermal radiation of objects for imaging, independent of lighting conditions. Furthermore, because the thermal radiation produced by targets such as people, vehicles, and boats differs greatly from the background, it is able to distinguish objects from their environment as well. These characteristics of infrared can be complemented with visible images, which are rich in color information but vulnerable to lighting conditions. Therefore, the fusion of IR and visible images can provide a better perception of the environment. In this paper, we propose a new infrared–visible fusion algorithm. It consists of three parts: feature extraction, fusion, and reconstruction. The attention mechanism is introduced into the feature extraction to better extract features and we propose a new way of describing the fusion task. The relationship between the two inputs is balanced by introducing a fused image obtained by summing the infrared and visible images. It is also optimized for sky layering and water surface ripples, which are common in water environments. The edge information is enhanced in the loss function and noise reduction is performed. Through comparison experiments, our algorithm achieves better results.

Keywords: image fusion; unsupervised learning; infrared; visible; encoder–decoder

check for updates

Citation: Lu, C.; Qin, H.; Deng, Z.; Zhu, Z. Fusion2Fusion: An Infrared–Visible Image Fusion Algorithm for Surface Water Environments. *J. Mar. Sci. Eng.* **2023**, *11*, 902. https://doi.org/10.3390/jmse11050902

Academic Editors: Mai The Vu and Hyeung-Sik Choi

Received: 14 March 2023
Revised: 13 April 2023
Accepted: 15 April 2023
Published: 23 April 2023

1. Introduction

Unmanned platforms, such as unmanned surface vehicles, have been widely applied due to the development of unmanned technology. The perception capability of unmanned systems is necessary to complete predefined tasks. As a basic sensor, visible light cameras are widely used because the information they collect can reflect object characteristics such as size, shape, and color. However, its disadvantages are also obvious, such as the imaging quality is vulnerable to lighting conditions, and it is difficult to collect effective data under poor lighting conditions. The thermal radiation imaging used by the infrared cameras is not affected by the lighting conditions. They can capture black and white images with features of object size and shape, but the color information is not as rich as that of visible light cameras. As a result, information fusion of infrared-visible images can combine the benefits of both, and provide better image information for subsequent sensing tasks.

It is well known that many signal processing methods have been applied to image fusion tasks to extract salient features of images, such as methods based on multi-scale decomposition [1]. First, salient features are extracted by image decomposition methods. Then, an appropriate fusion strategy is used to obtain the final fused image. The representation learning-based approaches have also received a lot of attention recently. Numerous fusion techniques have been proposed for the sparse domain, including joint sparse representation (JSR) [2], co-sparse representation [3], and sparse representation (SR) and histogram of oriented gradients (HOG)-based fusion methods [4]. Li and Wu [5] proposed a low-rank representation (LRR)-based fusion method in the low-rank domain.

Instead of using SR, they extract the features using LRR, and then they reconstruct the fused image using the l1-norm and the max selection strategy.

Convolutional neural networks (CNNs) have made a number of important strides in recent years in a variety of computer vision and image processing tasks, including image segmentation, super-resolution restoration, classification, and saliency detection, among others. Liu et al. [6] suggested a multifocus image fusion method using CNNs, and they trained a Siamese network [7] to categorize the focus and defocus patches and produce a precise binary decision map. The decision map and the corresponding source images were combined to create the fused image. On the basis of this, Tang et al.'s algorithm was improved by suggesting pixel-CNNs for the classification of focused and defocused pixels. By fusing a deep Boltzmann machine with a multi-scale transform, Wu et al. [8] presented a framework for fusing infrared and visible images. Li et al. [9] built a solid weight map for fusion by extracting multilayer deep features from the VGG network. They then used ImageNet [10] to train a dense network [11], a deep learning framework that was intended to carry out joint activity-level measurement and weight assignment. As a result, CNNs have not acquired the adaptive ability to combine or choose deep features. These techniques also produce excellent results, but they still rely on manual settings and do not offer complete solutions.

These methods have achieved state-of-the-art performance. There is no doubt that visible and infrared image fusion based on CNNs is worth researching. With the rapid advancement of deep learning, it is widely applied in the field of image fusion. Unlike tasks that require labeling (such as target detection and segmentation), image fusion is difficult to be labeled manually, so deep-learning-based unsupervised algorithms are well suited to image fusion. The good real-time performance and stable results determine that unsupervised deep learning algorithms are also better suited for deployment and application in unmanned surface systems with valuable computational resources.

One of the challenges of unsupervised deep-learning-based image fusion algorithms is how to define the task, particularly for infrared visible light fusion. As a multiple-input single-output activity, it is challenging to balance the relationship between each input and output. As a result, we propose a new approach for task description. Image I_C is created by combining two input images. The task is described by image I_C and the output image. The final result can also be adjusted to suit different environments by adjusting the proportion of the two input images during the summation of image I_C.

The surface water environment dominated by the sky and water has the most noticeable differences compared with the terrestrial environment. Different imaging methods of infrared cameras lead to color breaks in the sky and excessive ripples in the water, which will eventually cause interference to the fused image I_F. To address this feature, images are denoised and smoothed during the fusion process to reduce information interference.

The remainder of this paper is organized as follows. Section 2 briefly reviews the theory of related works. In Section 3, we present the proposed CNN-based image fusion. The experimental details are introduced and discussed in Section 4. Section 5 draws the conclusions.

2. Related Work

Many fusion algorithms have been proposed in recent years and these are mainly divided into traditional methods based on multi-scale decomposition- and representation-based learning, and deep-learning-based fusion algorithms.

Multi-scale transform (MST) methods have been the subject of in-depth study over the last few decades. Discrete wavelet transform (DWT) [12], Laplacian pyramid (LAP) [13], contourlet transform (CT) [14], nonsubsampled contourlet transform (NSCT) [15], nonsubsampled shearlet transform (NSST) [16], framelet transform (FT) [17], curvelet transform (CVT) [18], and discrete cosine transform (DCT) [19] are some of the traditional tools used in MST. Typically, MST-based infrared and visible image fusion schemes have three steps. The source images are divided into a number of multi-scale coefficients in the first step. The decomposed coefficients are then fused in accordance with predetermined guidelines. In the end, the corresponding inverse multi-scale transform is used to create the fused image. The key to these strategies is to pick a superior decomposition method and an advanced fusion rule, which frequently results in increased complexity.

Unlike representation learning- and multi-scale decomposition-based techniques, deep-learning-based algorithms employ a large number of pictures to train their neural networks, which are then used to extract significant features. Liu et al. [20] proposed a fusion method based on convolutional sparse representation (CSR). Although the CSR differs from CNN-based approaches, this algorithm is still based on deep learning because it also extracts deep features. In this method, authors use source images to learn several dictionaries which have different scales and employ CSR to extract multi-layer features, and then a fused image is generated by these features. For the task of multi-focus image fusion, Liu et al. [6] also presented a CNN-based fusion technique. To train the network and obtain a decision map, different blur versions of the input image are used in image patches. The decision map and the source images are then used to create the fused image. This approach, however, is only appropriate for multi-focus image fusion.

With the development and application of deep learning in the field of image fusion, many deep-learning-based image fusion algorithms have been proposed. Unlike learning-based tasks such as target detection and tracking, image fusion does not have explicit labels or ground truth. Therefore, unsupervised learning is widely used in the field of image fusion. The CNN-based network is employed to extract features and loss functions are constructed by comparing input and output images.

In ICCV 2017, Prabhakar et al. [21] proposed an unsupervised deep-learning-based algorithm for static multi-exposure image fusion. The proposed architecture has three components: feature extraction layer, fusion layer, and reconstruction layer. The two input images are encoded by a coding network consisting of two CNN layers, and the two feature map sequences obtained are fused by an addition strategy. The final fused image is reconstructed by a decoding layer consisting of three CNN layers. This algorithm constructs an unsupervised image fusion network and achieves good performance, but the network structure is simple, and the feature extraction ability is limited.

Li et al. [11] applied an unsupervised algorithm to the field of visible and infrared image fusion and also used the structure of encoder, fusion layer, and decoder. To extract multi-level features, Denseblock is selected as the encoder. In the fusion layer, the l1-Norm Strategy is introduced and compared with the original Addition Strategy.

Hou et al. [22] carried out further optimization for infrared characteristics. Firstly, *SSIM* was improved by deleting the factor related to global brightness due to the large difference between the global brightness of infrared and visible images. Secondly, the description in the task was improved to compare the brightness of local IR and visible images, and the part of both with higher brightness was selected to form the fused image.

3. Proposed Fusion Method

The proposed deep-learning-based algorithm Fusion2Fusion (F2F) is thoroughly introduced in this section. The advancement of unsupervised algorithms for describing infrared visible fusion task approaches will be paid special attention.

3.1. Network Architecture

The general framework of the F2F algorithm is shown in Figure 1, which contains three parts: feature extraction, fusion, and reconstruction. The input visible and infrared images are denoted as I_A and I_B, respectively. I_A and I_B use the same feature extraction network with shared weights. As shown in Figure 1, the feature extraction network includes D11, Denseblock [23], containing D12, D13, D14, and CBAM [24]. In feature extraction, our basic unit consists of three consecutive operations: batch normalization (BN) [25], rectified linear unit (ReLU) [26], and 3 × 3 convolution (Conv). Dense connectivity will improve information flow between layers.

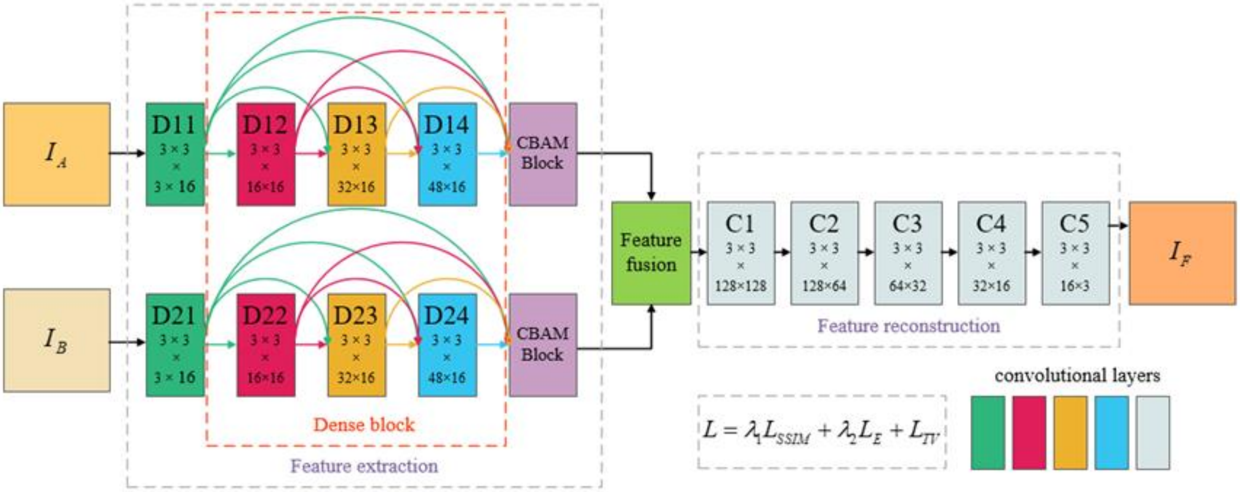

Figure 1. The architecture of the proposed F2F.

The convolutional block attention module (CBAM) represents the attention mechanism module of a convolutional module, which is an attention mechanism module that combines spatial and channel data. It is a simple and effective attention module for feedforward convolutional neural networks. Given an intermediate feature map, the CBAM module sequentially infers the attention map along two independent dimensions (channel and space) and then multiplies the attention map with the input feature map to perform adaptive feature optimization. CBAM is utilized in this paper to better extract features.

In the feature fusion section, the feature maps are connected rather than using the fusion strategy of fusing two feature maps into one to retain more information. The feature fusion stage directly superimposes the two sets of features in the channel direction in order to retain more information. Finally, the result of the fusion layer is passed through another five convolutional layers (C1, C2, C3, C4, and C5) to reconstruct the fused image denoted as I_F.

3.2. Loss Function

In this section, the construction of loss functions from *SSIM*, *TV*, and *MSE* will be discussed, and the fusion images I_C obtained by adding visible and infrared images will be introduced to better describe the unsupervised fusion task. The overall structure of the loss function is shown in Figure 2.

The loss function consists of three components, L_{SSIM}, L_E, and L_{TV}, where L_{SSIM} is the main component and is used to compare the similarity of input and output, L_E is the edge loss and is used to enhance the infrared edge information in the fused image, and L_{TV} is the noise reduction loss and is used for image noise reduction. λ_1 and λ_2 are two parameters referred to in Reference [22]. L_E and L_{TV} are 10^3 times different from L_{SSIM} in value. Therefore, the two parameters, λ_1 and λ_2, are introduced. The total loss function L is shown in Equation (1):

$$L = \lambda_1 L_{SSIM} + \lambda_2 L_E + L_{TV} \tag{1}$$

The *SSIM* loss L_{SSIM} is obtained by Equation (2)

$$L_{SSIM} = 1 - \frac{1}{N} \sum_{W=1}^{N} F_{SSIM}(W) \tag{2}$$

where the final fused image related to the infrared image and the visible image is represented by the F_{SSIM} composed of *SSIM*. *SSIM* [27] denotes a method for measuring the structural similarity of two images, which compares the structural similarity of each component of the two images by a sliding window. The sliding window W moves from the top left to the bottom right in the image to determine the overall similarity between the two. The parameter N represents the total number of sliding windows in a single image. As suggested in [27], the window size is set to 11 × 11 in our study.

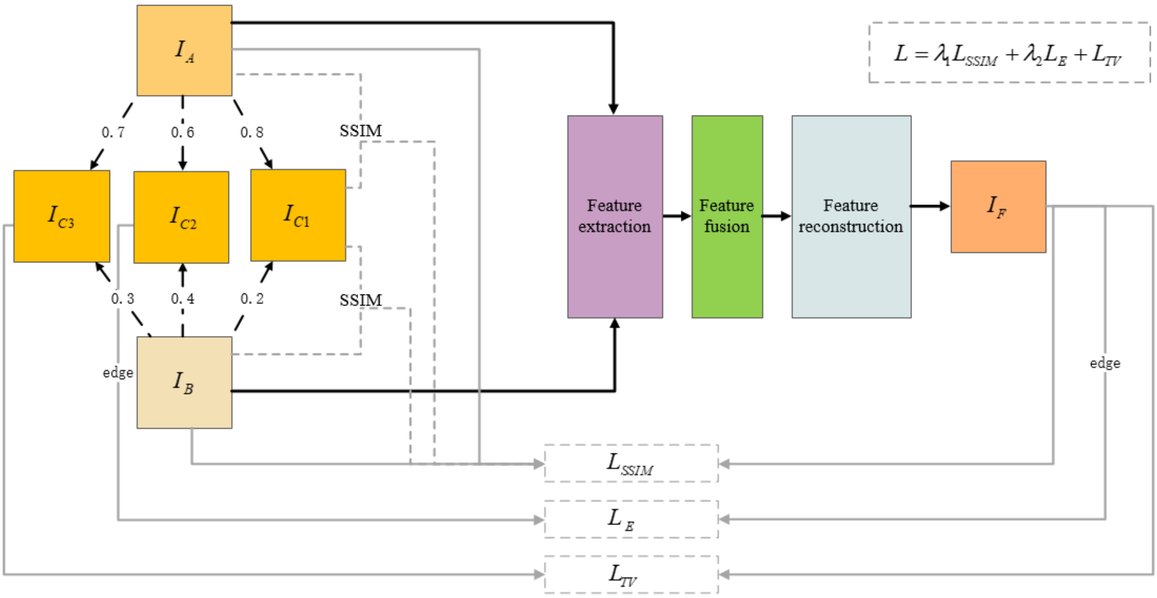

Figure 2. The loss function of the proposed F2F.

In F_{SSIM}, the image I_{C1} is obtained by adding image I_A and I_B, which can be changed by adjusting the contribution of I_A and I_B. Image I_{C1} is treated as a kind of scale, and its similarity with images I_A and I_B is compared in the sliding window. The parts of images I_A and I_B with high similarity to I_{C1} are selected to form a loss with the final fused image I_F. The F_{SSIM} is shown in Equation (3).

$$F_{SSIM}(W) = \begin{cases} SSIM(A, F|W) & , SSIM(A, C1|W) \geq SSIM(B, C1|W) \\ SSIM(B, F|W) & , SSIM(A, C1|W) < SSIM(B, C1|W) \end{cases} \tag{3}$$

When $SSIM(A, C1|W)$ exceeds or is equal to $SSIM(B, C1|W)$, more information from image I_A is presented in the sliding window of I_{C1}, and $SSIM$ directs the network to retain more information from image I_A in the region of the fused image I_F. By changing the proportion of images I_A and I_B in I_{C1}, the retention degree of infrared and visible information in the fused image can be altered. I_{C1} is used as a scale in the loss to determine whether this region of the fused image for comparison contains more visible or infrared information. For example, when I_{C1} has a higher similarity to the visible image, the fused image is made to contain more visible information in this region. The overall tendency of the final fused image is then determined by adjusting I_{C1}. According to the characteristics of the water surface image, in our study, image I_{C1} is obtained by adding 80% of visible image I_A and 20% of infrared image I_B.

L_E is introduced to represent more texture details in the fused image, as shown in Equation (4), and denote the edge texture of images I_F and I_{C2} extracted by the canny operator, where I_{C2} is obtained by summing 60% of image I_A and 40% of image I_B. Then, their mean square error (MSE) is calculated for the edge texture information of the two images.

$$L_E = MSE(edge_F, edge_{C2}) \tag{4}$$

To achieve gradient conversion and noise reduction, the total variation function is introduced as follows:

$$R(i, j) = I_{C3}(i, j) - I_F(i, j) \tag{5}$$

$$L_{TV} = \sum_{i,j} (||R(i, j+1) - R(i, j)||_2 + ||R(i+1, j) - R(i, j)||_2) \quad (6)$$

where $R(i, j)$ denotes the difference between the image I_{C3} and the fused image I_F, and $||\cdot||_2$ denotes the l2 distance. Since there are differences in the order of magnitude among the three loss functions, two hyperparameters λ_1 and λ_2 are set. Based on the outcomes of various experiments, λ_1 and λ_2 are set to 2000 and 10, respectively.

3.3. Training

A total of 2000 pairs of infrared and visible images were selected from the MIT Sea Grant Marine Autonomy Dataset Project. The dataset contains the viewpoint data of the ship in different locations and light conditions. The images were adjusted to 512 × 512 and used for training without any manual annotation. In addition, the network was trained for 200 epochs, and the loss was minimized using the Adam optimizer with a learning rate of 10^{-2}. Our network was implemented on PyTorch [28] and trained on a PC with an Intel(R) Xeon(R) Gold 5320 2.20 GHz CPU, 32 GB RAM, and an NVIDIA RTX A4000 GPU.

4. Experimental Results and Analysis

In this section, some experiments and comparisons are presented to verify the effectiveness of F2F. Three pairs of infrared visible source photos from various situations are selected, as shown in Figure 3, and named as Example A, Example B, and Example C, respectively. These three pairs of photos serve as examples for later comparison and investigation.

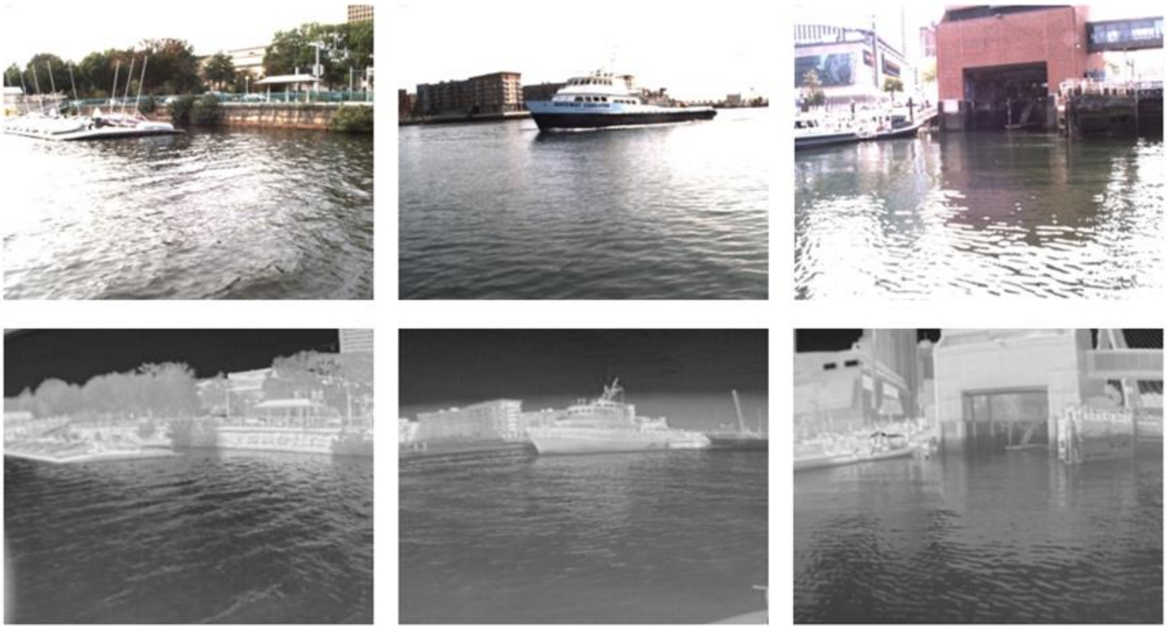

Figure 3. Three pairs of source images.

4.1. Experimental Setting

Several algorithms are selected for comparison in the experiments, including CBF [29], VSMWLS [30], Densefuse-Add [11], and Densefuse-L1 [11]. All four comparative methods were implemented based on publicly available codes, where the parameters were set according to the original papers. Several different I_{C1} images are compared to comprehend the impact of various I_{C1} images on the final fusion effect. For instance, Proposed-73 denotes that 70% of image I_A and 30% of image I_B are added to create I_{C1}. Proposed-82 and Proposed-91 are defined in the same way as Proposed-73. The comparison results are shown in Figures 4–6. Areas that need to be highlighted are marked by red dashed boxes.

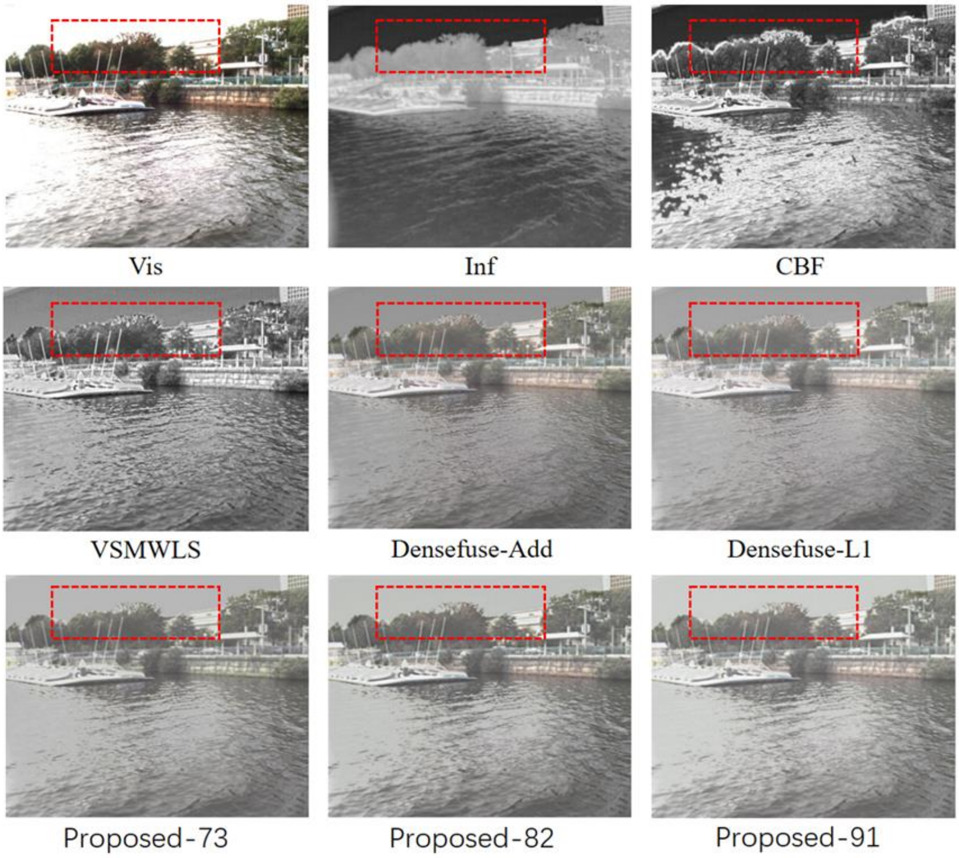

Figure 4. The fusion results of different algorithms in Example A.

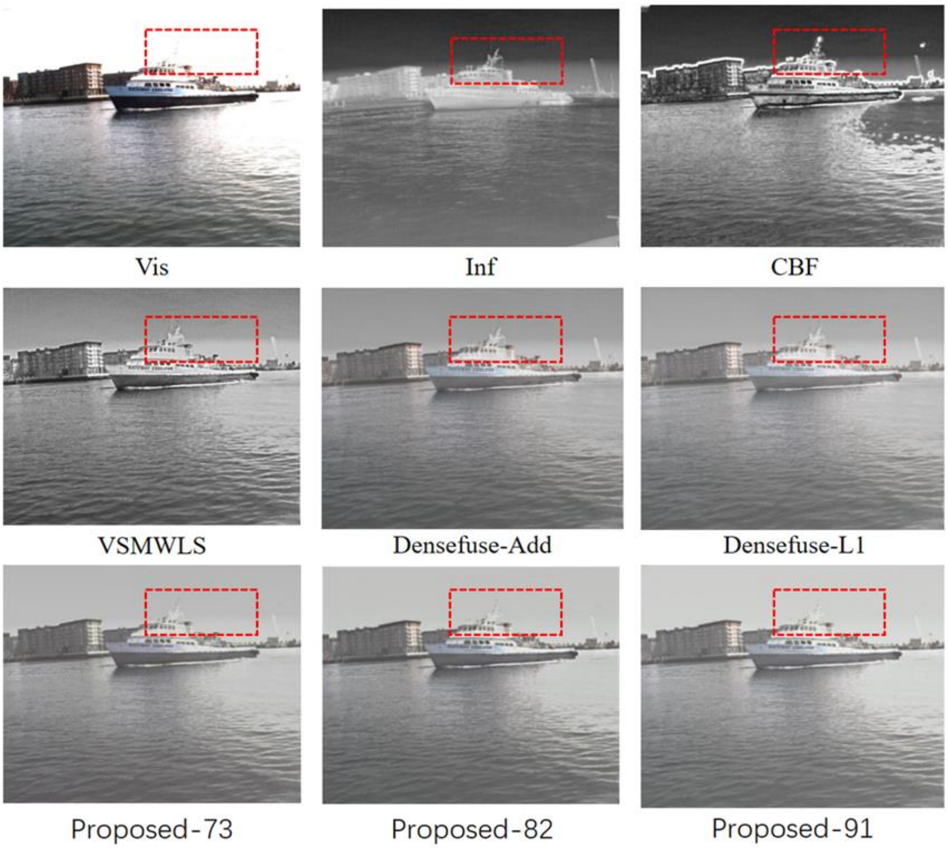

Figure 5. The fusion results of different algorithms in Example B.

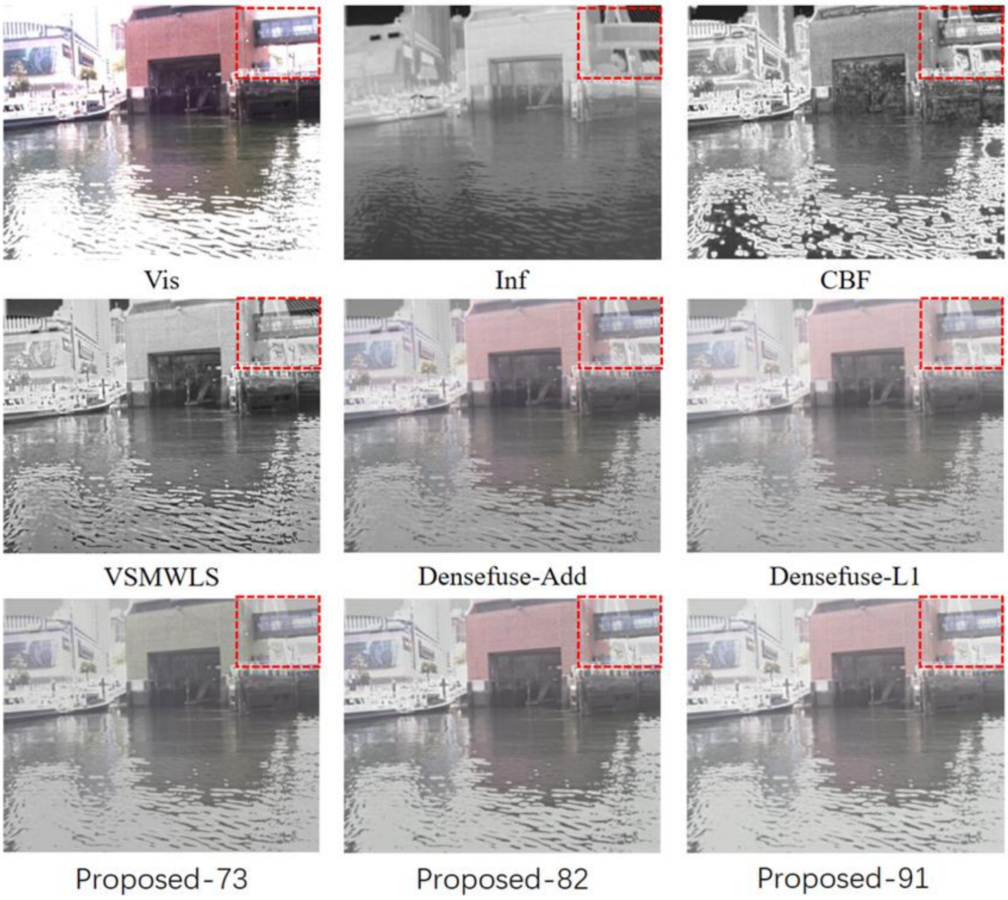

Figure 6. The fusion results of different algorithms in Example C.

The subjective visual assessment system is susceptible to human factors, such as eyesight, subjective preference, and individual emotion. Moreover, there is an insignificant difference among the fused results in most cases based on subjective evaluation. Thus, analyzing the fusion performance based on quantitative evaluation is indispensable. Seven quality metrics are utilized to quantitatively compare our fusion method with other existing algorithms: $Q^{AB/F}$ [31] the sum of the correlations of differences (SCD) [32]; modified structural similarity for no-reference image ($SSIM_a$) [11]; a new no-reference image fusion performance measure (*MS-SSIM*) [33]; and entropy (*En*) and nonlinear correlation information entropy (Q^{NCIE}) [34].

$Q^{AB/F}$ measures the amount of edges transferred from the source images to the fused image. It is defined as follows:

$$Q^{AB/F} = \frac{\sum\limits_{N}^{i=1}\sum\limits_{M}^{j=1}\left(Q^{AF}(i,j)w^{A}(i,j) + Q^{BF}(i,j)w^{B}(i,j)\right)}{\sum\limits_{N}^{i}\sum\limits_{M}^{j}(w^{A}(i,j) + w^{B}(i,j))} \tag{7}$$

where $Q^{AF}(i,j) = Q_g^{AF}(i,j)Q_o^{AF}(i,j)$, $Q_g^{AF}(i,j)$ and $Q_o^{AF}(i,j)$ are the edge strength and orientation preservation value at the location (i, j), respectively. N and M are the size of the image, and $Q^{BF}(i, j)$ is similar to $Q^{AF}(i, j)$, $w^A(i, j)$, and $w^B(i, j)$ denote the weights of $Q^{AF}(i, j)$ and $Q^{BF}(i, j)$, respectively.

The sum of the correlations of differences (SCD) is used to measure the information correlation between the fused image and the source image. The aim of the image fusion is to merge the images that depict the same scene with different sensing technology. The most meaningful fused image contains the maximum amount of complementary information transferred from the input images. It is formulated as follows:

$$\mathrm{SCD} = r(D_1, S_1) + r(D_2, S_2) \tag{8}$$

where $r(.)$ function calculates correlation between S_1 and D_1, and S_2 and D_2 as:

$$r(D_k, S_k) = \frac{\sum\limits_i \sum\limits_j \left(D_k(i,j) - \overline{D}_k\right)\left(S_k(i,j) - \overline{S}_k\right)}{\sqrt{\left(\sum\limits_i \sum\limits_j \left(D_k(i,j) - \overline{D}_k\right)^2\right)\left(\sum\limits_i \sum\limits_j \left(S_k(i,j) - \overline{S}_k\right)^2\right)}} \tag{9}$$

where k = 1, 2 and $\overline{S}_k$ and $\overline{D}_k$ are the average of the pixel values of S_k and D_k, respectively.

The *SSIM* [27] is an effective measure of structural similarity between two different images. It was first proposed by the University of Texas at Austin and the Laboratory of Image and Video Engineering, and combines the three components of luminance, structure, and contrast to comprehensively measure image quality. Let X be the reference image and Y be a test image, which is described as follows:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{\left(\mu_X^2 + \mu_Y^2 + C_1\right)\left(\sigma_X^2 + \sigma_Y^2 + C_2\right)} \tag{10}$$

where μ and σ denote the mean and standard deviation, respectively, and σ_{XY} is the cross-correlation between X and Y. C_1 and C_2 are stable coefficients when the mean value and the variance are close to zero, which are two small constants. The standard deviation of the Gaussian window is set to 1.5 in the calculation.

The *SSIMa* is calculated by Equation (10):

$$SSIM_a(F) = (SSIM(F, I_1) + SSIM(F, I_2)) \times 0.5 \tag{11}$$

where $SSIM(\cdot)$ denotes the structural similarity operation, F is the fused image, and I_1, and I_2 are source images. The value of $SSIM_a$ represents the ability to preserve structural information.

MS-SSIM is a multi-scale *SSIM* method for image quality assessment:

$$MS - SSIM(x,y) = [l_m(x,y)]^{\alpha M} \cdot \prod_{M}^{j=1} \left[c_j(x,y)\right]^{\beta_j} \left[s_j(x,y)\right]^{\gamma_j} \tag{12}$$

where $c_j(x,y)$ and $s_j(x,y)$ indicate contrast comparison and structure comparison, respectively, and 1 indicates brightness comparison. The exponents αM, β_j and γ_j are used to adjust the relative importance of different components.

Information entropy (En) reflects the richness of information in an image. In general, the greater the amount of information in an image, the greater the information entropy, whose mathematical expression is as follows:

$$En = -\sum_{L}^{x=0} p(x) \log_2 p(x) \tag{13}$$

Q^{NCIE} measures nonlinear correlation entropy between the source image and the fused image, and is defined as follows:

$$Q^{NCIE} = 1 + \sum_{3}^{i=1} \frac{\lambda_i}{3} \log_{256}\left(\frac{\lambda_i}{3}\right) \tag{14}$$

where λ_i is the eigenvalue of the nonlinear correlation matrix.

For each of these metrics, the largest score indicates the best fusion performance.

4.2. Fusion Method Evaluation

The fused images obtained from multiple methods are shown in Figures 4–6. Because CBF and VSMWLS only use grayscale maps for fusion, image color is not considered, and only texture information is compared.

As shown in the red box in Figure 4, the fused images obtained by CBF, VSMWLS, Densefuse-Add, and Densefuse-L1 do not naturally overlap objects such as trees. However, there is a natural excess of leaf edge in our algorithm.

The sky is frequently depicted as the background in images with water. Additionally, the typical fusion algorithms frequently result in color breaks in the sky, as shown in Figure 5, which does not occur in our algorithm. As can be seen from Figure 5, the ripples on the water surface are cleaner in our algorithm compared with other algorithms, reducing the interference in the image.

Figure 6 shows how different algorithms preserve some IR-specific information in the fused image. However, our algorithm can preserve as much infrared data as possible while keeping the brightness of the sky.

In our proposed fusion algorithm, Proposed-73, Proposed-82, and Proposed-91 have similar but different performance. Proposed-73 has some distortion in color performance as seen in Figure 4; Proposed-91 has some overexposure in bright areas; and Proposed-82 has the best performance in general.

In addition to the subjective evaluation, a few objective evaluation indicators were established. The average values of seven metrics for 18 fused images obtained by the existing methods and the proposed fusion method are shown in Table 1.

Table 1. The average values of quality metrics.

Methods	$Q^{AB/F}$	**SCD**	*SSIM*	*MS-SSIM*	*EN*	Q^{NCIE}
CBF	0.361383	0.6890	0.591628	0.744972	**7.626591**	0.810328431
VSMWLS	0.361481	1.4520	0.683922	0.831989	7.366922	0.810889691
Densefuse-Add	0.373791	1.2658	**0.729981**	0.845292	6.965533	0.811113586
Densefuse-L1	0.371051	1.0967	0.715435	0.841639	6.847064	0.811168666
Ours-91	0.397623	**1.2879**	0.712871	0.855669	6.790240	**0.818237040**
Ours-82	**0.400324**	1.2267	0.714010	**0.868377**	6.848320	0.817008603
Ours-73	0.369366	1.2059	0.715949	0.846516	6.827703	0.817501251

In Table 1, The best value in the objective indicator is shown in bold. It can be found that our algorithm has four optimal values in six metrics. The optimal values of $Q^{AB/F}$ and SCD indicate that our algorithm is more natural and has lower noise, while the optimal values of *MS-SSIM* and Q^{NCIE} indicate that the structural information is well retained.

Our algorithm performs better in both subjective evaluation and objective metrics, which demonstrates that it is an effective IR-visible fusion algorithm for the surface water environment. Our algorithm still has some shortcomings. During the operation of the machine used in this paper, it took about 0.156 s to test one image. Therefore, our algorithm is still some distance away from completing real-time video tasks.

5. Conclusions

In this study, a novel and effective deep learning architecture based on CNN and dense block is proposed for the infrared and visible image fusion task. By combining the intermediate image I_C created from two input images, it can more effectively describe the fusion task by balancing multiple inputs and a single output. The algorithm works well for water surface images.

The network consists of an encoder, a fusion layer, and a decoder, and the loss function consists of *SSIM*, *TV*, and *MSE*. The fused images I_{C1}, I_{C2}, and I_{C3} are introduced to better guide the network to complete the fusion task.

In addition, suitable parameters are set and compared with various methods, which achieved better results in some indicators. As a result, our algorithm produces better results and is specifically designed for the water surface scenario.

The goal of the infrared-visible fusion task is to combine the information from infrared and visible images while maintaining the color and texture information from each image separately. In addition to enhancing feature extraction capability, the challenge of this task is how to better describe the task objective. For future research, we aim to improve the real-time performance of the fusion algorithm and apply it to video fusion.

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

Funding: This research was funded by the National Natural Science Foundation of China, grants 52025111 and 51939003, and in part by the Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory, grant 2021JCJQ-SYSJJ-LB06911.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from MIT Sea Grant Marine Autonomy Dataset Project and are available from https://seagrant.mit.edu/auv-lab/ (accessed on 20 October 2022) with the permission of the MIT Sea Grant Marine Autonomy Dataset Project.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ben Hamza, A.; He, Y.; Krim, H.; Willsky, A. A multiscale approach to pixel-level image fusion. *Integr. Comput. Eng.* **2005**, *12*, 135–146. [CrossRef]
2. Zhang, Q.; Fu, Y.; Li, H.; Zou, J. Dictionary learning method for joint sparse representation-based image fusion. *Opt. Eng.* **2013**, *52*, 057006. [CrossRef]
3. Gao, R.; Vorobyov, S.A.; Zhao, H. Image Fusion with Cosparse Analysis Operator. *IEEE Signal Process. Lett.* **2017**, *24*, 943–947. [CrossRef]
4. Zong, J.-J.; Qiu, T.-S. Medical image fusion based on sparse representation of classified image patches. *Biomed. Signal Process. Control.* **2017**, *34*, 195–205. [CrossRef]
5. Li, H.; Wu, X.-J. Multi-focus Image Fusion Using Dictionary Learning and Low-Rank Representation. In *Image and Graphics*; Zhao, Y., Kong, X., Taubman, D., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 675–686. [CrossRef]
6. 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]
7. Melekhov, I.; Kannala, J.; Rahtu, E. Siamese network features for image matching. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 378–383. [CrossRef]
8. Wu, W.; Qiu, Z.; Zhao, M.; Huang, Q.; Lei, Y. Visible and infrared image fusion using NSST and deep Boltzmann machine. *Optik* **2018**, *157*, 334–342. [CrossRef]
9. Li, H.; Wu, X.-J.; Kittler, J. Infrared and Visible Image Fusion using a Deep Learning Framework. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018. [CrossRef]
10. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
11. Li, H.; Wu, X.-J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef]
12. Pu, T. Contrast-based image fusion using the discrete wavelet transform. *Opt. Eng.* **2000**, *39*, 2075. [CrossRef]
13. Burt, P.; Adelson, E. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [CrossRef]
14. Do, M.; Vetterli, M. The contourlet transform: An efficient directional multiresolution image representation. *IEEE Trans. Image Process.* **2005**, *14*, 2091–2106. [CrossRef]

15. Da Cunha, A.; Zhou, J.; Do, M. The Nonsubsampled Contourlet Transform: Theory, Design, and Applications. *IEEE Trans. Image Process.* **2006**, *15*, 3089–3101. [CrossRef]
16. Luo, X.; Zhang, Z.; Zhang, B.; Wu, X. Image Fusion with Contextual Statistical Similarity and Nonsubsampled Shearlet Transform. *IEEE Sensors J.* **2016**, *17*, 1760–1771. [CrossRef]
17. Yang, X.; Wang, J.; Zhu, R. Random Walks for Synthetic Aperture Radar Image Fusion in Framelet Domain. *IEEE Trans. Image Process.* **2017**, *27*, 851–865. [CrossRef]
18. Quan, S.; Qian, W.; Guo, J.; Zhao, H. Visible and infrared image fusion based on Curvelet transform. In Proceedings of the The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), Shanghai, China, 15–17 November 2014; pp. 828–832. [CrossRef]
19. Ahmed, N.; Natarajan, T.; Rao, K. Discrete Cosine Transform. *IEEE Trans. Comput.* **1974**, *C-23*, 90–93. [CrossRef]
20. 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]
21. Prabhakar, K.R.; Srikar, V.S.; Babu, R.V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4724–4732. [CrossRef]
22. Hou, R.; Zhou, D.; Nie, R.; Liu, D.; Xiong, L.; Guo, Y.B.; Yu, C. VIF-Net: An Unsupervised Framework for Infrared and Visible Image Fusion. *IEEE Trans. Comput. Imaging* **2020**, *6*, 640–651. [CrossRef]
23. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
25. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, 6–11 July 2015; JMLR.org: Lille, France, 2015; Volume 37, pp. 448–456.
26. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the Fourteenth International Con-ference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
27. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
28. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, 32.
29. Kumar, B.K.S. Image fusion based on pixel significance using cross bilateral filter. *Signal Image Video Process.* **2013**, *9*, 1193–1204. [CrossRef]
30. 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]
31. Xydeas, C.; Petrović, V. Objective image fusion performance measure. *Electron. Lett.* **2000**, *36*, 308–309. [CrossRef]
32. Aslantas, V.; Bendes, E. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-Int. J. Electron. Commun.* **2015**, *69*, 1890–1896. [CrossRef]
33. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003. [CrossRef]
34. Klonus, S.; Ehlers, M. Performance of evaluation methods in image fusion. In Proceedings of the 2009 12th International Conference on Information Fusion, Seattle, WA, USA, 6–9 July 2009; pp. 1409–1416.

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.