

Fine-grained multi-focus image fusion based on edge features

Bin Tian¹, Lichun Yang², and Jianwu Dang^{*3}

^{1,3}Lanzhou Jiaotong University, College of Electronics and Information Engineering, Gansu, Lanzhou, China, 730070

²Lanzhou Jiaotong University, Key Laboratory of Photoelectric Technology and Intelligent Control Ministry of Education, Gansu, Lanzhou, China, 730070

^{*}ylc1377759045@163.com

Abstract

Multi-focus image fusion is the process of fusing multiple images of different focus areas into a total focus image, which has important application value. In view of the shortcomings of the current fusion method in the global information retention effect of the original image, a fusion architecture based on two stages is designed. In the training phase, the polarization self-attention module is combined with the DenseNet network to design an encoder-decoder structure for image reconstruction tasks to enhance the global information attention ability of the model. In the fusion stage, combined with the edge feature information of the encoded feature map, a fusion strategy based on edge feature map is designed for image fusion task to enhance the detail attention ability of the fusion algorithm. Compared with 7 classical fusion algorithms, the proposed algorithm achieves advanced fusion performance in both subjective and objective evaluation, and the fused image has better information retention effect on the original image.

Keywords: Multi-focus Image Fusion; DenseNet; Polarizing Self-Attention; Edge Feature Map

1 Introduction

Image fusion refers to the process of extracting the most meaningful information from multiple different source images and generating an image with more information and more beneficial to subsequent applications[1]. As a branch of image fusion, multi-focus image fusion has important research significance in the field of imaging. Due to the limitations of optical lenses, it is difficult to focus all objects at different depths of field during camera shooting[2]. The main task

of multi-focus image fusion is to fuse multiple images of different focus areas to get a total focus image. Traditional image fusion methods can be classified into four categories : multi-scale transform, sparse representation, subspace analysis and hybrid method. However, these methods have the problems of poor robustness of feature extraction and complex fusion rules[3]. In recent years, with the development of deep learning, neural network has been widely used in the field of image fusion with its powerful feature extraction ability and information reconstruction ability..

Liu[4] et al.first introduced convolutional neural networks into the field of image fusion, and constructed a convolutional neural network architecture based on Siamese structure for multi-focus image fusion tasks. The convolution neural network is used to realize the discrimination task of clear pixels of image blocks, and the final fusion task is realized through a series of post-processing operations. Compared with the traditional algorithm, the experiment has achieved remarkable results and has better fusion performance. However, the training cost of this method is high, requiring millions of pixel blocks and thousands of iterations to achieve a good discriminative performance. Since then, Ram Prabhakar K[5] et al.introduced unsupervised learning into the field of image fusion and proposed DeepFuse as a general image fusion framework. DeepFuse adopts an encoder-decoder structure and uses a structural similarity function as a loss function for image reconstruction tasks. In the fusion stage, the original image is directly encoded and added by the encoder, and then sent to the decoder for decoding output, which realizes the end-to-end fusion of the image. However, DeepFuse's direct addition fusion method is rough, so its fused image still retains the blurred information of the defocused area.To this end, Li[6] et al.proposed the DenseFuse fusion task. DenseFuse uses the DenseNet structure to design the encoder, which is designed to preserve the multi-scale information of the original image. At the same time, DenseFuse improved the fusion strategy and proposed L_1 +norm fusion strategy, which achieved better results. SESF-Fuse proposed by Boyuan Ma[7] et al. considered the properties of multi-focus images, designed a fusion strategy based on feature gradient, and achieved better discrimination performance.The SDNet proposed by Zhang[8] et al. constructs the fusion task directly from the gradient information and intensity information of the original image.Different from the traditional encoder-decoder structure, the result of SDNet coding is the fused image. The decoding process is the decoding process from the fused image to the original image. SDNet has achieved advanced fusion performance on various fusion tasks, and has a good retention effect on the texture details and brightness information of the original image. Although it is feasible to use the gradient information and intensity information of the image to consider the multi-focus fusion problem, it is not enough to consider these two features. Therefore, the fused image has obvious color difference problem compared with the original image, and there is also a little loss of detail information.DIFNet proposed by Hyungjoo Jung[9] et al. used the eigenvector of the structure tensor to represent the directions of maximum and minimum contrast of multi-channel images and combined the intensity information of images to construct a general fusion architecture. Although DIFNet is

simple in structure. However, the fusion strategy of DIFNet is also rough, so the fused image still has the defocus information of the original image, failing to maximize the effective information. Zhang[10] et al. proposed MFF-GAN for multi-focus image fusion task based on image texture details. MFF-GAN uses generative adversarial networks and self-supervised learning for image fusion tasks, and achieves better texture detail retention, especially at the boundary of the focus area. However, because MFF-GAN pays too much attention to the texture details of the image, the attention to the global information of the image is reduced, resulting in the lack of overall information retention of the fused image.

In order to further improve the quality of multi-focus image fusion, the original image not only has better global information retention effect, but also has better ability to retain details. A two-stage image fusion algorithm based on fine-grained features and edge feature maps of the original image is proposed. Specific contributions are as follows.

- From the perspective of global information attention, an encoder-decoder structure network based on DenseNet network and polarized self-attention module is designed.
- From the perspective of local pixel discrimination, an image fusion strategy based on edge feature map is designed.

2 Methods

2.1 Overview of Methods

As shown in Figure 1, the fusion algorithm is divided into two stages. In the training phase, the DenseNet network structure[11] and the polarization self-attention module [12](PSA) are used to construct the encoder decoder structure for image reconstruction tasks.(Among them, the use of DenseNet network structure is mainly to retain the original image of the multi-level information as much as possible.) The polarization self-attention module is mainly to enhance the performance of the network on fine-grained pixel-level tasks. Under low computational overhead, the long-distance dependence of high-resolution input / output features is enhanced to estimate highly nonlinear pixel semantics.)In the fusion stage, the original image information is encoded by the trained encoder, and then cross-subtracted with the encoded feature map after Gaussian smoothing and guided filtering respectively to obtain the corresponding edge feature map. The gray variance product function is used to identify the focus area, and the final image fusion task is performed by weighting after mathematical morphology processing.

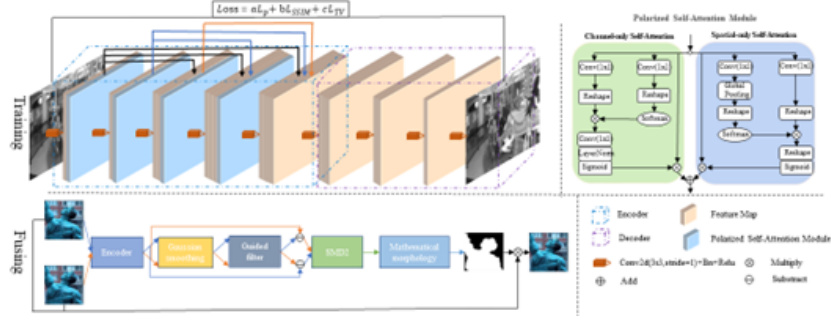


Figure 1: Fusion architecture diagram

2.2 Loss Function

This paper follows the comprehensive loss function used by TransMEF[13], but reconsiders the weight design of each loss function.(Given that TransMEF is primarily used for multi-exposure fusion tasks) the details are as follows.

$$L = aL_2 + bL_{ssim} + cL_{TV} \quad (1)$$

where L_2 represents the root mean square error loss function; L_{ssim} represents the structural similarity loss function ; L_{TV} represents the total variance loss function. a, b, c are the weight corresponding to each loss function. Different from TransMEF, this paper sets a, b, c as 1, 20, 1 respectively to enhance the model's ability to pay attention to the structural similarity of reconstructed images in image reconstruction training tasks.

The L_2 loss function is used to measure the degree of pixel loss in the reconstruction task. The calculation method is as follows:

$$L_2 = ||L_{out} - L_{in}||_2 \quad (2)$$

where L_{out} and L_{in} represent the input image and the reconstructed image, respectively, and $|| \cdot ||_2$ represents the L_2 norm.

The L_{ssim} loss function is used to measure the transfer of structural information in the reconstruction task. The calculation method is as follows:

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

L_{TV} comes from VIFNet[14]. It is mainly used to consider the gradient information retention of the reconstructed image and further eliminate the influence of noise. It is calculated as follows:

$$R(p, q) = I_{out}(p, q) - I_{in}(p, q) \quad (4)$$

$$L_{TV} = \sum_{p,q} (||R(p, q+1) - R(p, q)||_2 + ||R(p+1, q) - R(p, q)||_2) \quad (5)$$

where $R(p, q)$ represents the difference between the source image and the reconstructed image ; $\|\cdot\|_2$ represents the L_2 norm; p and q represent the horizontal and vertical coordinates of the image pixels, respectively.

2.3 Fusion Strategy

2.3.1 Decision Diagram Generation

In the fusion stage, a trained encoder is used to encode the fused image. Then, the encoded feature map is cross-subtracted with the original feature map using a Gaussian smoothing filter template with a size of 11×11 and guided filtering processing to obtain the corresponding edge feature map. The gray variance product function(SMD2) is used to determine the pixel level of image clarity according to the channel[15]. After mathematical morphology processing(the morphological operator of the element is used here, and the threshold of $0.1 \times H \times W$ is used to eliminate the independent small area), the final decision diagram is obtained.

$$f_r(x, y) = \sqrt{\sum_{a=-r}^r \sum_{b=-r}^r [F(x+a, y+b) - F(x+a+1, y+b)]^2} \quad (6)$$

$$f_c(x, y) = \sqrt{\sum_{a=-r}^r \sum_{b=-r}^r [F(x+a, y+b) - F(x+a, y+b+1)]^2} \quad (7)$$

$$F(x, y) = \sqrt{\frac{|f_r(x, y)| |f_c(x, y)|}{(2r+1)^2}} \quad (8)$$

$F_r(x, y)$ and $F_c(x, y)$ represent the row and column frequencies of the image, respectively.

Through the above method, the spatial frequency calculation results of images at different focusing positions are obtained respectively. The method of generating the initial weight map according to the frequency calculation results is shown in formula 9.

$$D(x, y) = \begin{cases} 1 & \text{if } F_1(x, y) \geq F_2(x, y) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

2.3.2 Weight Fusion

According to the generated final decision map, the final image fusion is performed in a weighted manner, and the calculation method is shown in Formula 10.

$$F(i, j) = D(i, j)A(i, j) + (1 - D(i, j))B(i, j) \quad (10)$$

2.4 Fusion Evaluation Indicators

So far, there is no evaluation method that can be applied to all fusion algorithms. Therefore, subjective evaluation and objective evaluation indexes are used to evaluate the fused image from two aspects. In the objective evaluation index, MI(mutual information)[16], SSIM(structural similarity)[17], PSNR(peak signal to noise ratio)[18], AG(evaluation gradient)[19], SF(spatial frequency)[20], Q^{MI} [21] and $Q^{AB/F}$ [22] were used to evaluate the fused image. The specific explanation is as follows.

MI(Mutual Information)

Mutual information represents how much information of the source image is transferred to the fused image. The larger the mutual information value, the better the image fusion effect.

SSIM(structural similarity)

SSIM is used to measure the structural similarity between the fused image and the source image. The larger the SSIM value, the more similar the structure of the fused image to the structure of the source image.

PSNR(peak signal-to-noise ratio)

The peak signal-to-noise ratio is used to measure the ratio between the effective information of the image and the noise, which can reflect whether the image is distorted. The greater the PSNR, the better the image quality.

AG(average gradient)

The average gradient can reflect the ability of the image to express the contrast of small details, and also reflect the characteristics of texture transformation in the image. Therefore, the greater the average gradient, the clearer the image.

SF(Spatial Frequency)

Spatial frequency reflects the overall activity of the image in the spatial domain, that is, the change rate of image gray. The larger the spatial frequency, the better the quality of fused image.

Q^{MI}

Q^{MI} eliminates the impact of information entropy on the fusion image quality assessment results, can better measure the amount of information transfer between the original image and the fused image. The larger the Q^{MI} , the better the image fusion.

$Q^{AB/F}$

$Q^{AB/F}$ can better reflect how much edge information in the source image is transmitted to the fused image. The greater the value of $Q^{AB/F}$ indicates that the fused image retains more edge information of the source image and the fusion effect is better.

3 Experimentats

3.1 Experimental Environment

In this paper, 10,000 images in the MS-COCO2014 dataset[23] were selected for training, and 38 pairs of public multi-focus images[24, 25] were selected as evaluation tests. The images used in the training phase were all subjected to data enhancement processing(randomly cropped to a size of 64×64 , randomly rotated, and randomly performed brightness, contrast, and tone conversion). The batchSize selected in this article is 16, the learning rate is 1×10^{-4} , and the number of iterations is 50. AdamW optimizer and Warm up learning rate strategy are used for training.

This paper uses Python language, Pycharm integrated development environment, mainly using Pytorch, Skimage and OpenCV library for experiments. The GPU platform used is NVIDIA 2060.

3.2 Image Fusion Evaluation

In order to measure the performance of the fusion algorithm in this paper, seven algorithms, CNN[4], DeepFuse[5], DenseFuse[6], SESF-Fuse[7], SDNet[8], DIFNet[9] and MFF-GAN[10], are selected here. The subjective and objective evaluation indexes are used to compare with the algorithm in this paper, as follows.

3.2.1 Subjective Assessment

The algorithm in this paper is used for image fusion, and the results of 7 pairs of images before and after fusion and corresponding decision diagrams are shown in Fig.2.



Figure 2: Fusion image results

It can be seen from Figure 2 that the algorithm in this paper can accurately

identify the clear and non-clear areas of the focused image, and the fused image is subjectively clearer. In order to further analyze the effect differences of various algorithms, several different fusion images are selected to compare the local image clarity after fusion and the difference between the fused image and the original image.

The comparison results of the local enlarged images of the fused images are shown in Figure 3-4.

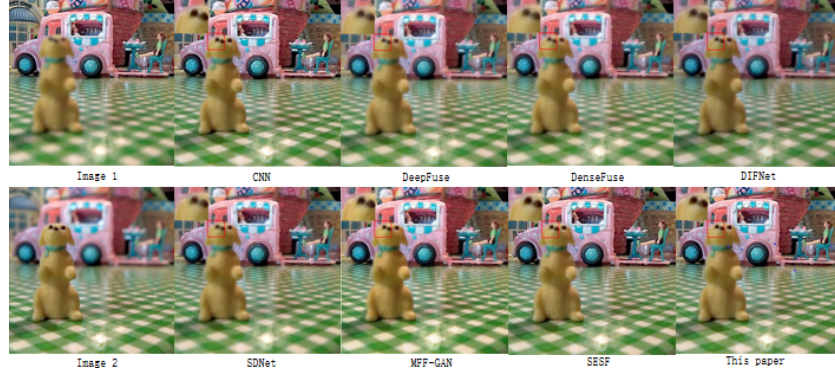


Figure 3: Image local comparison 1

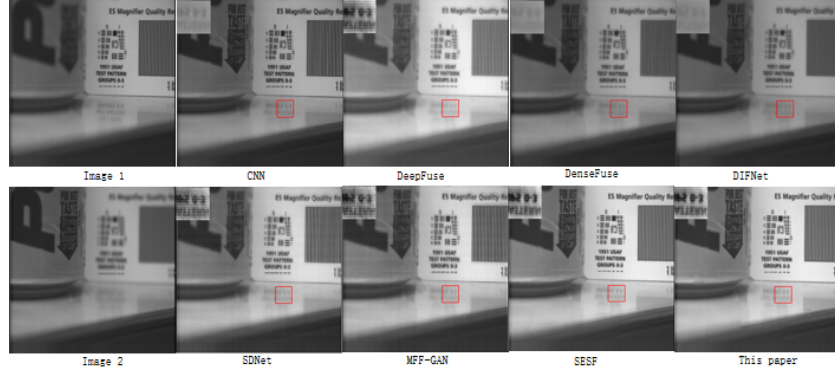


Figure 4: Image local comparison 2

It can be seen from Figure 3 that in the local enlarged image of the fused image, DeepFuse, DenseFuse, DIFNet and SDNet algorithms are used for multi-focus image fusion tasks, and the fused image has obvious blurring phenomenon. CNN, MFF-GAN, SESF and the proposed algorithm are relatively good. The reason is that DeepFuse and DenseFuse use a simple coding feature map addition method for fusion tasks in the fusion process, and the final fusion image is obtained after decoding by the decoder. This fusion method cannot consider the difference information between the original images well, and only plays a

comprehensive role in the information of the original images, which weakens the difference information between the original images to some extent. DIFNet also has the same problem in image fusion using concat, so the fused image still retains the information of the defocused area of the original image. SDNet introduces the gradient information and intensity information of the image to consider the fused image. Compared with DeepFuse and DenseFuse, the fusion effect is relatively better, but there is still some loss of detail information. MFF-GAN starts from the preservation of image texture information, so the fused image has a better effect on the preservation of detail information. However, it can be seen from Figure 4 that the effect of MFF-GAN on preserving the details of relatively flat areas in the original image is slightly inferior to that of SESF algorithm and this algorithm, and there is a little ambiguity. The fusion effect of CNN is similar to that of MFF-GAN. For a flat area like Figure 4, its fusion performance is still worse than that of SESF and the proposed algorithm.

In order to more accurately measure the information loss of the fused image of each algorithm, the difference image (the difference between the fused image and the original image) is considered. The specific results are shown in Figure 5-6.

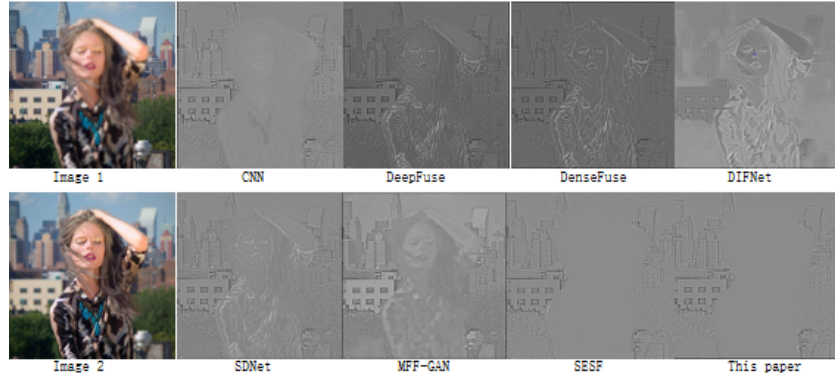


Figure 5: Comparison of difference between fused image and original image 1

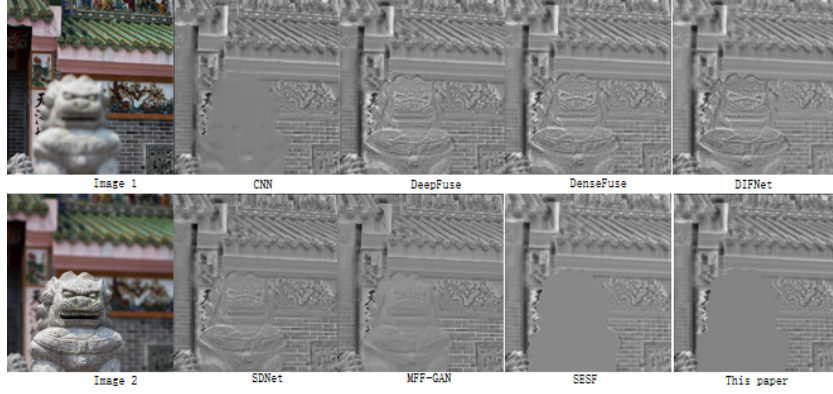


Figure 6: Comparison of difference between fused image and original image 2

From Figure 5, it can be seen that the fusion images of DeepFuse, DenseFuse and DIFNet have obvious loss of texture details of the original image, and more texture details of the original image are not preserved. Although SDNet has better retention effect on the texture information of the image than these two algorithms, there are still a lot of texture information loss. MFF-GAN retains the texture information of the original image better, but for other flat areas, because this part of the region has less texture information, its retention effect is relatively poor. CNN although the overall retention effect is good, but there are still obvious flaws. In comparison, SESF and the algorithm in this paper have a better effect on the overall information retention of the focus area of the original image, and Figure 6 also proves this conclusion.

3.2.2 Objective Appraisal

In order to better quantify and compare the fusion effects of various algorithms, MI(mutual information), SSIM(structural similarity), PSNR(peak signal-to-noise ratio), AG(average gradient), SF(spatial frequency), Q^{MI} and $Q^{AB/F}$ are used to compare the fusion images. The results are shown in Table 1.

Table 1: Comparison of various algorithms

Algorithms	Index Evaluation							
	EN	MI	SSIM	PSNR	AG	SF	Q^{MI}	$Q^{AB/F}$
CNN	7.1630	0.8399	0.8099	25.8275	8.4265	19.2010	6.0823	0.4431
DeepFuse	7.3671	0.7841	0.8503	26.2962	5.4042	12.0751	5.7598	0.3460
DenseFuse	7.2727	0.7889	0.8570	28.2387	5.1400	11.4805	5.7582	0.3360
SDNet	7.3786	0.7745	0.8339	26.3376	7.5673	17.4520	5.6940	0.3905
DIFNet	7.1488	0.7682	0.8428	25.4679	4.7213	10.5981	5.5582	0.2604
MFF-GAN	7.3856	0.7432	0.8116	24.7821	8.6212	19.8732	5.4468	0.3779
SESF	7.3683	1.1115	0.8058	26.2059	8.3985	19.2819	8.1600	0.4643
This paper	7.3654	1.1469	0.8054	26.2222	8.4013	19.2458	8.4180	0.4664

Red is the best, blue is the second best.

It can be seen from Table 1 that in the vertical comparison, the proposed algorithm achieves three optimal and one suboptimal fusion performance, which is better than the other seven fusion algorithms. In the horizontal comparison, compared with CNN, DeepFuse, DenseFuse, SDNet, DIFNet, MFF-GAN and SESF, the proposed algorithm has 5,5,5,5,6,4 and 5 indicators superior to them, respectively.

This algorithm has better information transfer ability and better edge retention effect. Compared with SESF, the algorithm in this paper is more effective and has been improved in many indicators. However, the algorithm in this paper is still slightly inferior to the MFF-GAN algorithm in the retention of texture details of the image. The reason is that the MFF-GAN algorithm focuses on the texture information of the original image. This method focuses on retaining the texture detail information of the original image, but it also has some defects in retaining the information of the flat area of the original image. As a result, the difference between the two becomes larger, so MFF-GAN achieves better performance on AG and SF metrics. However, this also has a negative impact, resulting in poor performance of the MFF-GAN algorithm on other indicators. This is consistent with the conclusion in the subjective evaluation analysis.

In summary, combining the results of subjective evaluation and objective evaluation, it can be seen that the proposed algorithm has better original information retention ability than other algorithms in multi-focus fusion tasks. Overall, better fusion performance was achieved.

3.3 Ablation Experiment

This paper explores the loss function, model construction and fusion strategy. In order to verify the validity of each study, the confirmatory analysis is carried out from the above three aspects, as follows.

3.3.1 Loss Function Analysis

In the image reconstruction task, this paper adopts the comprehensive loss function in TransMEF[13] and adjusts its weight. Analyzing the effect of each loss function, the mean and variance of reconstruction loss($L_2 + 20L_{ssim} + L_{TV}$) are shown in Table 2.

Table 2: Comparison of loss functions

Reconstruction loss	$L_2 + 20L_{ssim} + L_{TV}$	$20L_2 + L_{ssim} + 20L_{TV}$	$L_2 + 1000L_{ssim}$
Mean	0.1304	0.8373	0.4473
STD	0.1519	0.9499	0.5854

It can be seen from Table 2 that the effect of image reconstruction task using SESF loss function($L_2 + 1000L_{ssim}$) and TransMEF original loss function($20L_2 + L_{ssim} + 20L_{TV}$) is worse than the loss function used in this paper. Therefore, the loss function used in this paper is reasonable.

3.3.2 Attention Mechanism Analysis

In the image reconstruction task, multiple attention mechanisms are used to construct the model, and the corresponding model reconstruction loss is shown in Table 3.

Table 3: Performance comparison of various attention mechanisms

Reconstruction loss	PSA	CBAM[26]	SENet[27]	GCNet[28]
Mean	0.1304	0.2371	0.1653	0.1923
STD	0.1519	0.4118	0.2451	0.1152

From the reconstruction loss results of the original image encoding and decoding process using various attention mechanisms in Table 3, it can be seen that the construction of the model using PSA has a smaller reconstruction loss than other attention mechanisms. The reason is that CBAM uses a combination of space and channel to pay attention to image information. But this is often goal-based, and better at focusing on local information, the ability to focus on global information is insufficient. The way that SENet uses global context to weight different channels to adjust channel dependencies does not make full use of global context information. Although GCNet has stronger global attention ability, in this task, its attention effect has not been obviously reflected in the face of pixel blocks with a size of 64×64 after clipping. PSA combines the advantages of CBAM channel attention mechanism and spatial attention mechanism, and integrates the global modeling ability of self-attention mechanism.

Without increasing the computational complexity, it can improve the attention ability of the model to fine-grained pixels. Therefore, PSA has achieved better results in this task.

3.3.3 Fusion Strategy Analysis

In image fusion task, this paper proposes an image fusion strategy based on edge feature map. Without using mathematical morphology processing, it is compared with the SF(spatial frequency) fusion strategy used by SESF, and the results are shown in Figure 7.



Figure 7: Comparison of fusion strategies

It can be seen from Figure 7 that the fusion strategy based on edge feature map proposed in this paper is used for image fusion task. In the process of pixel-level clear pixel discrimination, the discrimination accuracy is higher. The reason is that the discrimination method based on spatial frequency is mainly based on the gradient information of the encoded feature map. For feature maps with noise interference, it often leads to inaccurate discrimination. The discrimination method based on edge feature map is preprocessed by Gaussian filtering and guided filtering, which reduces the noise information of the original image and increases the gradient information of the original image to a certain extent, so this method has achieved good discrimination effect. Overall, the fusion strategy proposed in this paper is effective.

4 Conclusion

This paper uses DenseNet and polarization self-attention module for unsupervised learning tasks of image reconstruction ; the fusion strategy is designed, and a fusion strategy based on edge feature map is proposed for image fusion task. In the subjective and objective evaluation compared with a variety of algorithms, achieved priority fusion performance. However, it is also found that the use of edge feature map information for discrimination also brings an increase in fusion time and there is still a certain discrimination error. How to further improve the rate and performance of image fusion is our future research direction. In addition, due to the blurring of the focus boundary of the original image, the use of decision graph for fusion will cause a certain gradient dispersion phenomenon in the decision boundary of the fused image. How to further improve the feature extraction performance and reduce the occurrence of gradient dispersion phenomenon needs further study.

5 Acknowledgements

This work was supported by the Process Evaluation of Online Virtual(62067006), the Natural Science Fund Project of Gansu Province(No.21JR7RA300) and the Center for Open project of Gansu Provincial Research (Dunhuang Heritage Conservation Cultural of (No.GDW2021YB15)).

6 Data Availability Statement

The training sets that support the findings of this study are available at <https://cocodataset.org/#download>. The test sets that support the findings of this study are available at <https://dsp.etfbl.net/mif/> and <https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset>.

7 Conflict of Interest Statement

The author's declared that they have no conflicts of interest to this work.

References

- [1] Wu K, Mei Y. Multi-focus image fusion based on unsupervised learning[J]. Machine Vision and Applications, 2022, 33(5): 1-14.
- [2] Zhang H, Xu H, Tian X, et al. Image fusion meets deep learning: A survey and perspective[J]. Information Fusion, 2021, 76: 323-336.
- [3] Xu H, Ma J, Jiang J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 502-518.

- [4] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network[J]. *Information Fusion*, 2017, 36: 191-207.
- [5] Ram Prabhakar K, Sai Srikar V, Venkatesh Babu R. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs[C]//*Proceedings of the IEEE international conference on computer vision*. 2017: 4714-4722.
- [6] Hui Li, Xiao-Jun Wu. DenseFuse: A Fusion Approach to Infrared and Visible Images.[J]. *IEEE Trans. Image Processing*, 2019, 28(5):
- [7] Boyuan Ma, Xiaojuan Ban, Haiyou Huang, Yu Zhu. SESF-Fuse: An Unsupervised Deep Model for Multi-Focus Image Fusion.[J]. *CoRR*, 2019, abs/1908.01703:
- [8] Zhang H, Ma J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion[J]. *International Journal of Computer Vision*, 2021, 129(10): 2761-2785.
- [9] H. Jung, Y. Kim, H. Jang, N. Ha and K. Sohn, "Unsupervised Deep Image Fusion With Structure Tensor Representations," in *IEEE Transactions on Image Processing*, vol. 29, pp. 3845-3858, 2020, doi: 10.1109/TIP.2020.2966075.
- [10] Zhang H, Le Z, Shao Z, et al. MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion[J]. *Information Fusion*, 2021, 66: 40-53.
- [11] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 4700-4708.
- [12] Liu H, Liu F, Fan X, et al. Polarized self-attention: towards high-quality pixel-wise regression[J]. *arXiv preprint arXiv:2107.00782*, 2021.
- [13] Qu L, Liu S, Wang M, et al. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. 2022, 36(2): 2126-2134.
- [14] Hou R, Zhou D, Nie R, et al. VIF-Net: an unsupervised framework for infrared and visible image fusion[J]. *IEEE Transactions on Computational Imaging*, 2020, 6: 640-651.
- [15] Li Yufeng, Chen Niannian, Zhang Jiacheng. A fast and high sensitivity focusing evaluation function [J].*Computer application research*, 2010, 27 (04) : 1534-1536.
- [16] Qu G, Zhang D, Yan P. Information measure for performance of image fusion[J]. *Electronics letters*, 2002, 38(7): 1.

- [17] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.
- [18] Petrović V. Subjective tests for image fusion evaluation and objective metric validation[J]. Information fusion, 2007, 8(2): 208-216.
- [19] Yu S, Zhongdong W, Xiaopeng W, et al. Tetrole transform images fusion algorithm based on fuzzy operator[J]. Journal of Frontiers of Computer Science and Technology, 2015, 9(9): 1132-1138.
- [20] Shreyamsha Kumar B K. Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform[J]. Signal, Image and Video Processing, 2013, 7(6): 1125-1143.
- [21] Hossny M, Nahavandi S, Creighton D. Comments on 'Information measure for performance of image fusion'[J]. Electronics letters, 2008, 44(18): 1066-1067.
- [22] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. Electronics letters, 2000, 36(4): 308-309.
- [23] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.
- [24] Nejati M, Samavi S, Shirani S. Multi-focus image fusion using dictionary-based sparse representation[J]. Information Fusion, 2015, 25: 72-84.
- [25] Savić S, Babić Z. Multifocus image fusion based on empirical mode decomposition[C]//Proc. IWSSIP. 2012: 1-4.
- [26] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
- [27] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
- [28] Cao Y, Xu J, Lin S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.