ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fbioe.2022.923364

A Two-To-One Deep Learning General Framework for Image Fusion

Pan Zhu[1,2,3], *Wanqi Ouyang*[1,2,3]*, *Yongxing Guo*[1,2,3] *and Xinglin Zhou*[1,2,3]

[1]*Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, Wuhan University of Science and Technology, Wuhan, China,* [2]*Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan, China,* [3]*Precision Manufacturing Institute, Wuhan University of Science and Technology, Wuhan, China*

The image fusion algorithm has great application value in the domain of computer vision, which makes the fused image have a more comprehensive and clearer description of the scene, and is beneficial to human eye recognition and automatic mechanical detection. In recent years, image fusion algorithms have achieved great success in different domains. However, it still has huge challenges in terms of the generalization of multi-modal image fusion. In reaction to this problem, this paper proposes a general image fusion framework based on an improved convolutional neural network. Firstly, the feature information of the input image is captured by the multiple feature extraction layers, and then multiple feature maps are stacked along the number of channels to acquire the feature fusion map. Finally, feature maps, which are derived from multiple feature extraction layers, are stacked in high dimensions by skip connection and convolution filtering for reconstruction to produce the final result. In this paper, multi-modal images are gained from multiple datasets to produce a large sample space to adequately train the network. Compared with the existing convolutional neural networks and traditional fusion algorithms, the proposed model not only has generality and stability but also has some strengths in subjective visualization and objective evaluation, while the average running time is at least 94% faster than the reference algorithm based on neural network.

Keywords: bionic vision, multi-modal image fusion, convolutional neural network, y-distribution structure, multi-convolution kernel, adaptive feature analysis

Edited by:
Tinggui Chen,
Zhejiang Gongshang University, China

Reviewed by:
Javad Hassannataj Joloudari,
University of Birjand, Iran
Yongfeng Li,
Henan Institute of Science and Technology, China

Correspondence:
Wanqi Ouyang
ouyangwanqi@163.com

Specialty section:
This article was submitted to Bionics and Biomimetics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *19 April 2022*
Accepted: *09 June 2022*
Published: *14 July 2022*

Citation:
Zhu P, Ouyang W, Guo Y and Zhou X (2022) A Two-To-One Deep Learning General Framework for Image Fusion. Front. Bioeng. Biotechnol. 10:923364. doi: 10.3389/fbioe.2022.923364

1 INTRODUCTION

Deep learning is a bio-inspired intelligent computing technology that is based on the principles of neurotransmission processes in the human brain, which resembles the pattern of connections between brain neurons (Xu et al., 2021). Unlike classical bionic techniques, i. e., ant colony algorithms (Deng et al., 2020), bee algorithms (Çil et al., 2020), etc., and particle swarm optimization (Elbes et al., 2019), etc., deep learning has an incredible and impressive ability to resolve the complexity of real-world problems, which has caused the attention of many scholars and has been successfully applied to practical problems (Chen et al., 2021b; Chen et al., 2022a; Chen et al., 2022c; Sun et al., 2022). In recent years, deep learning, especially neural networks, has become one of the most rapidly growing and widely applied artificial intelligence technologies. Several studies have demonstrated the superior performance of neural networks in target detection (Jiang et al., 2021a; Huang et al., 2021; Huang et al., 2022), image segmentation (Jiang et al., 2021b), data processing (Chen et al., 2021a; Chen et al., 2022b), and depth estimation (Jiang et al., 2019), etc. In addition, image fusion, which is an essential branch of neural network research, has been extensively

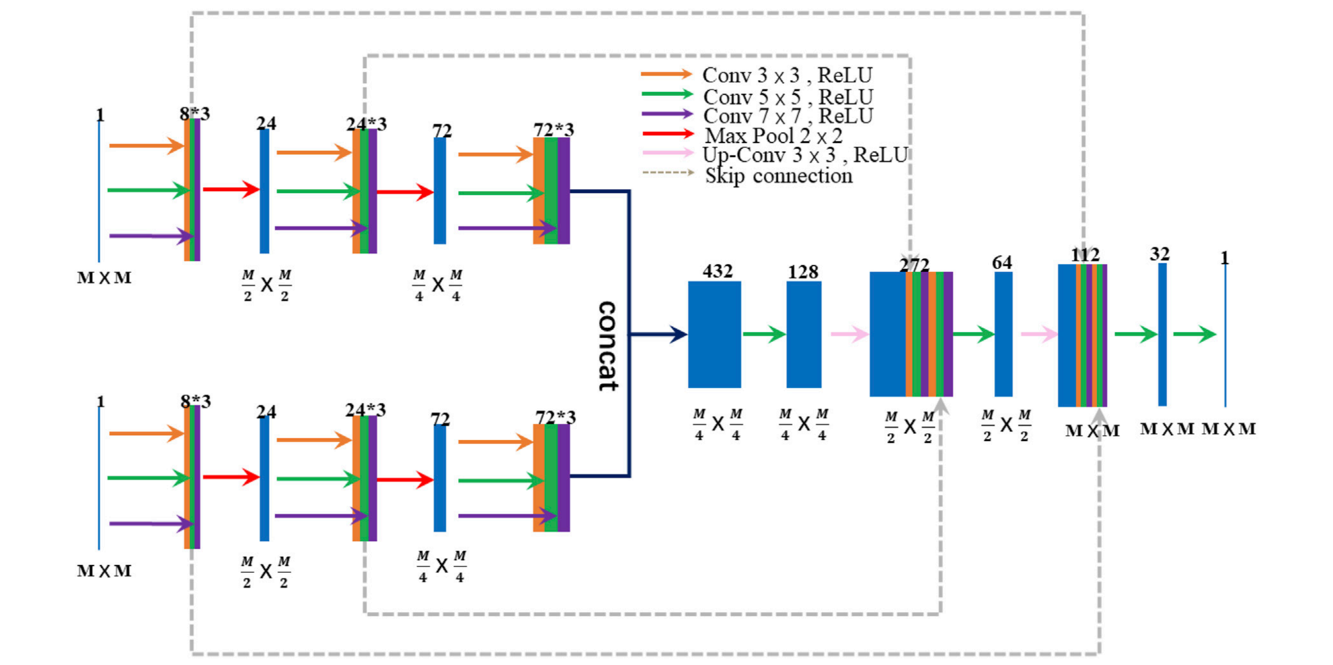
Conv 3 x 3 , ReLU
Conv 5 x 5 , ReLU
Conv 7 x 7 , ReLU
Max Pool 2 x 2
Up-Conv 3 x 3 , ReLU
Skip connection
1
8*3
24
24*3
72
72*3
M X M
$\frac{M}{2}$ X $\frac{M}{2}$
$\frac{M}{4}$ X $\frac{M}{4}$
concat
432
128
272
64
112
32
1
$\frac{M}{4}$ X $\frac{M}{4}$
$\frac{M}{4}$ X $\frac{M}{4}$
$\frac{M}{2}$ X $\frac{M}{2}$
$\frac{M}{2}$ X $\frac{M}{2}$
M X M
M X M
M X M
1
8*3
24
24*3
72
72*3
M X M
$\frac{M}{2}$ X $\frac{M}{2}$
$\frac{M}{4}$ X $\frac{M}{4}$

FIGURE 1 | The architecture of IY-Net. M represents the size of the feature map. The number at the top block represents the feature depth.

implemented in various areas, especially in civil, military, and industrial applications, since the research on neural networks has gradually advanced. For example, mobile phones often integrate with high dynamic range (Ma et al., 2015; Liu et al., 2018; Qi et al., 2021) or refocus algorithms (Saha et al., 2013; Bai et al., 2015; Zhang and Levine, 2016) to get stable and information-rich images. Visible and infrared image fusion can provide a more direct monitoring environment to the observers (Xue and Blum, 2003; Wan et al., 2009; Zhou et al., 2016; Zhang et al., 2017).

Convolutional neural network (CNN), which is a category of neural networks, usually is superior to traditional manual feature extractors in feature extraction (Yan et al., 2017; Li et al., 2018), and the number of convolutional filters is significantly larger than traditional filters. Therefore, CNN can capture richer image details and is frequently used for image feature extraction. As such a potent tool, CNN provides new ideas and directions for research on image fusion. In general, neural networks enable to excavate of implicit rules in massive datasets and then predict the result by the gained rules, which render the models with exceptional generalization ability (Cheng et al., 2021; Huang et al., 2021). For traditional image fusion algorithms, multi-modal image fusion usually implies different fusion rules and it is difficult to seek a harmonized approach. As for CNN, CNN is not fully exploited in most cases and is primarily applied for image feature extraction. Although a few fully convolutional neural networks, which don't need to impose preprocessing and fusion rules, can automate image fused, the fusion object is specified for single-modal images. Therefore, the study of the generality of multi-modal image fusion faces a tremendous challenge.

In this paper, a general CNN framework for image fusion, called IY-Net, is designed. The structure of IY-Net is shown in **Figure 1**. The proposed model has two innovations. First of all, the proposed model has the characteristics of a fully convolutional neural network with relatively good generality. It doesn't need to specify fusion rules and has a simple network structure. This is the key innovation point. Secondly, since the quality of training datasets constrains the model performance in the field of deep learning, the appropriate dataset is particularly critical. Theoretically, the performance of the model that is gained by using images of the same modal as the training dataset is more stable and accurate. However, this paper selects multi-modal images as the training dataset, and the proposed model can avoid the mutual influence of fusion results in some way. Thus, these two innovations can make the proposed model stand out from the current CNN methods.

The main contribution of this work is to propose a general image fusion framework. It is superior to many traditional algorithms and CNN methods in terms of image visual effects. The proposed model achieves excellent performance in multi-focus, infrared and visible, multi-exposure image fusion, etc. There are two more specific contributions. Firstly, a multi-feature extraction module is introduced, which effectively extends the perceptual field of the convolutional layer and thus captures more feature information. Secondly, a way of image reconstruction is constructed to effectively solve the problem of gradient disappearance and gradient explosion caused by CNN.

The rest of this paper is organized as follows. In **Section 2**, the paper discusses the related work. **Section 3** introduces the proposed model in detail. **Section 4** describes the experimental

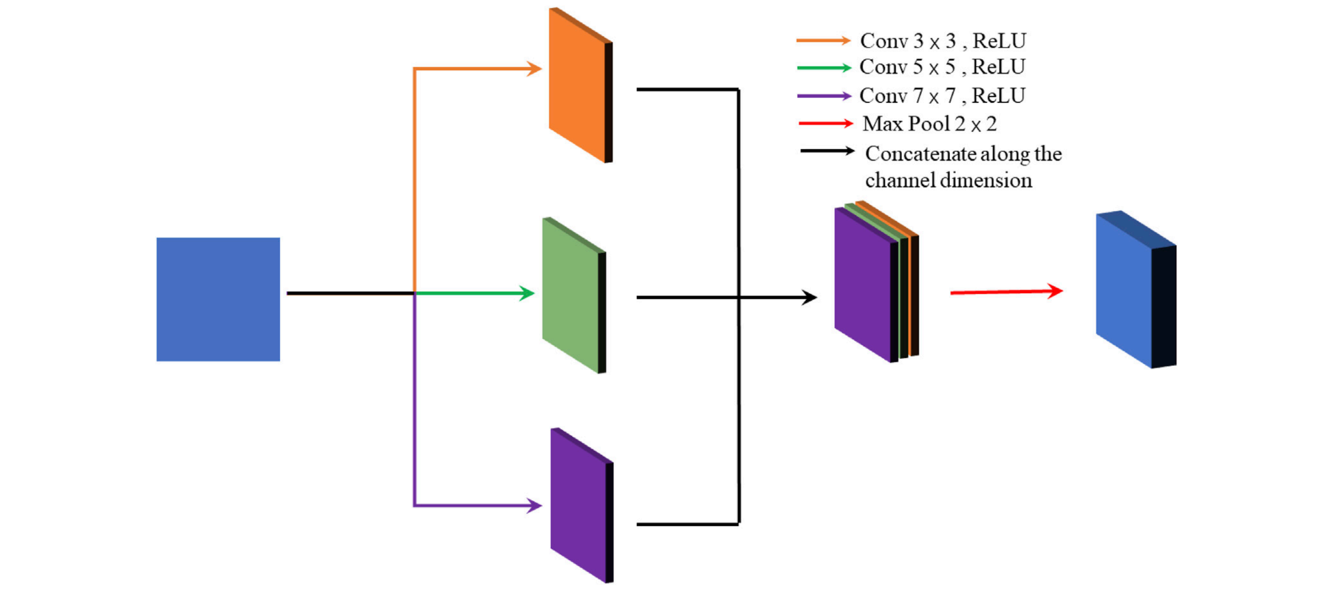

FIGURE 2 | Structure of multi-feature extraction layer.

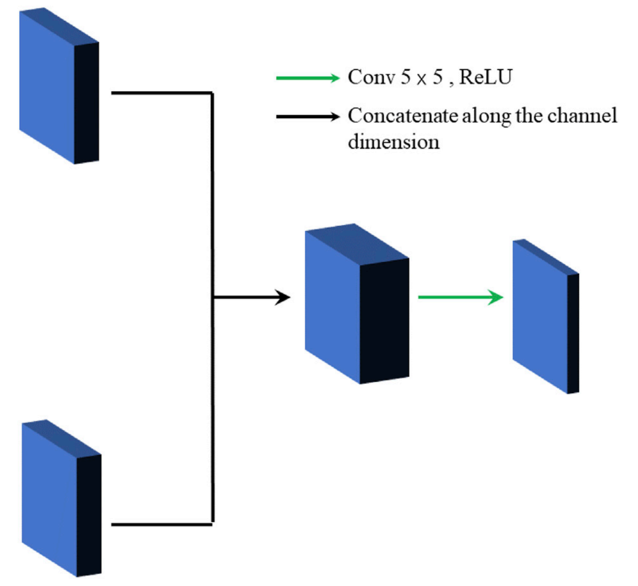

FIGURE 3 | Feature fusion structure.

results and discusses them. In **Section 5**, the paper shows the conclusion and future research directions.

2 RELATED WORK

Regarding CNN and traditional algorithms, despite several research results that have been achieved in image fusion algorithms, there is still space for optimization and improvement. In addition, most methods can only address image fusion of a few patterns and lack generality.

In general, traditional image fusion algorithms can be divided into two categories, i. e., spatial domain and transform domain algorithms. For image fusion algorithms in the spatial domain (Huang and Jing, 2007; Zhou et al., 2014; Zhang et al., 2017; Amin-Naji et al., 2022), the source image is divided into small pieces or regions according to certain criteria in the first step. Then the significance of the corresponding regions is evaluated, and finally, the most critical regions are fused. These algorithms are mainly applied to same-mode images, which may reduce the edge sharpness and contrast of the fused image or even produce halos at the edges. On the other hand, for the transform domain image fusion algorithm (Haghighat et al., 2011), the source image is decomposed into a feature domain by multi-scale geometry at the first step. Then, feature weighted fusion is achieved on multiple input images, and finally, the fused image is gained by the inverse transformation of the fused features. Among the current transform domain algorithms, multi-scale transform image fusion algorithms (MSTIF) are becoming increasingly popular. Examples of such transforms include pyramid-based decomposition (Liu et al., 2001), curvelet transform (Tessens et al., 2007), dual-tree complex wavelet transform (DTCWT) (Lewis et al., 2007), discrete wavelet transform (DWT) (Zheng et al., 2007; Tian and Chen, 2012) and non-subsampled contourlet transform (NSCT) (Moonon and Hu, 2015), etc. MSTIF relies on the selection of multi-scale decomposition methods and fusion strategies for multi-scale coefficient fusion. As a result, such algorithms have a relatively high manual factor, which leads to obvious weaknesses and lack of generality. For example, NSCT is weak at capturing curve details and curvelet transform is computationally complex, as well as it is terrible at multi-exposure and remote sensing image fusion. While fusing some modal images, pyramid-based decomposition will be distorted

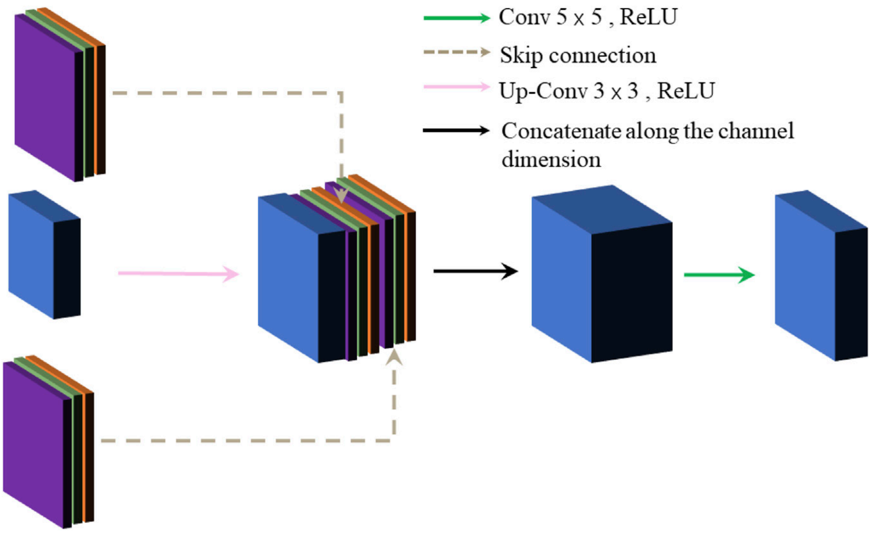

FIGURE 4 | Up-sampling operations and skip connection structure.

and laplace pyramid transform will incur redundant information, which is not available to infrared and visible image fusion. In conclusion, traditional MSTIF has a wide variety of filters, but it is always restricted in terms of the generality.

In recent years, image fusion methods based on neural networks have been rapidly growing (Liu et al., 2018). Firstly (Liu et al., 2017), regarded the fusion of multi-focus images as a classification task and used CNN to predict the focus image to obtain the fused image (Song et al., 2018). applied two neural networks to perform super-resolution processing of low-resolution terrestrial images and extract the feature map. Then high-pass modulation and weighting strategies are used to reconstruct the feature maps into fused images (Bhalla et al., 2022). integrated fuzzy theory with Siamese convolutional network to extract salient features of the source image as well as high-frequency information, and finally acquired fusion results by pixel strategy directly mapping to the source image. The above methods require pre-processing to generate fused images. In addition, they can only fuse images of a single-modal and lacks generality (Zhang et al., 2020). proposed a CNN-based image fusion framework that is trained in an end-to-end manner, and the parameters of the model can be jointly optimized without any subsequent processing. Although they designed a generalized model, it adopted human-selected fusion rules in the feature fusion phase, which led to the degradation of the model generality and the image fusion performance. For example, when infrared and visible images are fused, the model applies MAX fusion features to yield the best result. But when multi-exposure images are fused, it employs SUM fusion features to gain the best result. In summary, although CNN has achieved some success in the domain of image fusion, the majority of current models lack generality. In addition, most CNN is not designed end-to-end (Wang et al., 2019a) and requires additional steps to complete the task. Therefore, the CNN-based image fusion model has not been fully exploited, and there is still much potential to be boosted in terms of generality.

3 METHODS AND MATERIALS

3.1 Feature Extraction Module

The convolutional layer in CNN extracts different feature information from the training image by convolutional kernels and then updates the filter parameters automatically. Therefore, the selection of convolutional kernels is crucial for feature extraction. The specific structure is shown in **Supplementary Figure S1**. The small-size convolution kernel is used to extract the low-frequency and small detail information, while high-frequency and large detail information can’t be detected. Likewise, the large size of the convolution kernel is preferable for identifying high-frequency and large detail information.

As stated above, the paper utilizes multiple feature extraction layers, each of which has convolution kernels of sizes 3 × 3, 5 × 5, and 7 × 7, to capture low and high-frequency information. The specific structure is shown in **Figure 2**. The proposed model detects the feature information of the input image by three multiple feature extraction layers, but multiple convolutions can lead to over-fitting and increasing the training time. Therefore, this paper adds a max-pooling layer after both of the two previous multiple feature extraction layers to avoid such phenomena.

3.2 Feature Fusion Module

There are two general methods for feature fusion: 1) The feature maps are connected along with the number of channels. 2) The feature maps are fused according to certain fusion rules. If the second feature fusion way is chosen, it will lead to a decrease in the generality of the model. Therefore, the paper chooses the first method to get the fused feature map. The specific structure is shown in **Figure 3**.

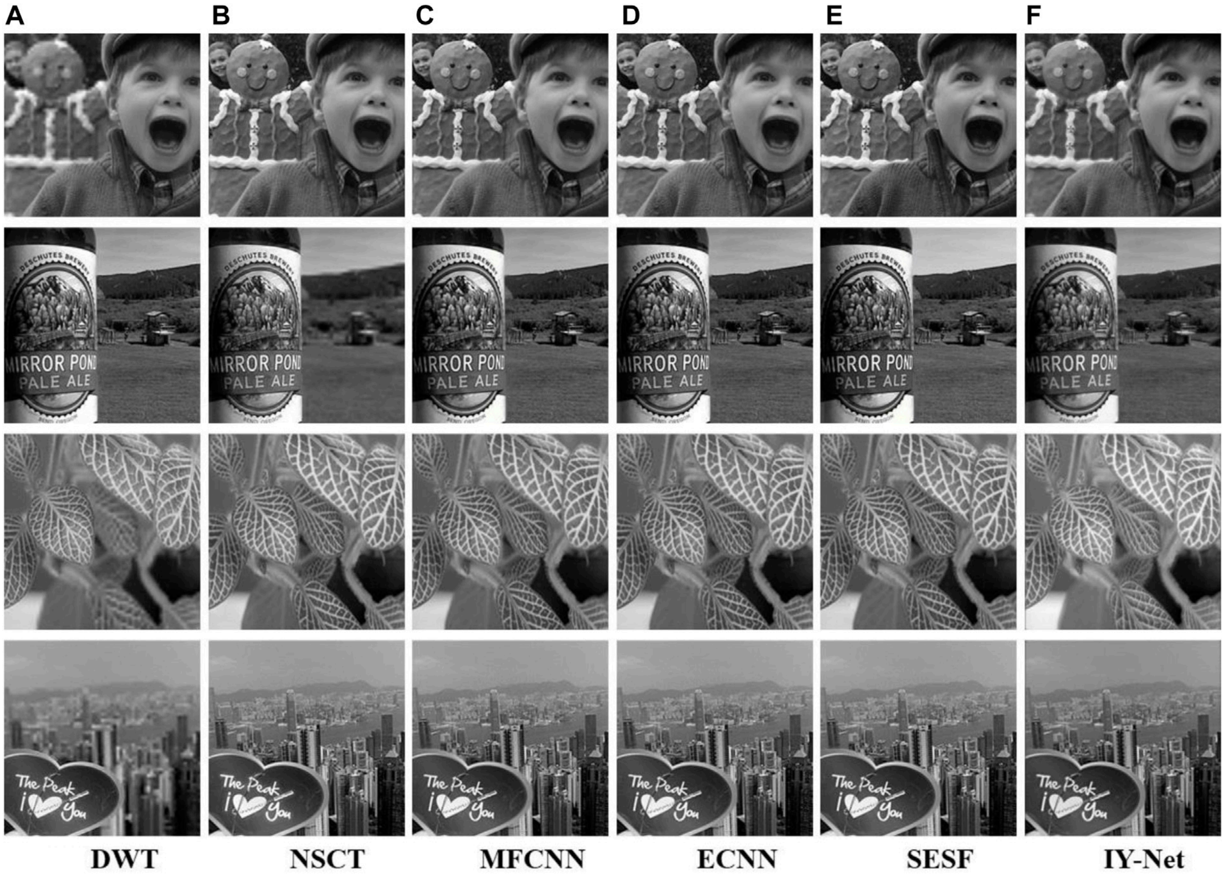

FIGURE 5 | Experiment on 4 pairs of multi-focus images. **(A)** DWT, **(B)** NSCT, **(C)** MFCNN, **(D)** ECNN, **(E)** SESF, **(F)** IY-Net.

Firstly, the feature maps are concatenated along the channel dimension to gain the initial feature fusion map, and then it is filtered by the convolution layer. Finally, it is down-dimensioned to produce the final cross-channel fused feature map.

3.3 Image Reconstruction Module

Under the effect of the pooling layer, the image size is changed from 256 × 256 to 64 × 64, which greatly reduces the resolution of the original image, and some features may be ambiguous. For restoring the size of the resource image, the paper applies the up-sampling operation (i.e., transposed convolution) to restore the resolution and optimize the image quality. However, it causes the image edge information to be dropped and blurred, so we deal with this problem by adding a skip connection based on the up-sampling operation, which can further enhance the image edge information. The module undergoes three up-sampling operations, which each time doubles the image size, and eventually produces a grayscale image with the original size. The specific up-sampling operations and skip connection structure are shown in **Figure 4**. Firstly, the feature map and the fused feature map are skip-connected, and then up-sampling operations are executed on them. Finally, the high-dimensional map is down-dimensioned to a low-dimensional map by convolutional layers.

3.4 Loss Function

Before training the model, it is necessary to optimize the model parameters using an appropriate loss function to compare the predicted values with the actual values. The proposed model aims to form a fused image by regression of two input images. Therefore, the paper chooses the structural similarity (SSIM) (Wang et al., 2004) to coping with this problem. As shown in the equation.

$$\boldsymbol{SSIM}(\boldsymbol{x},\boldsymbol{y}) = \frac{\left(2\boldsymbol{\mu}_x\boldsymbol{\mu}_y + \boldsymbol{C}_1\right)\left(2\boldsymbol{\sigma}_{xy} + \boldsymbol{C}_2\right)}{\left(\boldsymbol{\mu}_x^2 + \boldsymbol{\mu}_y^2 + \boldsymbol{C}_1\right)\left(\boldsymbol{\sigma}_x^2 + \boldsymbol{\sigma}_y^2 + \boldsymbol{C}_2\right)} \tag{1}$$

Where x is the real image, y is the predicted image, μ_x, μ_y is mean, σ_x, σ_y is variance, and σ_{xy} is covariance. $C_1 = (Lk_1)^2$, $C_2 = (Lk_2)^2$ are stable constants. L is the dynamic range of pixel values, $k_1 = 0.01$, $k_2 = 0.03$. The sliding window size is set as 11×11, it moves pixel by pixel in an image from top-left on an image.

Thus, SSIM loss function can be defined as:

$$\boldsymbol{L}_{ssim} = \frac{1}{\boldsymbol{n}} \sum 1 - \boldsymbol{SSIM}(\boldsymbol{x},\boldsymbol{y}) \tag{2}$$

Where n represents the total number of sliding windows.

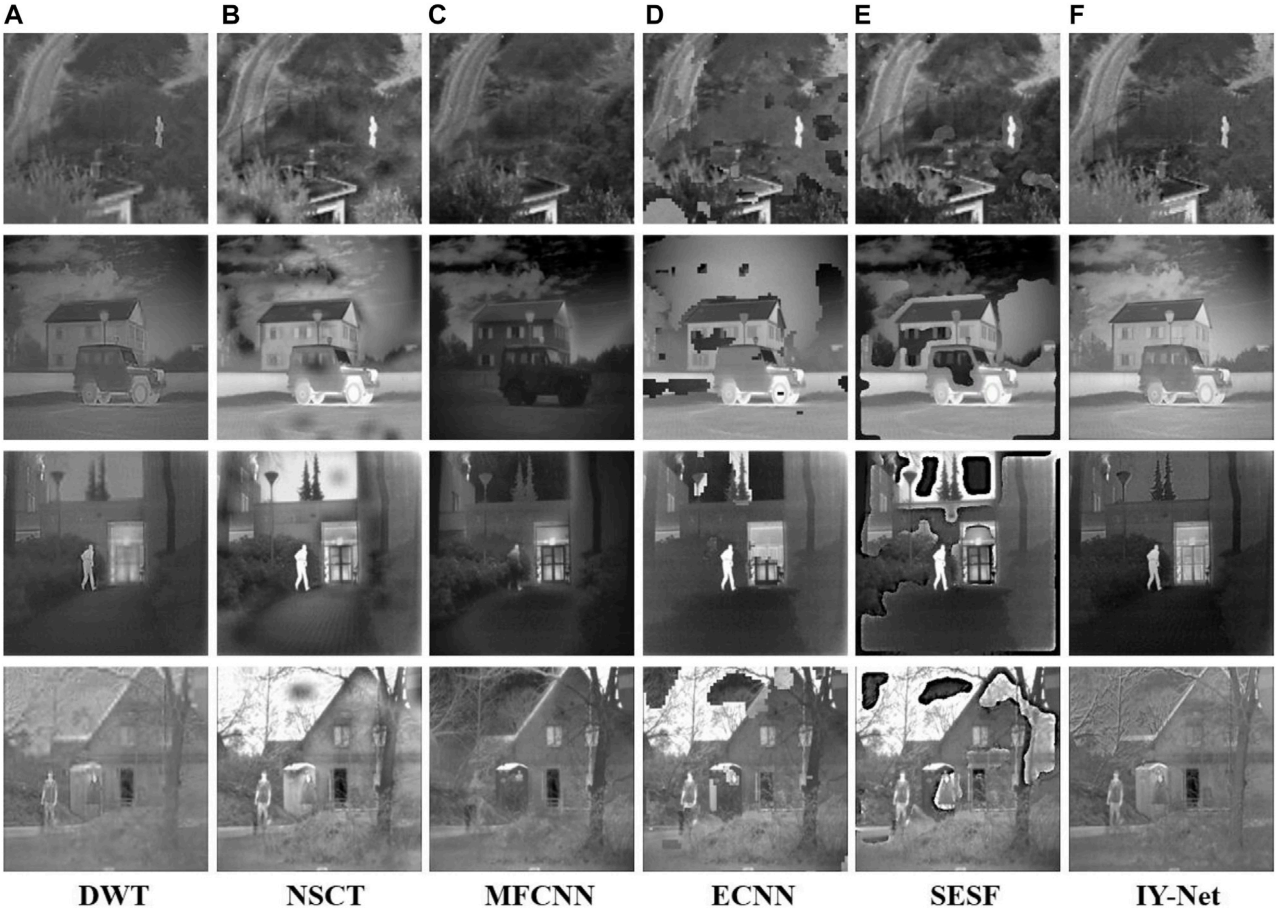

FIGURE 6 | Experiment on 4 pairs of infrared and visible images. **(A)** DWT, **(B)** NSCT, **(C)** MFCNN, **(D)** ECNN, **(E)** SESF, **(F)** IY-Net.

The proposed model has all components of the loss function that are differentiable, thus the model parameters of the paper can be updated by random gradient descent and back-propagation.

3.5 Training Dataset

It is well known that CNNs are data-driven. So large-scale image datasets are the basis for achieving favorable performance (Liu et al., 2017). randomly selected multi-focus images from the ImageNet dataset. And the focused images were obscured with a random scale of the Gaussian kernel to generate an image dataset consisting of 2 million pairs of images of size 16 × 16. Since no large-scale multi-exposure image dataset was available (Ram Prabhakar et al., 2017), randomly cropped 64 × 64 image segments from small multi-exposure images to generate a multi-exposure dataset.

As mentioned above, current experimental objects are composed mainly of small blocks of images as single-modal datasets, which can't fulfill the experimental requirements. Therefore, multi-focus images, multi-exposure images, and remote sensing images are selected from several datasets to form the training dataset with an image size of 256 × 256 in this paper. The images in the training dataset was are randomly rotated, randomly contrast shifted, and randomly stretched to boost diversity. The parts of multi-modal images in the dataset are shown in **Supplementary Figure S2.**

4 EXPERIMENTS AND RESULTS

4.1 Experimental Settings

IY-Net is implemented by Pytorch 1.8.1 based on Python 3.9.4. The proposed model is trained and tested on a computer equipped with an Intel i5-1035G1 CPU (1 GHz) and 2 GB GPU, and it is trained on the CPU. The paper trains 1826 pairs of images with an image size of 256 × 256 and a batch size of 40 in the training process. The whole process takes about 1 h. Concerning the learning rate, using the Adam optimizer (Wang et al., 2004) and the learning rate set to 0.0005.

In this paper, the proposed model is compared with traditional multi-scale transform algorithms, i. e., discrete wavelet transforms (DWT) (Zheng et al., 2007) and non-subsampled contourlet transform (NSCT) (Moonon and Hu, 2015). To further validate the advantages of the proposed model in the area of deep learning, it is compared with three current neural network-based image fusion models, i. e., multi-focus image fusion model (MFCNN) (Liu et al., 2017), CNN integration model for image fusion (ECNN) (Amin-Naji et al., 2019) and unsupervised depth model for image fusion (SESF) (Ma et al. 2021). To verify the generality of the proposed model, five types of

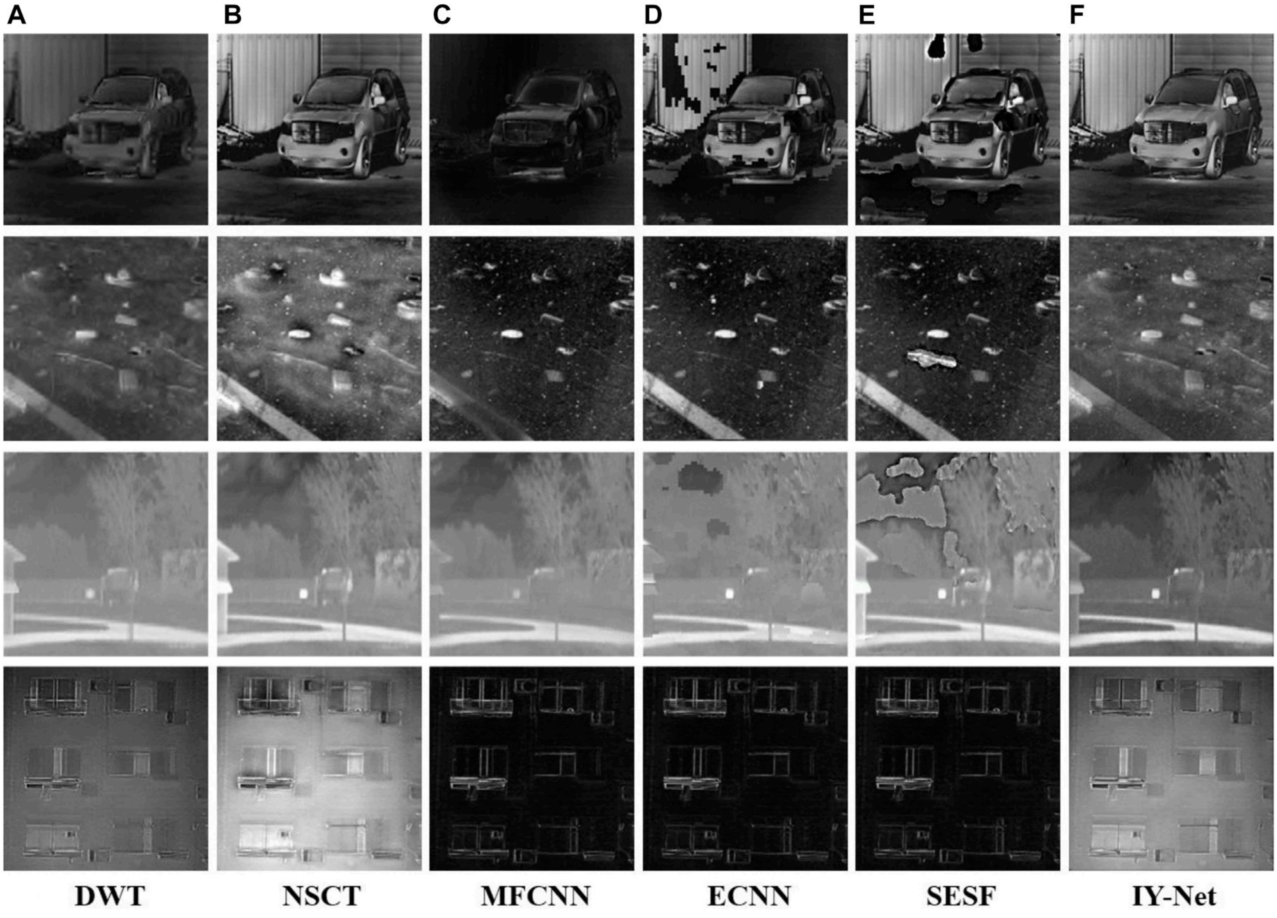

FIGURE 7 | Experiment on 4 pairs of infrared intensity and polarization images. **(A)** DWT, **(B)** NSCT, **(C)** MFCNN, **(D)** ECNN, **(E)** SESF, **(F)** IY-Net.

datasets (including multifocal images, infrared and visual images, etc.) are experimented and evaluated in the paper. The five image test datasets are shown in **Supplementary Figures S7,S8,S9,S10,S11**.

For the evaluation of the image fusion algorithm, the paper qualitatively judges the visual effect of the fused images. The performance of different image fusion methods can't be distinguished by visual effects alone. Therefore, five metrics are introduced to further estimate the quantitative manifestation of IY-Net on multi-modal image fusion. The five metrics are spatial frequency (SF), information entropy (IE), average gradient (AG) (Petrović, 2007), Peille index (Peille) (Piella and Heijmans, 2003), and edge preservation information (Q_{AB}) (Xydeas and Petrovic, 2000) respectively.

4.2 Experimental Results and Analysis

4.2.1 Multi-Focus Image Fusion

Experiments are conducted on multi-focus image test datasets as shown in **Supplementary Figure S3**. It is verified that the proposed model has a great performance in multi-focus image fusion. Taking "Boy" as shown in **Supplementary Figure S8 (A) and (B)** for example. The fusion result of DWT is blurred in some regions and fails to retain the complete details and features, but other algorithms can capture suitable feature information with better visual effects. **Figure 5** provides the fusion results of multi-focus image test datasets based on all algorithms. Experimental results show that the proposed model is practicable and stable in multi-focus image fusion visually.

4.2.2 Infrared and Visible Image Fusion

As shown in **Supplementary Figure S4**, four groups of infrared and visible images reveal different scene information. Experiments are carried on them to confirm the capability of IY-Net in infrared and visible image fusion. For simplicity, "Car" is used for detailed analysis in **Supplementary Figure S9**. Apparently, DWT basically preserves the infrared and visible features, but the fused image has relatively low contrast. MFCNN failed to capture the infrared features and the visual effect is weak. NSCT, ECNN, and SESF produce large areas of dark spots and shadows that generate no-desired results. Exhilaratingly, IY-Net acquires the most observable fusion results, which be provided with abundant visible details and infrared features as shown in **Supplementary Figure S9 (H)**. A similar situation occurs in **Figure 6** which is obtained from the images in **Supplementary Figure S4**. To all appearances, IY-Net not only has the best visual

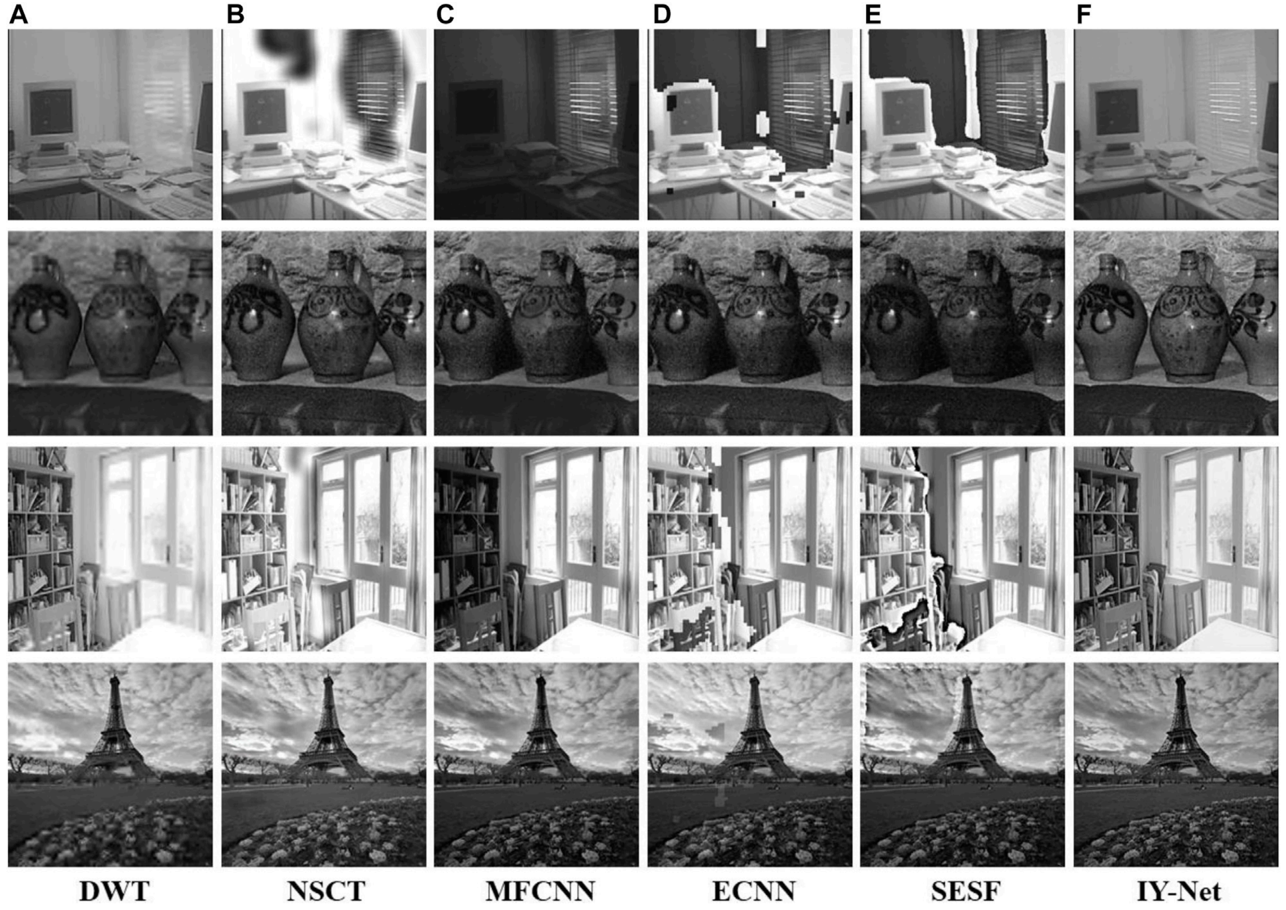

FIGURE 8 | Experiment on 4 pairs of multi-exposure images. **(A)** DWT, **(B)** NSCT, **(C)** MFCNN, **(D)** ECNN, **(E)** SESF, **(F)** IY-Net.

effect but also possesses evident stability and adaptability in infrared and visible image fusion.

4.2.3 Infrared Intensity and Polarization Image Fusion

Supplementary Figure S5 shows four pairs of infrared intensity and polarization images that are used to check the performance of the proposed model. A group of experimental results, taking "SUV" for example, is presented in **Supplementary Figure S10**. The source polarization and infrared intensity images are shown in **Supplementary Figure S10**. From the results of the experiment, we can see that DWT may maintain polarization and intensity information, but some parts are obscured, which results in poor visual effects. MFCNN cannot fuse the source image validly at all. ECNN and SESF can only combine the polarization and intensity information in part of the region and generate many pixel blocks and black spots, which seriously affects overall visual perception. In contrast, IY-Net and NSCT perfectly integrate these two kinds of images. It shows that NSCT and IY-Net could be employed availably in infrared intensity and polarization image fusion compared to other algorithms. The other fusion results are shown in **Figure 7** Experiments demonstrated that MFCNN, ECNN, and SESF failed to fuse infrared intensity and polarization images in a dark environment. In addition, it produces the phenomenon of image distortion and partial texture being blurred in bright environments. However, NSCT and IY-Net can be adapted for infrared intensity and polarization image fusion in different environments.

4.2.4 Multi-Exposure Image Fusion

Furthermore, fusion experiments are implemented in multi-exposure images as shown in **Supplementary Figure S6** to evaluate the capability of the proposed model. The source "Computer" image is shown in **Supplementary Figure S4 (A) and (B)**, and the two images show high and low exposure images. **Supplementary Figure S4 (C)-(H)** shows the fusion results of all algorithms. DWT can keep the source image features, but the region of the curtain is ambiguous. The fused results of NSCT, ECNN, and SESF appear with numerous black spots, and partial details of MFCNN failed to be preserved. In contrast, IY-Net saves the fully-featured texture and achieves great visual effect in multi-exposure image fusion. The results of all the test database fusion are shown in **Figure 8** DWT generates blurred textures in some regions. NSCT, MFCNN, and ECNN can effectively respond to fusion in dark environments, but they can lose efficacy for the images with bright information. SESF displays

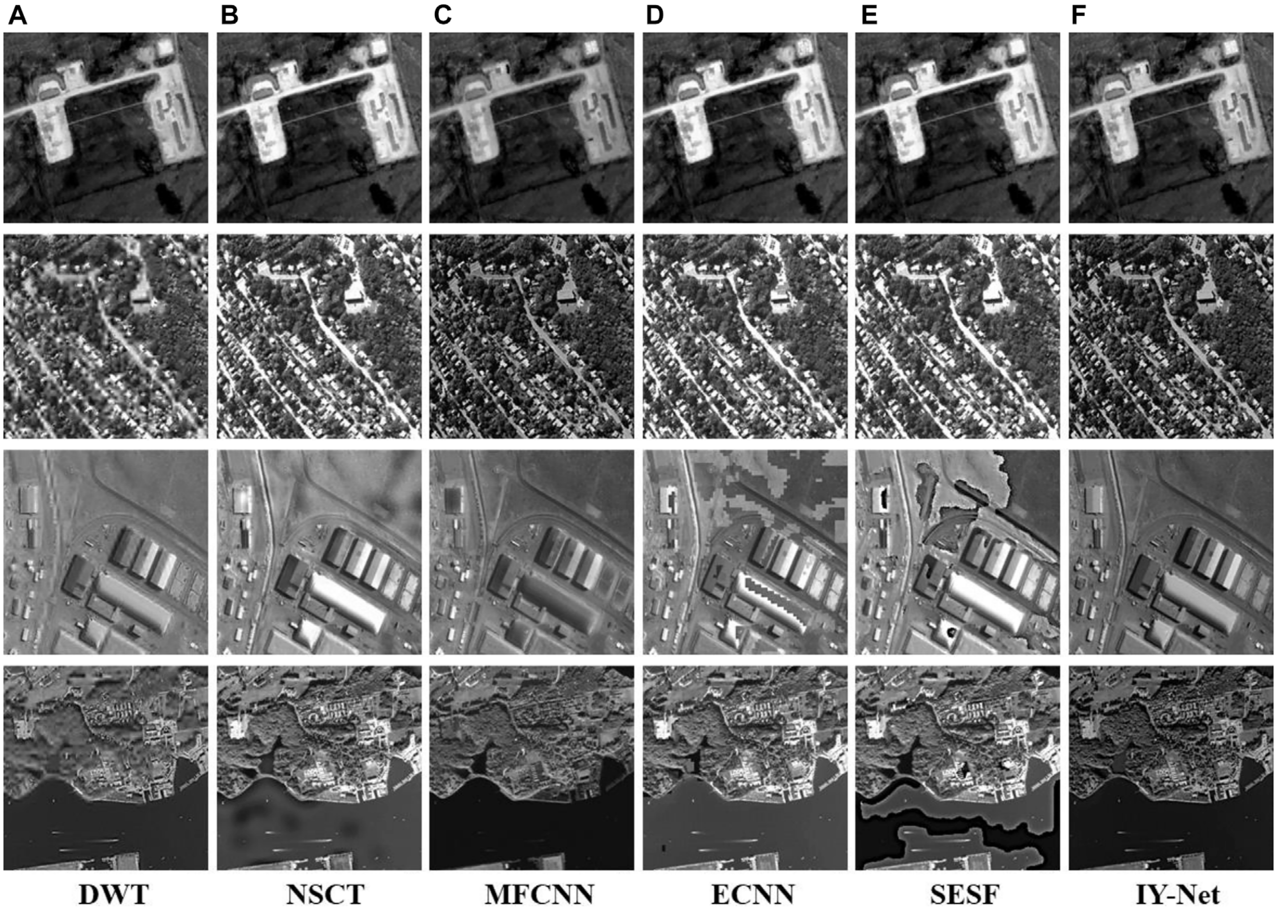

FIGURE 9 | Experiment on 4 pairs of remote sensing images. **(A)** DWT, **(B)** NSCT, **(C)** MFCNN, **(D)** ECNN, **(E)** SESF, **(F)** IY-Net.

TABLE 1 | Quantitative evaluation results of multi-focus image fusion.

Metrics	DWT	NSCT	MFCNN	ECNN	SESF	IY-Net
SF	21.41603	27.6822	27.5599	**29.5400**	29.4076	22.3491
AG	7.066	9.5068	9.3801	9.6744	**9.7212**	8.2074
IE	7.4132	7.4845	7.4694	**7.4783**	7.4713	7.4591
Q_{AB}	0.5012	0.7267	**0.7430**	0.7296	0.7212	0.6880
Peille	0.0076	0.0062	0.0065	0.0064	0.0072	**0.0090**

Bold indicates best values.

TABLE 2 | Quantitative evaluation results of infrared and visible image fusion.

Metrics	DWT	NSCT	MFCNN	ECNN	SESF	IY-Net
SF	8.1647	12.7831	9.5506	18.3357	**24.9147**	12.5291
AG	3.0915	5.0239	3.6153	5.4813	**7.2602**	4.8389
IE	6.4426	7.166	6.6088	7.1048	**7.3101**	6.8087
Q_{AB}	0.328	0.5085	0.4563	**0.5811**	0.5695	0.451
Peille	0.0064	0.0043	**0.0189**	0.0052	0.0129	0.0065

Bold indicates best values.

TABLE 3 | Quantitative evaluation results of infrared intensity and polarization image fusion.

Metrics	DWT	NSCT	MFCNN	ECNN	SESF	IY-Net
SF	8.8792	13.7983	11.1476	17.0852	**19.5567**	14.2838
AG	3.0686	5.1073	3.9309	5.4514	**6.1073**	5.1259
IE	6.4152	**7.1872**	5.987	6.2169	6.4883	6.9192
Q_{AB}	0.3268	0.528	0.5246	0.6061	**0.6156**	0.4627
Peille	0.0073	0.0042	**0.0358**	0.0242	0.0241	0.0049

Bold indicates best values.

TABLE 4 | Quantitative evaluation results of multi-exposure image fusion.

Metrics	DWT	NSCT	MFCNN	ECNN	SESF	IY-Net
SF	15.5429	23.285	19.9213	29.3432	**30.4046**	22.0839
AG	5.4503	8.7552	6.8542	9.6245	**9.8643**	8.0654
IE	7.1778	7.2096	7.1206	**7.3695**	7.2344	7.2672
Q_{AB}	0.4376	0.7668	0.6826	**0.7916**	0.7453	0.7074
Peille	0.0048	0.0027	**0.0103**	0.0036	0.0042	0.0037

Bold indicates best values.

TABLE 5 | Quantitative evaluation results of remote sensing image fusion.

Metrics	DWT	NSCT	MFCNN	ECNN	SESF	IY-Net
SF	24.9828	31.6788	25.4795	34.3604	**36.9394**	30.3675
AG	9.7509	12.6092	10.1474	12.9351	**13.7295**	11.9796
IE	7.0700	**7.2978**	6.7970	6.8664	6.9975	6.9814
Q_{AB}	0.4699	0.6895	0.6557	**0.7131**	0.7049	0.6580
Peille	0.0063	0.0048	**0.0106**	0.0054	0.0069	0.0103

Bold indicates best values.

TABLE 6 | Average running time of various algorithms (Time unit: second).

Method	DWT	NSCT	MFCNN	ECNN	SESF	IY-Net
Runtime	0.76	2.025	0.38	0.34	0.31	**0.16**

Bold indicates best values.

terrible results for different environments, for example, the fused images appear with extensive black spots and distortion of textures. In contrast to these reference algorithms, the proposed model is suitable for multi-exposure image fusion, and the fusion results reflect clearer features and appropriate visual perception.

4.2.5 Remote Sensing Image Fusion

Finally, this paper confirms the performance of the proposed model in remote sensing image fusion, and the test dataset is shown in **Supplementary Figure S7**. The source "Building" images are shown in **Supplementary Figures S12 (A) and (B). Supplementary Figures S12 (C)-(H)** show the fusion results of all algorithms. DWT, ECNN, SESF, and NSCT retain most of the detailed features, but some small details are vague. MFCNN and IY-Net can completely detect textures and details, nevertheless, IY-Net has higher contrast and more obvious intensity information than MFCNN. Concerning remote sensing image fusion, IY-Net has a better visual effect. Other fusion results are shown in **Figure 9**. Experiments reveal that DWT appears to texture distortion, and NSCT has too high contrast and thus obscures some feature information. MFCNN has only a visual perception of single-source image feature information, and ECNN and SESF have a lot of shadows and black spots locally. Obviously, the proposed model has a good visual effect and proper contrast.

4.3 Quantitative Comparison and Discussion

Table 1, **Table 2**, **Table 3**, **Table 4**, **Table 5** shows the quantitative metrics corresponding to the above multi-modal image fusion results respectively. In these tables, each value represents the average measured value of the dataset, and the best values are bolded. These metrics can be used to fairly and objectively reveal the fusion performance of all the algorithms from an objective perspective combined with subjective vision. As shown in **Table 1**, IY-Net acquires the optimum Peille metric, which denotes the proposed model is highly correlated with original images compared to these reference algorithms. Although the proposed model failed to yield optimal values for other metrics, the values achieved by the proposed model are acceptable.

As can be noticed from the objective metrics in **Table 2**, SESF acquires the greatest SF, AG and IE values, while ECNN gains the best Q_{AB} value. However, their fusion images present undesirable visual effects as shown in **Figure 6**. Although the congeneric values of the proposed model are not optimal, they are totally acceptable, especially combining the visual properties of the fusion results. It exhibits that the fusion result with the proposed method is equipped with rich detail and feature information from resource images.

Similar to the objective values in **Tables 1** and **2**, although the SESF obtained the best values in SF, AG and Q_{AB} in **Table 3**, it was also mainly caused by unreasonable distortion as shown in **Figure 7**. There are similar situations in DWT, MFCNN, and ECNN. Even though NSCT can achieve a similar visual effect to the proposed model, the SF, AG, and Peilla values are lower than IY-Net, which indicates that the proposed model has richer image sharpness and edge information, and is highly relevant to the source images.

In **Table 4**, although the best SF and AG values are attained by SESF and the best Q_{AB} and IE values were yielded by ECNN, it is resulting from the distorted and discordant fusion results as shown in **Figure 8**. In contrast to these reference algorithms, the proposed model is always stable in the expression of fusion results and the objective metrics are also acceptable, despite IY-Net being unable to highlight the advantages in every metric.

Similar to **Table 4**, SESF and ECNN in **Table 5** also produce abnormal SF, AG and Q_{AB} values caused by partial loss and distortion of image edge information. NSCT achieves a great IE value since some of the fusion results produce redundant feature information. Unlike these reference algorithms, the proposed model can provide excellent visual perception with sound objective values.

In addition to the visual analysis and objective evaluation metrics discuss, the average running time is an important indicator for evaluating algorithm performance. In **Table 6**, the average running times of all kinds of algorithms are displayed, where the shortest value is bolded. Apparently, the average running time of IY-Net is significantly optimal compared with these reference algorithms, and the proposed neural network model is at least 94% faster than these reference network algorithms. In general, the proposed model has a significant advantage in terms of average running time, compared to these reference algorithms.

Although the reference algorithms yield the best metrics for some modal images, the majority are overestimated due to the incongruous texture features in their fusion results, and they lack generality and stability for different patterns of images. For example, MFCNN, SESF, and ECNN achieved acceptable visual effects only in multi-focus image fusion, and DWT yielded favorable visual effects only in multi-exposure image fusion. As for NSCT, it is also inadequate in generality despite acquiring valuable visual effects in infrared intensity and polarization image fusion and multi-focus image fusion. In contrast, IY-Net can gain reasonable and acceptable quantitative metrics, and it also has significant strengths in the visual effects of multi-modal image fusion, while the computational speed is much faster than these reference algorithms. It reveals that the proposed model has premium generality, stability and rapidity. With the quantitative analysis and running time comparison, it is not difficult to realize that IY-Net achieves outstanding metrics in certain aspects, but there is still much progress to be expected.

5 CONCLUSION

In this paper, a general CNN framework for image fusion is proposed. Compared to current image fusion models, the proposed model has three main advantages: 1) Since it is fully convolutional, the model can be trained end-to-end and without pre-processing. 2) Although the training dataset is comprised of multi-modal images, the fused images not only have outstanding visual effects but also are not impacted by other modal images. 3) Its structure is similar to MSTIF, hence, it has outstanding generality in multi-modal image fusion. To summarize, IY-Net is superior to partial traditional multi-scale algorithms and existing neural network image fusion methods in terms of generality.

The proposed model provides the optimal visual effects compared to these reference algorithms through numerous fusion experiments, but the quantitative metrics are slightly inadequate. There are still several problems to be resolved to get a better-performing image fusion model. Firstly, this paper has a small training dataset, and increasing the large-scale sample may raise the model performance. Secondly, the proposed model consists of only three multiple feature extraction layers, which is relatively simplified, and the efficiency of the model can be enhanced by using a deeper network structure. Thirdly, the loss functions of the model are relatively simple, and the construction of more complex and optimized loss functions may enhance the stability and adaptability of the model.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

PZ provided the algorithmic ideas and theoretical analysis. WO performed the data processing and manuscript editing. YG guided the writing of the manuscript. All authors read and contributed to the manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (Grant No: 61901310, E080703, 51778509).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.923364/full#supplementary-material

REFERENCES

Amin-Naji, M., Aghagolzadeh, A., and Ezoji, M. (2019). Ensemble of CNN for Multi-Focus Image Fusion. *Inf. fusion* 51, 201–214. doi:10.1016/j.inffus.2019.02.003

Amin-Naji, M., Aghagolzadeh, A., and Mahdavinataj, H. (2022).Fast Multi Focus Image Fusion Using Determinant, 2022 International Conference on Machine Vision and Image Processing (MVIP). IEEE, 1–6. Ahvaz, Iran, Islamic Republic of, 23-24 Feb. 2022. doi:10.1109/MVIP53647.2022.9738555

Bai, X., Zhang, Y., Zhou, F., and Xue, B. (2015). Quadtree-based Multi-Focus Image Fusion Using a Weighted Focus-Measure. *Inf. Fusion* 22, 105–118. doi:10.1016/j.inffus.2014.05.003

Bhalla, K., Koundal, D., Bhatia, S., Khalid Imam Rahmani, M., Tahir, M., and Tahir, M. (2022). Fusion of Infrared and Visible Images Using Fuzzy Based Siamese Convolutional Network. *Comput. Mat. Con* 70, 5503–5518. doi:10.32604/cmc.2022.021125

Chen, T., Jin, Y., Yang, J., and Cong, G. (2022a). Identifying Emergence Process of Group Panic Buying Behavior under the COVID-19 Pandemic. *J. Retail. Consumer Serv.* 67, 102970. doi:10.1016/j.jretconser.2022.102970

Chen, T., Peng, L., Yang, J., Cong, G., and Li, G. (2021a). Evolutionary Game of Multi-Subjects in Live Streaming and Governance Strategies Based on Social Preference Theory during the COVID-19 Pandemic. *Mathematics* 9 (21), 2743. doi:10.3390/math9212743

Chen, T., Qiu, Y., Wang, B., and Yang, J. (2022b). Analysis of Effects on the Dual Circulation Promotion Policy for Cross-Border E-Commerce B2B Export Trade Based on System Dynamics during COVID-19. *Systems* 10 (1), 13. doi:10.3390/systems10010013

Chen, T., Rong, J., Yang, J., and Cong, G. (2022c). Modeling Rumor Diffusion Process with the Consideration of Individual Heterogeneity: Take the Imported Food Safety Issue as an Example during the COVID-19 Pandemic. *Front. Public Health* 10, 781691. doi:10.3389/fpubh.2022.781691

Chen, T., Yin, X., Yang, J., Cong, G., and Li, G. (2021b). Modeling Multi-Dimensional Public Opinion Process Based on Complex Network Dynamics Model in the Context of Derived Topics. *Axioms* 10 (4), 270. doi:10.3390/axioms10040270

Cheng, Y., Li, G., Yu, M., Jiang, D., Yun, J., Liu, Y., et al. (2021). Gesture Recognition Based on Surface Electromyography-Feature Image. *Concurrency Comput. Pract. Exp.* 33 (6), e6051. doi:10.1002/cpe.6051

Çil, Z. A., Li, Z., Mete, S., and Özceylan, E. (2020). Mathematical Model and Bee Algorithms for Mixed-Model Assembly Line Balancing Problem with Physical Human–Robot Collaboration. *Appl. soft Comput.* 93, 106394. doi:10.1016/j.asoc.2020.106394

Deng, W., Zhao, H., Song, Y., and Xu, J. (2020). An Effective Improved Co-evolution Ant Colony Optimisation Algorithm with Multi-Strategies and its Application. *Int. J. Bio-Inspired Comput.* 16 (3), 158–170. doi:10.1504/ijbic.2020.10033314

Elbes, M., Alzubi, S., Kanan, T., Al-Fuqaha, A., and Hawashin, B. (2019). A Survey on Particle Swarm Optimization with Emphasis on Engineering and Network Applications. *Evol. Intel.* 12 (2), 113–129. doi:10.1007/s12065-019-00210-z

Haghighat, M. B. A., Aghagolzadeh, A., and Seyedarabi, H. (2011). Multi-focus Image Fusion for Visual Sensor Networks in DCT Domain. *Comput. Electr. Eng.* 37 (5), 789–797. doi:10.1016/j.compeleceng.2011.04.016

Huang, L., Chen, C., Yun, J., Sun, Y., Tian, J., Hao, Z., et al. (2022). Multi-scale Feature Fusion Convolutional Neural Network for Indoor Small Target Detection. *Front. Neurorobotics* 85. 881021, doi:10.3389/fnbot.2022.881021

Huang, L., Fu, Q., He, M., Jiang, D., and Hao, Z. (2021). Detection Algorithm of Safety Helmet Wearing Based on Deep Learning. *Concurrency Comput. Pract. Exp.* 33 (13), e6234. doi:10.1002/cpe.6234

Huang, W., and Jing, Z. (2007). Evaluation of Focus Measures in Multi-Focus Image Fusion. *Pattern Recognit. Lett.* 28 (4), 493–500. doi:10.1016/j.patrec.2006.09.005

Jiang, D., Li, G., Sun, Y., Hu, J., Yun, J., and Liu, Y. (2021a). Manipulator Grabbing Position Detection with Information Fusion of Color Image and Depth Image Using Deep Learning. *J. Ambient. Intell. Hum. Comput.* 12 (12), 10809–10822. doi:10.1007/s12652-020-02843-w

Jiang, D., Li, G., Tan, C., Huang, L., Sun, Y., and Kong, J. (2021b). Semantic Segmentation for Multiscale Target Based on Object Recognition Using the Improved Faster-RCNN Model. *Future Gener. Comput. Syst.* 123, 94–104. doi:10.1016/j.future.2021.04.019

Jiang, D., Zheng, Z., Li, G., Sun, Y., Kong, J., Jiang, G., et al. (2019). Gesture Recognition Based on Binocular Vision. *Clust. Comput.* 22 (6), 13261–13271. doi:10.1007/s10586-018-1844-5

Lewis, J. J., O'Callaghan, R. J., Nikolov, S. G., Bull, D. R., and Canagarajah, N. (2007). Pixel- and Region-Based Image Fusion with Complex Wavelets. *Inf. fusion* 8 (2), 119–130. doi:10.1016/j.inffus.2005.09.006

Li, L., Zhang, S., Yu, X., and Zhang, L. (2018). PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching. *IEEE Trans. Circuits Syst. Video Technol.* 28 (3), 679–692. doi:10.1109/TCSVT.2016.2628782

Liu, Y., Chen, X., Peng, H., and Wang, Z. (2017). Multi-focus Image Fusion with a Deep Convolutional Neural Network. *Inf. Fusion* 36, 191–207. doi:10.1016/j.inffus.2016.12.001

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., and Wang, X. (2018a). Deep Learning for Pixel-Level Image Fusion: Recent Advances and Future Prospects. *Inf. Fusion* 42, 158–173. doi:10.1016/j.inffus.2017.10.007

Liu, Z., Blasch, E., Bhatnagar, G., John, V., Wu, W., and Blum, R. S. (2018b). Fusing Synergistic Information from Multi-Sensor Images: an Overview from Implementation to Performance Assessment. *Inf. Fusion* 42, 127–145. doi:10.1016/j.inffus.2017.10.010

Liu, Z., Tsukada, K., Hanasaki, K., Ho, Y. K., and Dai, Y. P. (2001). Image Fusion by Using Steerable Pyramid. *Pattern Recognit. Lett.* 22 (9), 929–939. doi:10.1016/s0167-8655(01)00047-2

Ma, B., Zhu, Y., Yin, X., Ban, X., Huang, H., and Mukeshimana, M. (2021). Sesf-fuse: An Unsupervised Deep Model for Multi-Focus Image Fusion. *Neural Comput. Applic* 33 (11), 5793–5804. doi:10.1007/s00521-020-05358-9

Ma, K., Kai Zeng, K., and Zhou Wang, Z. (2015). Perceptual Quality Assessment for Multi-Exposure Image Fusion. *IEEE Trans. Image Process.* 24 (11), 3345–3356. doi:10.1109/tip.2015.2442920

Moonon, A.-U., and Hu, J. (2015). Multi-focus Image Fusion Based on NSCT and NSST. *Sens. Imaging* 16 (1), 1–16. doi:10.1007/s11220-015-0106-3

Petrović, V. (2007). Subjective Tests for Image Fusion Evaluation and Objective Metric Validation. *Inf. Fusion* 8 (2), 208–216. doi:10.1016/j.inffus.2005.10.002

Piella, G., and Heijmans, H. (2003).A New Quality Metric for Image Fusion, Proceedings 2003 international conference on image processing (Cat. No. 03CH37429). IEEE. Barcelona Spain, 14-17 Sep. 2003. doi:10.1109/ICIP.2003.1247209

Qi, Y., Zhou, S., Zhang, Z., Luo, S., Lin, X., Wang, L., et al. (2021). Deep Unsupervised Learning Based on Color Un-referenced Loss Functions for Multi-Exposure Image Fusion. *Inf. Fusion* 66, 18–39. doi:10.1016/j.inffus.2020.08.012

Ram Prabhakar, K., Sai Srikar, V., and Venkatesh Babu, R. (2017). Deepfuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. Proceedings of the IEEE international conference on computer vision, 4714–4722. Venice Italy, 22-29 Oct. 2017. doi:10.1109/ICCV.2017.505

Saha, A., Bhatnagar, G., and Wu, Q. M. J. (2013). Mutual Spectral Residual Approach for Multifocus Image Fusion. *Digit. Signal Process.* 23 (4), 1121–1135. doi:10.1016/j.dsp.2013.03.001

Song, H., Liu, Q., Wang, G., Hang, R., and Huang, B. (2018). Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11 (3), 821–829. doi:10.1109/jstars.2018.2797894

Sun, Y., Zhao, Z., Jiang, D., Tong, X., Tao, B., Jiang, G., et al. (2022). Low-illumination Image Enhancement Algorithm Based on Improved Multi-Scale Retinex and ABC Algorithm Optimization. *Front. Bioeng. Biotechnol.*, 10, 396. doi:10.3389/fbioe.2022.865820

Tao Wan, T., Canagarajah, N., and Achim, A. (2009). Segmentation-driven Image Fusion Based on Alpha-Stable Modeling of Wavelet Coefficients. *IEEE Trans. Multimed.* 11 (4), 624–633. doi:10.1109/tmm.2009.2017640

Tessens, L., Ledda, A., Pizurica, A., and Philips, W. (2007).Extending the Depth of Field in Microscopy through Curvelet-Based Frequency-Adaptive Image Fusion, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. IEEE. I-861-I-864. Honolulu, HI, USA, 15-20 Apr. 2007. doi:10.1109/icassp.2007.366044

Tian, J., and Chen, L. (2012). Adaptive Multi-Focus Image Fusion Using a Wavelet-Based Statistical Sharpness Measure. *Signal Process.* 92 (9), 2137–2146. doi:10.1016/j.sigpro.2012.01.027

Wang, K., Dou, J., Kemao, Q., Di, J., and Zhao, J. (2019a). Y-net: a One-To-Two Deep Learning Framework for Digital Holographic Reconstruction. *Opt. Lett.* 44 (19), 4765–4768. doi:10.1364/ol.44.004765

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 13 (4), 600–612. doi:10.1109/tip.2003.819861

Xu, S., Chen, Z., Cao, W., Zhang, F., and Tao, B. (2021). Retinal Vessel Segmentation Algorithm Based on Residual Convolution Neural Network. *Front. Bioeng. Biotechnol.* 9, 786425. doi:10.3389/fbioe.2021.786425

Xue, Z., and Blum, R. S. (2003).Concealed Weapon Detection Using Color Image Fusion, Proceedings of the 6th Internation Conference On Information fusion. IEEE, 622–627. Cairns, QLD, Australia, 08-11 July 2003. doi:10.1109/ICIF.2003.177504

Xydeas, C. S., and Petrovic´, V. (2000). Objective Image Fusion Performance Measure. *Electron. Lett.* 36 (4), 308–309. doi:10.1049/el:20000267

Yan, H., Yu, X., Zhang, Y., Zhang, S., Zhao, X., and Zhang, L. (2017). Single Image Depth Estimation with Normal Guided Scale Invariant Deep Convolutional Fields. *IEEE Trans. Circuits Syst. Video Technol.* 29 (1), 80–92. doi:10.1109/TCSVT.2017.2772892

Zhang, Q., and Levine, M. D. (2016). Robust Multi-Focus Image Fusion Using Multi-Task Sparse Representation and Spatial Context. *IEEE Trans. Image Process.* 25 (5), 2045–2058. doi:10.1109/tip.2016.2524212

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L. (2020). IFCNN: A General Image Fusion Framework Based on Convolutional Neural Network. *Inf. Fusion* 54, 99–118. doi:10.1016/j.inffus.2019.07.011

Zhang, Y., Zhang, L., Bai, X., and Zhang, L. (2017). Infrared and Visual Image Fusion through Infrared Feature Extraction and Visual Information Preservation. *Infrared Phys. Technol.* 83 (1), 227–237. doi:10.1016/j.infrared.2017.05.007

Zheng, Y., Essock, E. A., Hansen, B. C., and Haun, A. M. (2007). A New Metric Based on Extended Spatial Frequency and its Application to DWT Based Fusion Algorithms. *Inf. Fusion* 8 (2), 177–192. doi:10.1016/j.inffus.2005.04.003

Zhou, Z., Li, S., and Wang, B. (2014). Multi-scale Weighted Gradient-Based Fusion for Multi-Focus Images. *Inf. Fusion* 20, 60–72. doi:10.1016/j.inffus.2013.11.005

Zhou, Z., Wang, B., Li, S., and Dong, M. (2016). Perceptual Fusion of Infrared and Visible Images through a Hybrid Multi-Scale Decomposition with Gaussian and Bilateral Filters. *Inf. Fusion* 30 (c), 15–26. doi:10.1016/j.inffus.2015.11.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.