WaveFuse: A Unified Unsupervised Framework for Image Fusion with Discrete Wavelet Transform

Shaolei Liu[1,2], Manning Wang[1,2(✉)], and Zhijian Song[1,2(✉)]

[1] Digital Medical Research Center, School of Basic Medical Science,
Fudan University, Shanghai, China
{slliu,mnwang,zjsong}@fudan.edu.cn
[2] Shanghai Key Laboratory of Medical Image Computing and Computer Assited
Intervention, Shanghai, China

Abstract. We propose an unsupervised image fusion architecture for multiple application scenarios based on the combination of multi-scale discrete wavelet transform through regional energy and deep learning. To our best knowledge, this is the first time that a conventional frequency method has been combined with deep learning for feature maps fusion. The useful information of feature maps can be utilized adequately through multi-scale discrete wavelet transform in our proposed method. Compared with other state-of-the-art fusion methods, the proposed algorithm exhibits better fusion performance in both subjective and objective evaluation. Moreover, it's worth mentioning that comparable fusion performance trained in COCO dataset can be obtained by training with a much smaller dataset with only hundreds of images chosen randomly from COCO. Hence, the training time is shortened substantially, leading to the improvement of the model's performance both in practicality and training efficiency.

Keywords: Multi-scene image fusion · Unsupervised learning · Discrete wavelet transform · Regional energy

1 Introduction

Image fusion is the technique of integrating complementary information from multiple images obtained by different sensors of the same scene, so as to improve the richness of the information contained in one image [3]. Image fusion can compensate for the limitation of single imaging sensors, and this technique has developed rapidly in recent years because of the wide availability of different kinds of imaging devices [3]. For example, in medical imaging applications, images of different modalities can be fused to achieve more reliable and precise medical diagnosis [23]. In military surveillance applications, image fusion integrates information from different electromagnetic spectrums (such as visible and infrared bands) to achieve night vision [10].

The extraction of feature maps and the selection of fusion rules are the two key factors determining the quality of the fused image [15], and most studies focus on proposing new methods based on these two factors. Before the overwhelming application of deep learning in image processing, many conventional approaches were used in feature extraction for image fusion, which can be divided into two categories: transform domain algorithms and spatial domain algorithms [10]. In transform domain algorithms, the source images are transformed to a specific transform domain, such as the frequency domain, where the feature maps are represented by the decomposition coefficients of the specific transform domain. In feature maps fusion, max-rule and averaging are commonly used for high and low frequency bands, respectively, and then the fused image is reconstructed by the inverse transform from the fused feature maps [1,6]. Unlike transform domain algorithms, spatial domain algorithms employ the original pixel of source images as feature maps and directly calculate the weighted average of the source images to obtain the final fused image without dedicated feature maps extraction, where the weights are selected according to image blocks [11] or gradient information [9]. Consequently, the conventional approaches can be regarded as designing some hand-crafted filters to process the source images, and it is difficult for them to adapt to images of different scenes or parts with different visual cues in one image.

Nowadays, deep learning has been the state-of-the-art solution in most tasks in the fields of image processing and computer vision. Recently, deep learning has also been used in image fusion and achieved higher quality than conventional methods. For example, CNN can be used to automatically extract useful features and can learn the direct mapping from source images to feature maps. In recent image fusion research based on deep learning [2,7,8,15,16,20,23], fusion using learned features through CNN achieved higher quality than conventional fusion approaches. According to the different fusion framework utilized, deep learning based methods can be divided into the following three categories: CNN based methods [8,14,15,29], encoder-decoder based methods [7,20,23,27] and generative adversarial network (GAN) based methods [18]. CNN based methods merely apply several convolutional layers to obtain the weight map for source images. Encoder-decoder based methods introduce encoder-decoder architecture to extract deep features, and the deep features are fused by weighted average or concatenation. Furthermore, GAN based methods leverage conditional GAN to generate the fused image, where the concatenated source images are directly input to the generator. In these studies, the feature maps obtained through deep learning are usually simply fused by weighted averaging, and we will show that this is not optimal. More importantly, the neural networks used in these studies [7,8,15,20,23] usually need to be trained on large image dataset, which is time consuming.

In this paper, we propose an image fusion algorithm by combining the deep learning based approaches with conventional transform domain based approaches. Concretely, we first train an encoder-decoder network and extract feature maps from the source images by the encoder. Inspired by the multi-scale

transform [10], discrete wavelet transform (DWT) is utilized to transform the feature maps into the wavelet domain, and adaptive fusion rules are used at low and high frequencies, thus making the beat use of the information of feature maps. Finally, inverse wavelet transform is used to reconstruct the fused feature map, which is decoded by the decoder to obtain the final fused image. Experiments show that with the additional processing of the feature maps by DWT, the quality of the fused image is remarkably improved. To the best of our knowledge, this is the first time to adopt conventional transform domain approaches to fuse the feature maps obtained from deep learning approaches.

The main contributions are summarized as follows:

(1) A generalized and effective unsupervised image fusion framework is proposed based on the combination of multi-scale discrete wavelet transform and deep learning.
(2) With multi-scale decomposition in DWT, the useful information of feature maps can be fully utilized. Moreover, a region-based fusion rule is adopted to capture more detail information. Extensive experiments demonstrate the superiority of our network over the state-of-the-art fusion methods.
(3) Our network can be trained in a smaller dataset with low computational cost to achieve comparable fusion performance compared with existing deep learning based methods trained on full COCO dataset. Our experiments show that the quality of the fused images and the training efficiency are improved sharply.

2   Proposed Method

2.1   Network Architecture

WaveFuse is a typical encoder-decoder architecture, consisting of three components: an encoder, a DWT-based fusion part and a decoder. As shown in Fig. 1(b), the inputs of the network are spatially aligned source images I_k, where $k = 1,2$ is used to index the images. Feature maps F_k are obtained by extracting features from the input source images I_k through the encoder. The feature maps F_k are first transformed into the wavelet domain, and an adaptive fusion rule is used to obtain the fused feature maps $F^{'}$. Finally, the fused feature maps are input into the decoder to obtain the final fused image I_F. The encoder is composed of three ConvBlocks, where two CNNs and a Relu layer are included. The kernel size of CNNs are all 3×3. After encoding, 48D feature maps are obtained for fusion. In the DWT-based fusion part, to take 1 layer wavelet decomposition and one dimension of the feature maps F_k for example, the feature maps are decomposed to different wavelet components C_k, including one low-frequency component C_{kL}, namely L_{1k} and three high-frequency components C_{kH}: horizontal component H_{1k}, vertical component V_{1k} and diagonal component D_{1k}, respectively. Different fusion rules are employed for different components to obtain the fused wavelet components F, where the low-frequency component L_2 is obtained from the fusion of L_{11} and L_{12}, and the high-frequency components
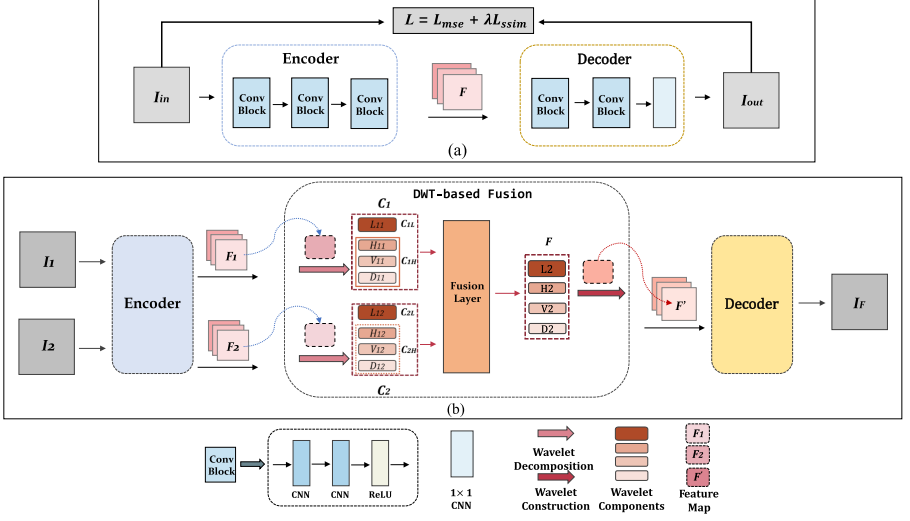
Fig. 1. (a) The framework of the training process. (b) Architecture of the proposed WaveFuse image fusion network. The feature maps learned by the encoder from the input images are processed by multi-scale discrete wavelet transform, and finally the fused feature maps are utilized to the fused image reconstruction by the decoder.

H_2, V_2 and D_2 are obtained from the fusion of H_{1k}, V_{1k} and D_{1k}, respectively. Finally, the fused low-frequency component and high-frequency components are integrated by wavelet reconstruction to obtain the final fused feature map F'. The decoder is mainly composed of two ConvBlocks and one 1×1 CNN, where the fused image is finally reconstructed.

2.2 Loss Function

The loss function L used to train the encoder and the decoder in WaveFuse is a weighted combination of pixel loss L_p and structural similarity loss L_{ssim} with a weight λ , where λ is assigned as 1000 according to [7]. The loss function L, pixel loss L_p and structural similarity loss L_{ssim} are defined as follows:

$$L = L_p + \lambda L_{ssim}, \tag{1}$$

$$Lp = ||I_{out} - I_{in}||_2, \tag{2}$$

$$L_{ssim} = 1 - SSIM(I_{out}, I_{in}), \tag{3}$$

where I_{in} and I_{out} represent the input image to the encoder and the output image of the decoder, respectively. The structural similarity (SSIM) is a widely used perceptual image quality metric, which combines the three components of luminance, structure and contrast to comprehensively measure image quality [25].

2.3   Training

We trained our network shown in Fig. 1(a) using COCO [12] containing 70,000 images, and all of them were resized to 256×256 and transformed to gray images. The batch size and epochs were set as 64 and 50, respectively. Learning rate was 1×10^{-4}. The proposed method was implemented on Pytorch 1.1.0 with Adam as the optimizer and a NVIDIA GTX 2080 Ti GPU for training. In our practical training process, we found that using comparatively small dataset, containing 300–700 images chosen randomly from COCO, still achieved a comparable fusion quality. The parameters for small dataset are as follows: learning rate was set as 1×10^{-4}, and the batch size and epochs were 16 and 100, respectively.

2.4   Fusion Rule

The selection of fusion rules largely determines the quality of fused images [15]. Existing image fusion algorithms based on deep learning usually calculate the sum of the feature maps directly, leaving the information of feature maps not fully mined.

In our method, two complementary fusion rules based on DWT are adopted for wavelet components C_k transformed by feature maps F_k, including adaptive rule based on regional energy [26] and *l1-Norm* rule [7], and the fused wavelet components are denoted as F_r and F_{l1}, respectively. In adaptive rule based on regional energy, different fusion rules are employed for different frequency components, that is, the low-frequency components C_{kL} adopts an adaptive weighted averaging algorithm based on regional energy, and for the high-frequency components C_{kH}, the one with larger variance between C_{1H} and C_{2H} will be selected as the fused high-frequency components. Due to the page limitation, the detailed description and futher equations can be found in [13]. Additionally, to preserve more structural information and make our fused image more natural, we apply *l1-Norm* rule [7] to our fusion part, where both low and high frequency components are fused by the same rule to obtain global and general fused wavelet components.

3   Experimental Results and Analysis

In this section, to validate the effectiveness and generalization of our WaveFuse, we first compare it with several state-of-the-art methods on four fusion tasks, including mult-exposure (ME), multi-modal medical (MED), multi-focus (MF) and infrared and visible (IV) image fusion. There are 20 pairs of images in each scenario, and all the images are from publicly available datasets [7,15,18,22]. For quantitative comparison, we use nine metrics to evaluate the fusion results. Then, we evaluate the fusion performance of the proposed method trained with small datasets. Finally, we also conduct the fine-tuning experiments on wavelet parameters for the further improvement of fusion performance.

3.1   Compared Methods and Quantitative Metrics

WaveFuse is compared against nine representative peer methods including discrete wavelet transform (DWT) [6], cross bilateral filter method (CBF) [5], convolutional sparse representation (ConvSR) [17], GAN-based fusion algorithm (FusionGAN) [18], DenseFuse [7] IFCNN [29], and U2Fusion [27]. All the nine comparative methods were implemented based on public available codes, where the parameters were set according to the original papers.

The commonly used evaluation methods can be classified into two categories: subjective evaluation and objective evaluation. Subjective evaluation is susceptible to human factors, such as eyesight, subjective preference and individual emotion. Furthermore, no prominent difference among the fusion results can be observed in most cases based on subjective evaluation. In contrast, objective evaluation is a relatively accurate and quantitative method on the basis of mathematical and statistical models. Therefore, in order to compare fairly and comprehensively with other fusion methods, we choose the following nine metrics:EN [21], cross entropy(CE), FMI_pixel [4], FMI_dct [4], FMI_w [4], Q^{NICE} [24]), $Q^{AB/F}$ [28], variance(VARI) and subjective similarity (MS-SSIM [19]). Each of them reflects different image quality aspects , and the larger the nine quality metrics are, the better the fusion results will be.

3.2   Comparison to Other Methods

Table 1. Quantitative comparison of WaveFuse with existing multi-scene image fusion methods. Red ones are the best results, and blue ones mark the second results. For all metrics, larger is better.

ME	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
U2Fusion	6.9525	4.7649	0.8529	0.4005	0.4411	0.8146	0.4319	41.5841	0.9349
IFCNN	6.5269	2.1552	0.8590	0.4580	0.4892	0.8134	0.4788	34.4859	0.9478
Wavefuse	6.9224	4.7309	0.8661	0.4941	0.5265	0.8240	0.5149	39.8423	0.9582
MED	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
U2Fusion	5.1257	7.5716	0.8430	0.2876	0.3471	0.8067	0.2879	57.8377	0.8980
IFCNN	5.0101	1.2406	0.8581	0.3611	0.4450	0.8075	0.3240	73.6683	0.9434
Wavefuse	5.2814	7.8243	0.8650	0.4052	0.3848	0.8094	0.3265	67.0572	0.9007
MF	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
U2Fusion	7.4552	0.3424	0.8706	0.4018	0.4653	0.8269	0.6586	55.5703	0.9574
IFCNN	7.2756	0.0558	0.8903	0.4687	0.5389	0.8356	0.7447	48.6216	0.9947
Wavefuse	7.3681	0.3483	0.8912	0.5041	0.5619	0.8368	0.7743	51.7256	0.9896
IV	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
U2Fusion	6.9092	1.1991	0.8795	0.3158	0.3546	0.8055	0.4775	37.1806	0.9196
IFCNN	6.5600	1.3125	0.8850	0.3450	0.3912	0.8062	0.5456	30.5133	0.8883
Wavefuse	6.8877	1.4285	0.8892	0.3702	0.4202	0.8084	0.5525	39.3585	0.8770

Subjective Evaluation. Examples of the original image pairs and the fusion results obtained by each comparative method for the four scenarios are shown in Fig. 2.

Multi-scene Image Fusion: We first compare the proposed WaveFuse with existing multi-scene image fusion algorithms U2Fusion [27] and IFCNN [29] in all the four different fusion scenarios, and the results of the objective metrics are shown in Table 1. From Table 1, we can see that WaveFuse achieves the best results in almost all scenarios. In some scenarios that it does not achieve the highest metric, our method is still close to the highest one.

Multi-exposure Image Fusion: The multi-exposure image fusion aims to combine different exposures to generate better subjective images in both dark and bright regions. From Fig. 2 (c1-j1, c2-j2), we can observe that CBF and ConvSR generate many artifacts. JSRSD, DWT, DeepFuse and IFCNN suffer from low brightness and blurred details. U2Fusion and WaveFuse achieve better fusion reluslts considerding both dark and bright factors.

Multi-modal Medical Image Fusion: Multi-modal medical image fusion can offer more accurate and effective information for biomedical research and clinical applications. Better multi-modal medical fused image should provide combined features sufficiently and preserve both significant textural features. As shown in Fig. 2 (c3-j3, c4-j4), JSR, JSRSD and ConvSR shows obvious artifacts in the whole image. DWT and CBF fail to preserve the crucial features of the source images. U2Fusion shows better visual results than the above-mentioned methods. However, DenseFuse still weakens the details and brightness. Information-rich fused images can be obtained by IFCNN. In contrast, our method preserves the details and edge information of both source images, which is more in line with the perception characteristics of the human vision compared to other fusion methods.

Multi-focus Image Fusion: The multi-focus image fusion aims to reconstruct a fully focused image from partly focused images of the same scene. From Fig. 2 (c5-j5, c6-j6), we can observe that JSR and JSRSD shows obvious blurred artifacts. DWT shows low brightness in the fusion results. Other compared methods perform well.

Infrared/Visible Image Fusion: Visible images can capture more detail information compared to infrared images. However, the interested objects can not be easily observed in visible image especially when it is under low contrast circumstance and the light is insufficient. Infrared images can provide thermal radiation information, making it easy to detect the salient object even in complex background. Thus, the fused image can provide more complementary information. Figure 2 (c7-j7, c8-j8) show fusion results of infrared and visible images with the comparison methods. JSR, JSRSD and FusionGAN exhibit significant artifacts., and U2Fusion shows unclear salient objects. The results in DWT, DenseFuse and IFCNN weaken the contrast. We can see that, WaveFuse preserves more details in high contrast and brightness.

Objective Evaluation. From Fig. 2, we can observe that the fusion results of CBF and ConvSR in multi-exposure images, the fusion results of CBF, JSR,
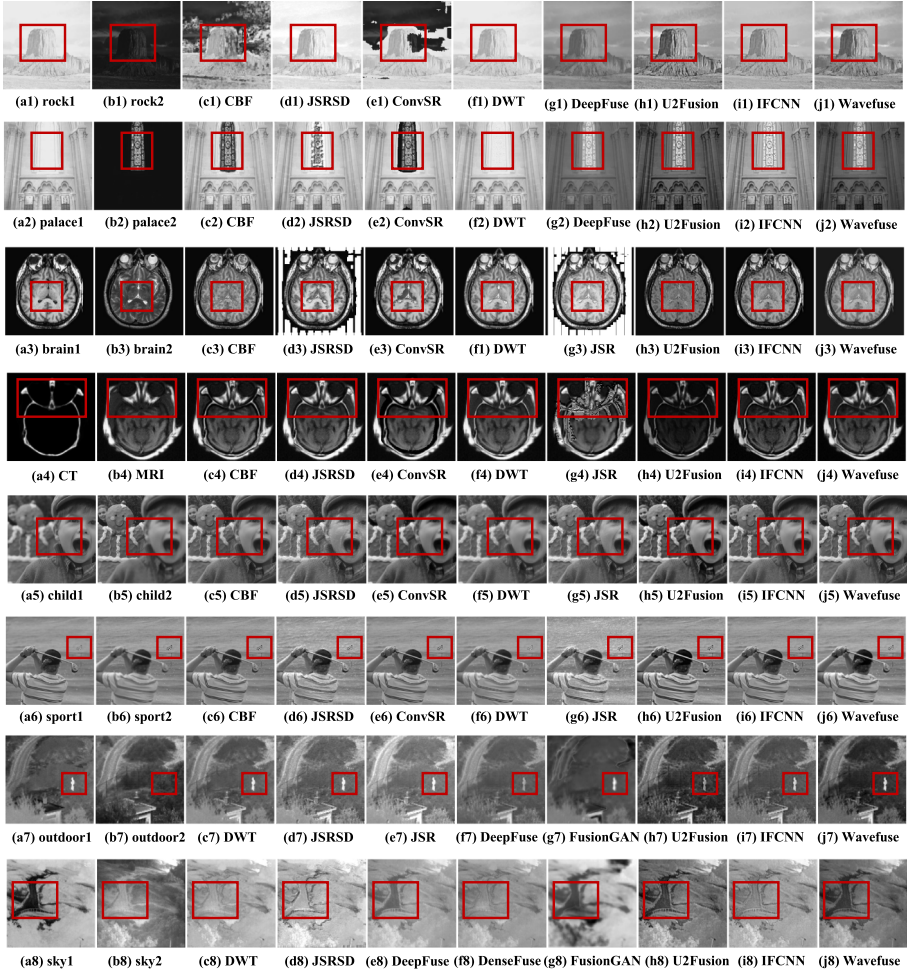
Fig. 2. Fusion results by different methods. (a1)-(b1),(a2)-(b2) are two pairs of multi-exposure source images and (c1)-(j1),(c2)-(j2) are the fusion results of them by different methods; (a3)-(b3),(a4)-(b4) are two pairs of multi-modal medical source images and (c3)-(j3),(c4)-(j4) are the fusion results of them by different methods; (a5)-(b5),(a6)-(b6) are two pairs of multi-focus source images and (c5)-(j5),(c6)-(j6) are the fusion results of them by different methods; (a7)-(b7),(a8)-(b8) are two pairs of infrared and visible source images and (c7)-(j7),(c8)-(j8) are the fusion results of them by different methods;

JSRSD and ConvSR in multi-modal medical images and the fusion results of JSRSD in multi-focus images contain poor visual effects owing to considerable artificial noise, and in this case their objective quality metrics will not be calculated for the quantitative evaluation.

Table 2. The average values of fusion quality metrics for fused images of four different scenarios. Red ones are the best results, and blue ones mark the second results. For all metrics, larger is better.

ME	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
DWT	5.9089	0.7402	0.8510	0.3592	0.3826	0.8165	0.4603	33.8954	0.8936
JSRSD	6.4366	1.4055	0.8533	0.2649	0.3026	0.8150	0.4257	36.3555	0.9001
DeepFuse	6.5352	0.8456	0.8556	0.4360	0.4462	0.8163	0.4573	30.7345	0.9115
U2Fusion	6.9525	4.7649	0.8529	0.4005	0.4411	0.8146	0.4319	41.5841	0.9349
IFCNN	6.5269	2.1552	0.8590	0.4580	0.4892	0.8134	0.4788	34.4859	0.9478
WaveFuse	6.9224	4.7309	0.8661	0.4941	0.5265	0.8240	0.5149	39.8423	0.9582
MED	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
DWT	5.1344	1.1957	0.8522	0.2472	0.3730	0.8090	0.3171	82.7802	0.8974
U2Fusion	5.1257	7.5716	0.8430	0.2876	0.3471	0.8067	0.2879	57.8377	0.8980
IFCNN	5.0101	1.2406	0.8581	0.3611	0.4450	0.8075	0.3240	73.6683	0.9434
WaveFuse	5.2814	7.8243	0.8650	0.4052	0.3848	0.8094	0.3265	67.0572	0.9007
MF	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
DWT	7.2436	0.0406	0.8811	0.3716	0.4908	0.8360	0.7530	46.9477	0.9706
DeepFuse	7.2488	0.4975	0.8791	0.4340	0.4868	0.8292	0.7361	47.8298	0.9812
U2Fusion	7.4552	0.3424	0.8706	0.4018	0.4653	0.8269	0.6586	55.5703	0.9574
IFCNN	7.2756	0.0558	0.8903	0.4687	0.5389	0.8356	0.7447	48.6216	0.9947
WaveFuse	7.3681	0.3483	0.8912	0.5041	0.5619	0.8368	0.7743	51.7256	0.9896
IV	EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{AB/F}$	VARI	MS-SSIM
DWT	6.6684	0.9913	0.8799	0.2244	0.3117	0.8009	0.5416	34.9068	0.7703
JSRSD	6.7557	1.3340	0.8482	0.1439	0.1924	0.8056	0.3976	36.1247	0.7636
DeepFuse	6.1759	1.0604	0.8799	0.3346	0.3718	0.8051	0.5044	22.9167	0.8473
DenseFuse	6.3550	1.0577	0.8909	0.3592	0.4157	0.8083	0.5458	27.8602	0.8131
FusionGAN	6.6208	2.3257	0.8729	0.2334	0.2937	0.8080	0.3266	40.1621	0.3963
U2Fusion	6.9092	1.1991	0.8795	0.3158	0.3546	0.8055	0.4775	37.1806	0.9196
IFCNN	6.5600	1.3125	0.8850	0.3450	0.3912	0.8062	0.5456	30.5133	0.8883
WaveFuse	6.8877	1.4285	0.8892	0.3702	0.4202	0.8084	0.5525	39.3585	0.8770

Table 2 shows the average values of the fusion quality metrics among four different fusion tasks by different fusion methods. In multi-exposure image fusion, our method ranks first in FMI_pixel, FMI_dct, FMI_w, Q^{NICE}, $Q^{AB/F}$ and MS-SSIM, and ranks second in EN, CE and VARI. In multi-modl medical image fusion, our method ranks first in EN, CE, FMI_pixel, FMI_dct, Q^{NICE} and $Q^{AB/F}$, and ranks second in FMI_w and MS-SSIM. In multi-focus image fusion, our method ranks first in FMI_pixel, FMI_dct, FMI_w, Q^{NICE} and $Q^{AB/F}$, and ranks second in EN, CE, VARI and MS-SSIM . In the infrared and visible image fusion, our method obtains the highest metrics in FMI_dct, FMI_w, Q^{NICE} and $Q^{AB/F}$, and ranks second in EN, FMI_pixel and VARI. Furthermore, from the value of the last row among three image fusion task in Overall, compared with other peer methods, our proposed method achieves the highest values in most fusion quality metrics and the second in the remaining metrics.

3.3   Comparison of Using Different Training Dataset

In order to further demonstrate the effectiveness and robustness of our network, we conducted experiments on another three different training minisets: MINI1-MINI3, each of which contains 0.5%, 1% and 2% images respectively chosen

randomly from COCO, and the fusion results difference can be found among the subjective fused images, we compared the objective fusion results of WaveFuse on COCO and MINI1-MINI3. The fusion performance was compared and analyzed by the averaged fusion quality metrics. In WaveFuse, higher performance is even achieved by training on minisets. Due to page limitation, further details can be found in [13].

Furthermore, we can observe that WaveFuse is trained on minisets within one hour, where the GPU memory utilization is just 4085 MB, so it can be trained with lower computational cost compared with that trained in COCO (7.78h and 17345 MB). Accordingly, we can learn that our proposed network is robust both to the size of the training dataset and to the selection of training images.

3.4 Ablation Studies

- **DWT-based feature fusion.** In this section, we attempt to explain why DWT-based feature fusion module can improve fusion performance. DWT has been a poweful multi-sacle analysis tool in signal and image processing since it was proposed. DWT transforms the images into different low and high frequencies, where low frequencies represent contour and edge information and high frequencies represent detailed texture information [6]. In this way, DWT-based fusion methods first transform the images into low and high frequencies, and then fuse them in the wavelet domain, achieving promising fusion results. Inspired by DWT methods, we apply DWT-based fusion module to deep feature fusion extracted by deep-learning models, so as to fully utilize the information contained in deep features. We conducted the ablation study about DWT-based feature fusion module, and the results is shown in Table 3. As we can see, when we apply the module, the fusion performance is indeed improved largely.

Table 3. Ablation study on the DWT-based feature fusion module. Red ones are the best results. For all metrics, larger is better.

		EN	CE	FMI_pixel	FMI_dct	FMI_w	Q^{NICE}	$Q^{A/BF}$	VARI	MS-SSIM
ME	BaseLine	6.8230	**4.9708**	0.8473	0.2500	0.2875	0.8153	0.4451	37.5291	0.9566
	WaveFuse	**6.9224**	4.7309	**0.8661**	**0.4941**	**0.5265**	**0.8240**	**0.5149**	**39.8423**	**0.9582**
MED	BaseLine	**5.5631**	**7.8961**	0.8505	0.2098	0.2274	0.8078	0.2911	66.7267	0.8970
	WaveFuse	5.2814	7.8243	**0.8650**	**0.4052**	**0.3848**	**0.8094**	**0.3265**	**67.0572**	**0.9007**
MF	BaseLine	**7.3759**	0.3276	0.8663	0.2686	0.3314	0.8290	0.6821	51.0500	0.9812
	WaveFuse	7.3679	**0.3498**	**0.8909**	**0.5028**	**0.5609**	**0.8368**	**0.7746**	**51.6778**	**0.9889**
IV	baseline	6.7367	1.4337	0.8776	0.1919	0.2443	0.8054	0.5036	33.1956	0.9182
	WaveFuse	**6.7805**	**1.4595**	**0.8891**	**0.3853**	**0.4122**	**0.8057**	**0.5449**	**34.2832**	**0.9268**

- **Experiments on Different Wavelet Decomposition Layers and Different Wavelet Bases.** In wavelet transform, the number of decomposition layers and the selection of different wavelet bases could exert great impacts on the effectiveness of wavelet transform. We also conducted the ablation

study on different settings. Due to the page limitation, we just give our final conclusion, when the number of decomposition layers and the wavelet base are set as 2 and *db1* respectively, we achieve the best fusion results. More details can be found in [13].

4   Conclusions

In this paper, we propose a novel image fusion method through the combination of a multi-scale discrete wavelet transform based on regional energy and deep learning. To our best knowledge, this is the first time that a conventional technique is integrated for feature maps fusion in the pipeline of deep learning based image fusion methods, and we think there are still a lot of possibilities to explore in this direction.

Our network consists of three parts: an encoder, a DWT-based fusion part and a decoder. The features of the input image are extracted by the encoder, then we use the adaptive fusion rule at the fusion layer to obtain the fused features, and finally reconstruct the fused image through the decoder. Compared with existing fusion algorithms, our proposed method achieves better performance. Additionally, our network has strong universality and can be applied to various image fusion scenarios. At the same time, our network can be trained in smaller datasets to obtain the comparable fusion results trained in large datasets with shorter training time and higher efficiency, alleviating the dependence on large datasets. Extensive experiments on different wavelet decomposition layers and bases demonstrate the possibility of further improvement of our method.

References

1. Da Cunha, A.L., Zhou, J., Do, M.N.: The nonsubsampled contourlet transform: theory, design, and applications. IEEE Trans. Image Process. **15**(10), 3089–3101 (2006)
2. Du, C., Gao, S.: Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network. IEEE Access **5**, 15750–15761 (2017)
3. Goshtasby, A.A., Nikolov, S.: Image fusion: advances in the state of the art. Inform. Fus. **2**(8), 114–118 (2007)
4. Haghighat, M., Razian, M.A.: Fast-fmi: non-reference image fusion metric. In: 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–3. IEEE (2014)
5. Shreyamsha Kumar, B.K.: Image fusion based on pixel significance using cross bilateral filter. Sign. Image Video Process. **9**(5), 1193–1204 (2013). https://doi.org/10.1007/s11760-013-0556-9
6. Li, H., Manjunath, B., Mitra, S.K.: Multisensor image fusion using the wavelet transform. Graph. Models Image Process. **57**(3), 235–245 (1995)
7. Li, H., Wu, X.J.: Densefuse: a fusion approach to infrared and visible images. IEEE Trans. Image Process. **28**(5), 2614–2623 (2018)
8. Li, H., Wu, X.J., Durrani, T.S.: Infrared and visible image fusion with resnet and zero-phase component analysis. Infrared Phys. Technol. **102**, 103039 (2019)

9. Li, S., Kang, X.: Fast multi-exposure image fusion with median filter and recursive filter. IEEE Trans. Consum. Electron. **58**(2), 626–632 (2012)

10. Li, S., Kang, X., Fang, L., Hu, J., Yin, H.: Pixel-level image fusion: a survey of the state of the art. Inform. Fus. **33**, 100–112 (2017)

11. Li, S., Kwok, J.T., Wang, Y.: Combination of images with diverse focuses using the spatial frequency. Inform. Fus. **2**(3), 169–176 (2001)

12. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

13. Liu, S.: Fusion rule description (2021). https://github.com/slliuEric/WaveFuse_code

14. Liu, Y., Chen, X., Cheng, J., Peng, H.: A medical image fusion method based on convolutional neural networks. In: 2017 20th International Conference on Information Fusion (Fusion), pp. 1–7. IEEE (2017)

15. Liu, Y., Chen, X., Peng, H., Wang, Z.: Multi-focus image fusion with a deep convolutional neural network. Inform. Fus. **36**, 191–207 (2017)

16. Liu, Y., Chen, X., Wang, Z., Wang, Z.J., Ward, R.K., Wang, X.: Deep learning for pixel-level image fusion: recent advances and future prospects. Inform. Fus. **42**, 158–173 (2018)

17. Liu, Y., Chen, X., Ward, R.K., Wang, Z.J.: Image fusion with convolutional sparse representation. IEEE Sign. Process. Lett. **23**(12), 1882–1886 (2016)

18. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: a generative adversarial network for infrared and visible image fusion. Inform. Fus. **48**, 11–26 (2019)

19. Ma, K., Zeng, K., Wang, Z.: Perceptual quality assessment for multi-exposure image fusion. IEEE Trans. Image Process. **24**(11), 3345–3356 (2015)

20. Prabhakar, K.R., Srikar, V.S., Babu, R.V.: Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: ICCV, pp. 4724–4732 (2017)

21. Roberts, J.W., Van Aardt, J.A., Ahmed, F.B.: Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J. Appl. Remote Sens. **2**(1), 023522 (2008)

22. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)

23. Song, X., Wu, X.-J., Li, H.: MSDNet for medical image fusion. In: Zhao, Y., Barnes, N., Chen, B., Westermann, R., Kong, X., Lin, C. (eds.) ICIG 2019. LNCS, vol. 11902, pp. 278–288. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34110-7_24

24. Wang, Q., Shen, Y.: Performances evaluation of image fusion techniques based on nonlinear correlation measurement. In: Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No. 04CH37510), vol. 1, pp. 472–475. IEEE (2004)

25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

26. Xiao-hua, S., Yang, G.S., Zhang, H.l.: Improved on the approach of image fusion based on region-energy. Journal of Projectiles, Rockets, Missiles and Guidance, vol. 4 (2006)

27. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)

28. Xydeas, C., Petrovic, V.: Objective image fusion performance measure. Electron. Lett. **36**(4), 308–309 (2000)
29. Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: Ifcnn: a general image fusion framework based on convolutional neural network. Inform. Fus. **54**, 99–118 (2020)