Multimedia Tools and Applications (2022) 81:9277–9287
https://doi.org/10.1007/s11042-021-11549-w

1212: DEEP LEARNING TECHNIQUES FOR INFRARED IMAGE/VIDEO UNDERSTANDING

An infrared and visible image fusion algorithm based on ResNet-152

Liming Zhang[1,2] · Heng Li[1,2] · Rui Zhu[1,2] · Ping Du[1,2]

Received: 24 September 2020 / Revised: 29 July 2021 / Accepted: 9 September 2021 /
Published online: 3 January 2022

Abstract
The fusion of infrared and visible images can obtain a combined image with hidden objective and rich visible details. To improve the details of the fusion image from the infrared and visible images by reducing artifacts and noise, an infrared and visible image fusion algorithm based on ResNet-152 is proposed. First, the source images are decomposed into the low-frequency part and the high-frequency part. The low-frequency part is processed by the average weighting strategy. Second, the multi-layer features are extracted from high-frequency part by using the ResNet-152 network. Regularization L1, convolution operation, bilinear interpolation upsampling and maximum selection strategy on the feature layers to obtain the maximum weight layer. Multiplying the maximum weight layer and the high-frequency as new high-frequency. Finally, the fusion image is reconstructed by the low-frequency and the high-frequency. Experiments show that the proposed method can obtain more details from the image texture by retaining the significant features of the images. In addition, this method can effectively reduce artifacts and noise. The consistency in the objective evaluation and visual observation performs superior to the comparative algorithms.

Keywords Image processing · Image fusion · ResNet-152 · Infrared image · Visible image

1 Introduction

Image fusion is an image enhancement technology whose purpose is to combine the information captured by different types of sensors to generate a single image with more information and clearer details. The fusion image will be more informative for image processing computer vision and other research fields. The interactance, reflectance, and transmission

✉ Liming Zhang
zlm@lzjtu.edu.cn; zhanglm8@gmail.com

[1] Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China

[2] National-Local Joint Engineering Research Center of Technologies and Applications for National Geographic State Monitoring, Lanzhou 730070, China

modes of visible and infrared are very different. Infrared images are imaged by capturing thermal radiation, whereas the visible image is transformed from the optical correlator. The resolution of the visible image is higher than the infrared image but are easily disturbed by weak illumination, fog, and other harsh weather conditions. Although infrared images do not have these defects, but the detail information are not clear, the contrast are low and the texture are poorer [9, 16, 23]. The image after fusion has more information than any of the images before fusion [7, 12]. Therefore, the fusion of those two types images can achieve information complementation [17]. At present, fusion image has important applications in remote sensing, military reconnaissance, security monitoring, medical health, industrial production and other fields [24]. There are several types of image fusion methods for infrared and visible image, such as multi-scale transformation, sparse representation, and neural network.

The multi-scale transform method is most widely used in image fusion. Huang et al. [4] proposed an infrared and visible image fusion method based on curve transformation and visual attention mechanism. The model can improve the signal-to-noise ratio of the fused image and work well for blur objects. Zhu et al. [27] proposed an improved multi-scale high-hat transform model infrared image fusion method which can highlight the target of infrared image and better preserve the details in the visible image.

The image fusion method based on sparse representation has better fusion effect than the multi-scale transformation method, and has become a very active research direction in the current image fusion field. The key of this method lies in construct a dictionary and coding an image via sparse representation. Yin et al. [25] proposed a multi-scale-dictionary learning method combining wavelet transform with dictionary learning. This method can make full use of the advantages of multi-scale representation and dictionary learning. Kim et al. [5] proposed a dictionary learning method based on image block clustering and principal component analysis which not only eliminates the redundancy of learning dictionary but also guarantees the result of the fused image. The shortcoming of dictionary learning is the computational complexity, these algorithms always undergo several iterations.

Deep learning has been widely used in image fusion in recent years [9, 11, 18]. Deep learning models can extract deep features of image [21]. Approving results can be achieved by fusing these features and the source images. Prabhakar et al. [18] proposed a multi-exposure image fusion method based on convolutional neural network (CNN). The network structure is a weight-sharing twin network. After the source image is input to the encoder, two feature mapping sequences are obtained. The fusion feature map is obtained by the addition fusion strategy. Then, the fusion feature map is reconstructed by the three-layer convolution layer of the decoder. Liu et al. [11] proposed an image fusion method based on CNN. The image block and its blurred block are input to the network for training the model, so the model has classification ability. The network output is a classified average score table. Overlapping blocks, binarization, and two consistency checking strategies are used to determine the mapping, and finally the mapping is determined as the source map weight to reconstruct the image [1]. Although the method achieves better performance, there are still two disadvantages: (1) This method is currently only suitable for multi-focus image fusion, and its use is limited; (2) It only uses the calculation result of the last layer of the neural network whereas the middle layer information is not fully utilized which is useful.

In this paper, an infrared and visible image fusion method based on ResNet-152 [3] is proposed. Firstly, the source images are processed by mean filtering, then two-scale decomposition [8] was used to obtain a low-frequency part containing large-scale features and a high-frequency part containing texture features. Secondly, a new low-frequency part is obtained from the low-frequency part by using the average weighting strategy and the multi-layer features is extracted from high-frequency part by using the ResNet-152 network. L1 regularization, convolution operation and bilinear interpolation upsampling on each layer to obtain the weight layer. Then, by choosing the maximum of multiple weight layers to obtain the new weight layer. The new weight layer multiplies the high-frequency part to get a new high-frequency. Finally, the image is reconstructed with the new low-frequency and the high-frequency.

The rest parts of this paper are organized as follows: In Sect. 2, the structure element of residual network and the ResNet-152 are introduced, the fusion framework of the proposed algorithm and details are given, respectively. In Sect. 3, 20 groups of TNO's Data fusion experiments are performed with five comparative fusion methods to verify the effectiveness of the proposed method, and the fusion results and analysis are provided. Finally, conclusions are drawn in Sect. 4.

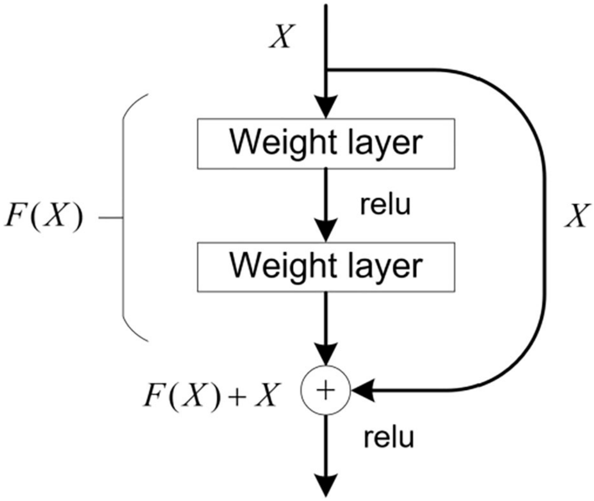
2 The proposed fusion method

2.1 Advantages of ResNet

The number of hidden layers in the neural network is very important for the extraction of image features. However, the gradient vanishing and the unexpected degradation emerged with the network depth increasing [3]. This problem can be successfully solved by the residual neural network (ResNet) that with shortcut connections [3] and speed up the training. It became a backbone network in many computer vision tasks for feature extraction. Figure 1 shows a unit of ResNet.

In Fig. 1, In order to solve the problem of vanishing/exploding gradients, a skip/shortcut connection is added, and the input is added to the output after several weight layers. X represents the input of the residual network, $F(X)F(X)$ is the function which mapping X and two weight layers, Relu is the activation function and the output of the residual block is $F(X) + X$. The weight layer is actually learning a kind of residual mapping. Even if the

Fig. 1 The residual block



gradient of the weight layer disappears, we still have X that can be transferred back to the earlier layer. Thus, every layer information of the network can be effectively used. Some trained networks like ResNet-152 has been widely applied in image processing. The depth of ResNet-152 is 8 times of VGG-19 [19], but its complexity is lower and the ability to extract features is stronger [14]. Therefore, ResNet-152 can be used as the basic network for feature extraction in image fusion.

The method we proposed is featured in Fig. 2. The image is decomposed into a base layer and a detail layer, and different layers use different fusion methods. For the mean filtering is a very common method of smoothing images, it can eliminate sharp noise of image. So, the mean filter is employed firstly in this method. By using the two-scale decomposition [8] on I_k to obtain the low-frequency part I_k^b and the high-frequency image I_k^d, where I_k, $k \in \{1{,}2\}$ are the input images, Z is the mean filter of size 31×31, it was chosen to achieve a balance of effect and efficiency. I_k^b is obtained from I_k and Z through formula (1), I_k^d is obtained by the formula (2). The low-frequency part is fused with the average weight method to obtain F_b, and the high-frequency part is processed by ResNet-152 to obtain the maximum weight layer, then, the high-frequency of fusion image F_d is obtained from the former and the high-frequency part of the input image. Finally, the fusion image is reconstructed from F_b and F_d.

$$I_k^b = I_k * Z \tag{1}$$

$$I_k^d = I_k - I_k^b \tag{2}$$

2.2 The fusion rule for low-frequency

The low-frequency component of the image usually represents the background information of the image. The fusion of the low-frequency component often uses the averaging method, which can effectively suppress the influence of noise, but it reduces the contrast of the image to a certain extent, causing the loss of information. Through experimental comparison, the low-frequency weight of infrared images is higher than that of visible light images, and the results are relatively ideal. As shown in formula (3), $I_1^b(x,y)$ and $I_2^b(x,y)$ are the

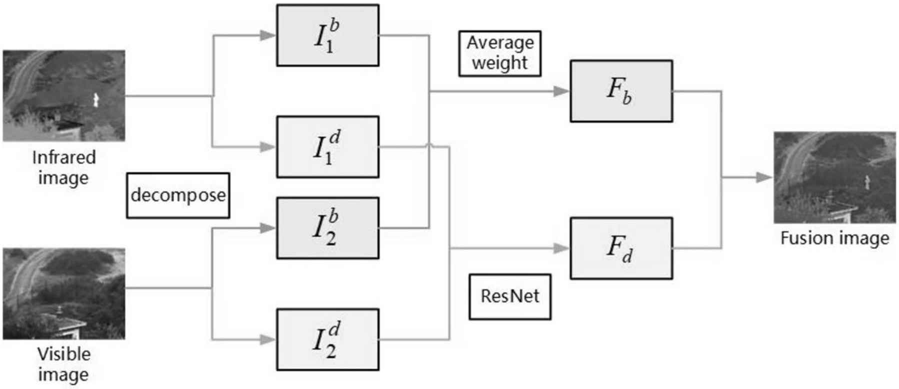


Fig. 2 The framework of proposed method

values in the low-frequency (x, y) of the input images, and $F_b(x, y)$ is the value in the low-frequency part (x, y) of the fused image, λ_1 and λ_2.are the weight of I_1^b and I_2^b. In order to preserve the large-scale features in the low-frequency part as much as possible, the values in the experiment are 0.6 and 0.4 respectively.

$$F_b(x, y) = \lambda_1 I_1^b(x, y) + \lambda_2 I_2^b(x, y) \tag{3}$$

2.3 The fusion rule for high-frequency

The high-frequency components of the image carrys the sharply changing parts of the image, that is, the edges (contours) or noise and details of the image. The usual fusion method is to compare the absolute values of the high-frequency coefficients at the corresponding positions of the two images, and take the larger absolute value as the new coefficient. This method can retain more image information. The procedure of the fusion process for high-frequency is given in Fig. 3. For the high-frequency, by using the ResNet-152 to obtain the feature layer from I_1^d and I_2^d. Then, the maximum weighting layer is obtained

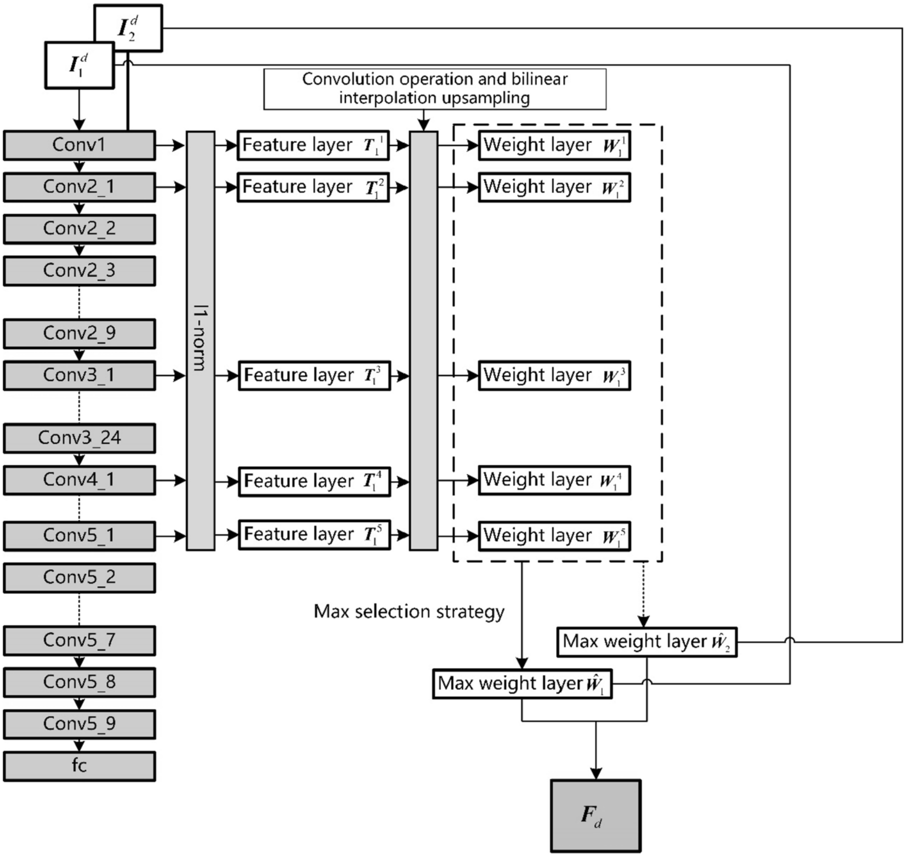


Fig. 3 The procedure of high-frequency fusion

from the feature layer through the maximum selection strategy. Finally, the maximum weight layer is used as the I_1^d and I_2^d weight coefficients to obtain the high-frequency of the fused image F_d.

3 The specific steps of high-frequency part fusion are as follows:

Step 1: Extracting features. The pre-trained models can be used for feature extraction. So, the ResNet-152 is employed to extract features, conv1, conv2_1, conv3_1, conv4_1, conv5_1 of ResNet-152 are chosen for the feature extraction. In formula (4), $H_i(\cdot)$ is the function for the feature extraction in the ResNet-152. The extracted feature layer is $L_k^{i,m}$, $i \in \{1,2,3,4,5\}$.

$$L_k^{i,m} = H_i\left(I_k^d\right) \tag{4}$$

Step 2: L1 regularization. $L_k^{i,m}$ represents the i th extracted feature from k th feature layer through the $H_i(\cdot)$ in the high-frequency part, where $m = 64 \times 2^{i-1}$, m represents the number of channels at the i th feature layer. So, $L_k^{i,m}(x, y)$ represents a m dimension vector. As shown in Eq. (5), T_k^i is the result of L1 regularization on $L_k^{i,m}(x, y)$. Specifically, T_k^i is the result of the i th feature layer of the k th high-frequency part by L1 regularization.

$$T_k^i = \|L_k^{i,m}(x, y)\|_1 \tag{5}$$

Step 3: Convolution operation. For fused images are rich in texture information, the convolution kernel A is adopted, show as the formula (6), the step size is 1. The convolution operation process is shown in Fig. 4. Then, convolution operation on T_k^i to get $\widehat{T}_k^i$. The weight layer $W_k^i(x, y)$ is obtained from $\widehat{T}_k^i$ by using formula (7) where $n = 2$, $\widehat{T}_k^i(x, y)$ are the

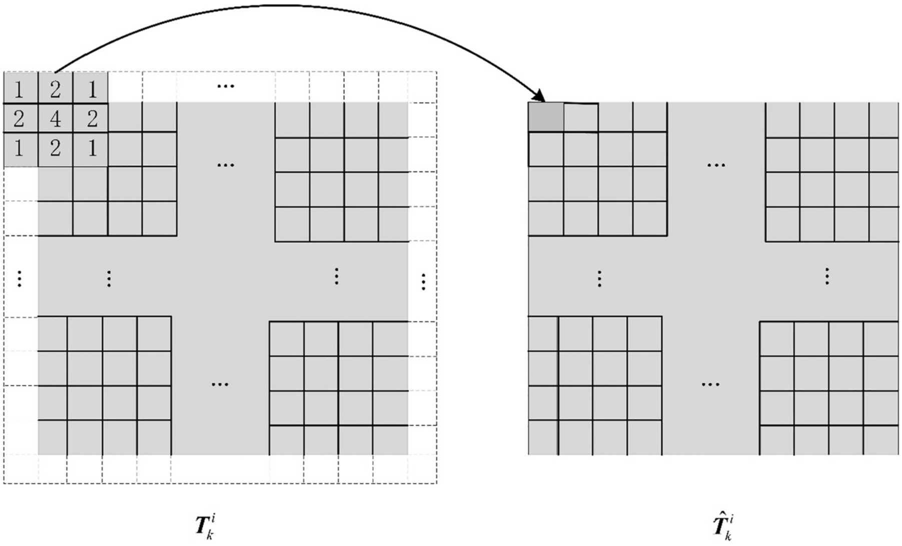


Fig. 4 The process of the convolution operation

values of the i th feature layer at (x, y) from the k th high-frequency part by convolving, and $W_k^i(x, y)$ represents the i th weight value at (x, y) from the k th high-frequency part.

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} \tag{6}$$

$$W_k^i(x, y) = \frac{\widehat{T}_k^i(x, y)}{\sum_{k=1}^{n} \widehat{T}_k^i(x, y)} \tag{7}$$

Step 4: Bilinear interpolation upsampling. W_k^i is obtained by ResNet-152, the width and height of W_k^i are $\left(m_i, n_i\right)$, the width and height of I_k^d are (M, N), the relationship between them shows as Eq. (8). Bilinear interpolation upsampling on W_k^i to obtain $\widehat{W}_k^i$, $i \in \{1,2,3,4,5\}$, also, it can make $\left(m_i, n_i\right)$ equal to (M, N).

$$\left(m_i, n_i\right) = (M, N) \times \frac{1}{2^i} \tag{8}$$

Step 5: The maximum selection strategy. $\widehat{W}_k^i(x, y)$ is the value of the i th weights layer at (x, y) which bilinear interpolation on the k th high-frequency partial. So, $\widehat{W}_k^{1:5}(x, y)$ is a 5-dimensional vector, such as Eq. (9). $\widehat{W}_k(x, y)$ is obtained by using the maximum selection strategy on $\widehat{W}_k^{1:5}(x, y)$. $\widehat{W}_k$ is the maximum weighting layer of the k th high-frequency part. The final high-frequency F_d is obtained by Eq. (10).

$$\widehat{W}_k(x, y) = max\left(\widehat{W}_k^{1:5}(x, y)\right) \tag{9}$$

$$F_d = \sum_{k=1}^{n} \widehat{W}_k I_k^d \tag{10}$$

3.1 Image reconstruct

After obtaining the low-frequency part F_b and the high-frequency part F_d, then, the fusion image can be reconstructed by (11). where $F_b(x, y)$ represents the value of the low-frequency part at (x, y) and $F_d(x, y)$ represents the high-frequency of the fusion image at (x, y). $F(x, y)$ represents the pixel value of the fusion image at (x, y).

$$F(x, y) = F_b(x, y) + F_d(x, y) \tag{11}$$

4 Fusion experiments and quality analysis

4.1 3.1 Experiment dataset and settings

The TNO dataset [20] is used in ours experiments. 20 pairs of visible and infrared images are selected for the experiments. The image resolution is 360×270, 632×496, and 620×450, respectively. In order to quantitatively evaluate the performance of different fusion methods, we use five commonly adopted objective fusion metrics. They are cross-bilateral filtering fusion method (CBF) [6], joint sparse representation model (JSR) [26], joint detection based sparse representation model (JSRSD) [13], weighted least-squares optimization method (WLS) [15] and convolution sparse representation model (ConvSR) [10]. All parameters of the above the five algorithms are consistent with the literature.

4.2 3.2 Subjective visual analysis

Our method is compared with the above five methods and the results are analyzed. Five groups of images numbered as (a, b, c, d, e) are selected for descriptions respectively. Figure 5 shows that the CBF has more noise and vignetting, the salient features are not clear.

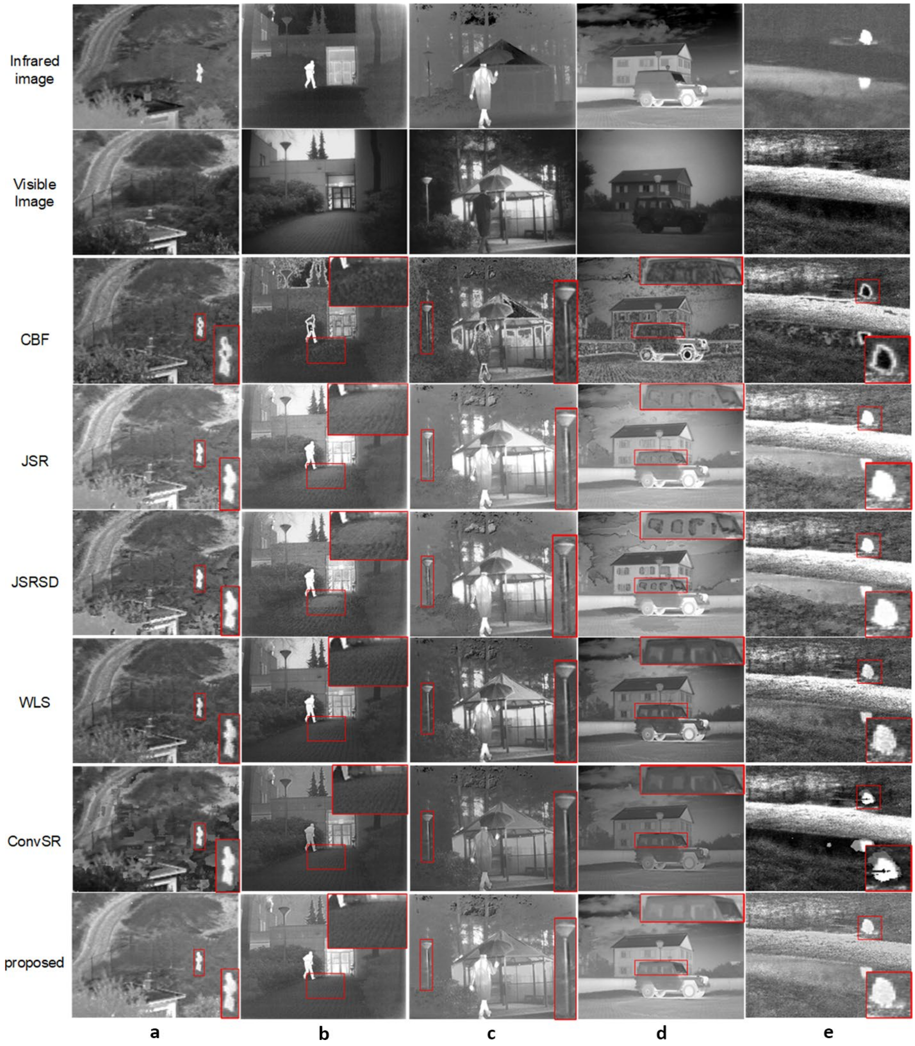


Fig. 5 Fused images obtained by different methods

In the JSR, the details of other regions are blurred and the contrast is not high, also the salient features are not clear. After the fusion of JSRSD and WLS, the image retains more features of the infrared image, the brightness is too high, and the area is excessively unnatural. In ConvSR, it has more artifacts around the salient features and the blockiness is obvious, and the perception is not good. Compared with these five methods, the proposed method is clearer, the contrast is higher, and the vignette and block effect are not obvious, which is more suitable for human visual observation.

4.3 3.3 Objective evaluation metrics

In order to quantitatively compare the proposed method and the comparison methods, Five indicators: FMI_{pixel}[2, 2]FIM_{dct}, FIM_w[2, 17]$SSIM$ and N_{abf}[22] is used for evaluation. FMI_{pixel}, FIM_{dct} and FIM_w can be calculated for mutual information of features from pixel features, discrete cosine features and wavelet features, the larger the value, the higher the information correlation between the input image and the fusion image, and the less information loss in the process. $SSIM$ indicates how similar the input image is to the fusion image. N_{abf} represents the noise or artifacts produced in the fusion image. The smaller the value, the fewer artifacts, and noise the fused image contains. In our experiments, the five groups of images in Fig. 5 were selected to quantitatively evaluate our method and the comparison methods. The results show in Table 1.

The best results of the five indicators are marked in bold. It can be seen from the results that our method has a wonderful performance in the almost five indicators. From Table 1, Most of indicators of the proposed method are better than the comparison methods except FMI_w and N_{abf} of ConvSR in the image e. This shows that compared with the other five contrast methods, the fusion image by our method largely retain the texture detail features of the source image, and reduce the artifacts and noise. It is clearer and more natural, which is consistent with the subjective evaluation.

5 Conclusions

This paper presents a novel image fusion algorithm based on ResNet-152. First, the source images are decomposed into a low-frequency part and a high-frequency part. The low-frequency part is fused by using the average weight. The high-frequency part is processed by the ResNet-152 network to obtain the maximum weight layer, and then the high-frequency part and the maximum weight layer are multiplied to obtain the fused high-frequency part. Finally, the image is reconstructed from the low-frequency part and the high-frequency part. The experimental results demonstrate that our method not only preserves the texture features of the source image well but also greatly reduces the artifacts and noise of the fusion image. In conclusion, our method performs superior in both subjective evaluation and objective evaluation. At last, how to use different networks (VGG19, ResNet50 and ResNet101) and different norms(l1-norm, and l2-norm]) to further improve the fusion effect can be further researched.

Table 1 The objective evaluation values of image fusion

Images	Metrics	CBF	WLS	JSR	JSRSD	ConvSR	Proposed
a	FMI_{pixel}	0.8701	0.8847	0.8528	0.8339	0.8336	**0.8856**
	FMI_{dct}	0.2450	0.2745	0.1538	0.1315	0.1422	**0.3378**
	FMI_{w}	0.2984	0.3351	0.2113	0.1812	0.3205	**0.3828**
	N_{abf}	0.2317	0.1621	0.2332	0.3285	0.1224	**0.0250**
	$SSIM$	1.2476	1.4442	1.2146	1.0792	1.1968	**1.5324**
b	FMI_{pixel}	0.8968	0.9305	0.9189	0.8932	0.9160	**0.9345**
	FMI_{dct}	0.2229	0.2890	0.1686	0.1443	0.1388	**0.3563**
	FMI_{w}	0.2920	0.3495	0.2326	0.2036	0.3813	**0.3906**
	N_{abf}	0.2554	0.2091	0.2430	0.3419	0.0572	**0.0374**
	$SSIM$	1.2994	1.4618	1.2562	1.1330	1.3608	**1.5454**
c	FMI_{pixel}	0.8645	0.8918	0.8792	0.8632	0.8477	**0.8927**
	FMI_{dct}	0.2299	0.2981	0.1699	0.1454	0.1342	**0.3477**
	FMI_{w}	0.2633	0.3464	0.2205	0.1956	0.3424	**0.3836**
	N_{abf}	0.3607	0.2156	0.1769	0.2417	0.0868	**0.0387**
	$SSIM$	1.0740	1.3834	1.2560	1.1874	1.2084	**1.46□20**
d	FMI_{pixel}	0.8386	0.8977	0.8885	0.8653	0.8380	**0.9128**
	FMI_{dct}	0.1835	0.3236	0.1591	0.1325	0.1065	**0.3649**
	FMI_{w}	0.2407	0.3773	0.2266	0.2078	0.2992	**0.4007**
	N_{abf}	0.5278	0.2821	0.2109	0.2801	0.1145	**0.0376**
	$SSIM$	0.9150	1.3538	1.2506	1.1616	1.1500	**1.4478**
e	FMI_{pixel}	0.8569	0.8471	0.8481	0.8434	0.8629	**0.8634**
	FMI_{dct}	0.3781	0.3908	0.2360	0.2099	0.3061	**0.4429**
	FMI_{w}	0.4650	0.4428	0.2840	0.2565	**0.5144**	0.4476
	N_{abf}	0.1523	0.2820	0.1873	0.2198	**0.0348**	0.0483
	$SSIM$	1.0472	1.0916	0.9764	0.9374	1.1292	**1.1808**

Acknowledgements This work is funded by the Natural Science Foundation Committee, China (No. 41761080, and No. 41930101) and Industrial Support and Guidance Project of Gansu Colleges and Universities, No. 2019C-04.

Author contributions LZ conceived, designed, and also wrote the manuscript; HL performed the experiments; RZ supervised the study; PD offered helpful suggestions and reviewed the manuscript. RZ and PD analyzed and evaluated the results.

Data availability The code and the test vector map data associated with this paper can be found at https://github.com/diylife/imagefusion_deeplearning.git.

References

1. Du P et al (2020) Advances of four machine learning methods for spatial data handling: a review. J Geovis Spat Anal 4(1):13
2. Haghighat M, Razian MA (2014) Fast-FMI: non-reference image fusion metric. IEEE
3. He K, et al (2016) Deep residual learning for image recognition
4. Huang Y, et al (2017) Infrared and visible image fusion with the target marked based on multi-resolution visual attention mechanisms. In: Selected Papers of the Chinese Society for Optical Engineering Conferences held October and November 2016. International Society for Optics and Photonics

5. Kim M, Han DK, Ko H (2016) Joint patch clustering-based dictionary learning for multimodal image fusion. Inf fusion 27:198–214
6. Kumar BS (2015) Image fusion based on pixel significance using cross bilateral filter. SIViP 9(5):1193–1204
7. Li H, Wu X-J (2018) Densefuse: a fusion approach to infrared and visible images. IEEE Trans Image Process 28(5):2614–2623
8. Li S, Kang X, Hu J (2013) Image fusion with guided filtering. IEEE Trans Image Process 22(7):2864–2875
9. Li H, Wu X-J, Durrani TS (2019) Infrared and visible image fusion with ResNet and zero-phase component analysis. Infrared Phys Technol 102:103039
10. Liu Y et al (2016) Image fusion with convolutional sparse representation. IEEE Signal Process Lett 23(12):1882–1886
11. Liu Y et al (2017) Multi-focus image fusion with a deep convolutional neural network. Inf Fusion 36:191–207
12. Liu SP, Fang Y (2007) Infrared image fusion algorithm based on contourlet transform and improved pulse coupled neural network. J Infrared Millim Waves 26(3):217–221
13. Liu C, Qi Y, Ding W (2017) Infrared and visible image fusion method based on saliency detection in sparse domain. Infrared Phys Technol 83:94–102
14. Liu S, Tian G, Xu Y (2019) A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. Neurocomputing 338:191–206
15. Ma J et al (2017) Infrared and visible image fusion based on visual saliency map and weighted least square optimization. Infrared Phys Technol 82:8–17
16. Ma J et al (2020) Infrared and visible image fusion via detail preserving adversarial learning. Inf Fusion 54:85–98
17. Ma J, Ma Y, Li C (2019) Infrared and visible image fusion methods and applications: a survey. Inf Fusion 45:153–178
18. Prabhakar, KR, Srikar VS, Babu RV (2017) DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image Pairs
19. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. http://arxiv.org/abs/1409.1556
20. Toet A (2014) TNO Image fusion dataset. Figshare. data
21. Wang M et al (2019) Scene classification of high-resolution remotely sensed image based on ResNet. J Geovis Spat Anal 3(2):16
22. Wang Z, Bovik AC (2002) A universal image quality index. IEEE Signal Process Lett 9(3):81–84
23. Wu Y, Wang Z (2017) Infrared and visible image fusion based on target extraction and guided filtering enhancement. Acta Opt Sin 37(8):0810001
24. Xu L, Cui GM, Zheng CP (2017) Fusion method of visible and infrared images based on multi-scale decomposition and saliency region. Laser Optoelectron Prog 54(11):111–120
25. Yin H (2015) Sparse representation with learned multiscale dictionary for image fusion. Neurocomputing 148:600–610
26. Zhang Q et al (2013) Dictionary learning method for joint sparse representation-based image fusion. Opt Eng 52(5):057006
27. Zhu P, Ma X, Huang Z (2017) Fusion of infrared-visible images using improved multi-scale top-hat transform and suitable fusion rules. Infrared Phys Technol 81:282–295