AN IMAGE FUSION ALGORITHM USING GLOBAL-LOCAL FEATURE AGGREGATION AND ENHANCEMENT

Bin YANG[1], Qingchun ZHENG[1], Peihao ZHU[1,2]

This study presents GLAFusion, a hybrid Restormer-CNN image fusion technique for multi-exposure and infrared-visible image fusion. To efficiently perceive global information from the channel aspect, the Restormer module is employed in our algorithm, while our algorithm can capture multi-scale local features through a fully connected attention network. Ultimately, the feature aggregation and improvement module that was created acquires the fused image. The superiority of the proposed GLAFusion is confirmed by designed ablation experiments and comparison studies with other cutting-edge algorithms.

Keywords: Restormer, image fusion, multimodal, CNN.

1. Introduction

To create comprehensive images with multi-modal image composite characteristics is the aim of image fusion. This technology is widely employed in the military, industrial, and surveillance domains. Following years of study and development, there are two main categories for image fusion: deep learning approaches [1][2][3] and classical methods [4][5]. Traditional methods are superior in computational speed, however, their modelling techniques are intricate and necessitate the manual creation of fusion and feature extraction procedures. This drawback can be avoided with the advancement of artificial intelligence with the use of deep learning in the field. Most deep learning algorithms use convolution to extract useful information. Although convolution can effectively extract local details, it is not good at sensing global dependencies. Therefore, the current advanced algorithms use Transformer [6] to realize the fusion task and achieve excellent results. High computational complexity is a drawback of Transformer, current approaches are unable to effectively sense global information while also utilizing multi-scale information.

We propose a multimodal image fusion method based on global-local feature aggregation and enhancement which is named GLAFusion to address the above challenges, which can be implemented for infrared-visible and multi-exposure image fusion tasks. To efficiently perceive the long-range dependencies of multimodal images, we employ a variant of Transformer, the Restormer module [7], to address this problem. Restormer, which was first used for high-resolution image restoration, uses multi-Dconv head transposed attention instead of the self-

[1] Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control, School of Mechanical Engineering, Tianjin University of Technology, Tianjin, 300384, China.
[2] Corresponding author: Peihao Zhu, e-mail: zhupeihao_gp@163.com

attention mechanism based on sliding window strategy, by computing cross-channel cross covariance to obtain the attention features, which largely reduces the computational complexity. We use it for the multimodal image fusion missions. We design a fully connected attention network for extracting the local details of the multimodal images. A modified parallel-structured convolution block attention module (CBAM) [8] is embedded in this module, and we employ a multi-scale convolutional kernel in the spatial attention path to further perceive the local features. In addition, at the end of the model we design a feature aggregation enhancement module (FAEM) for the fusion of local and global information. If our GLAFusion reduces any of these components, it will lead to a degraded fusion quality. Overall, the key contributions of this work are elaborated below:

(1) GLAFusion, a new image fusion algorithm is proposed. Restormer in the algorithm is used to construct long-distance dependencies, and fully connected attention network is designed to extract multi-scale local details. By combining the properties of both, our algorithm can retain both global and local information.
(2) We designed a feature enhancement module for effective integration and enhancement of global information and multi-scale local details simultaneously. In addition, we design suitable intensity loss functions for infrared-visible and multi-exposure image fusion tasks, max and mean-max, respectively, and experimentally prove that the method can lead to better fusion quality.
(3) Our algorithm can be implemented for infrared-visible and multi-exposure image fusion tasks. The superiority of the proposed GLAFusion is confirmed by designed ablation experiments and comparison studies with other cutting-edge algorithms.

2. Related work

Deep learning models with excellent learning and data processing capabilities are suitable for processing text, images, and other tasks, which greatly advance the development process of multimodal image fusion technology. In 2018, Liu et al. [9] utilized the designed convolutional networks to obtain weight maps with pixel activity information of the multimodal source images to avoid artificial activity level measurement and the assignment of weight. In 2018, Ma et al. [10] performed the first application of GAN to image fusion task. Since FusionGAN contains only one discriminator, which cannot retain the information of multimodal images more comprehensively, Ma et al. [11] again proposed a dual discriminator GAN network in 2020, which takes the images of two modalities as inputs to two discriminators respectively. In 2019, Li et al. [12] used dense connection to construct encoder network to extract information from multimodal source images well. In 2020, this research group again proposed the NestFuse model [1] to extract source image features from multi-scale aspect and use spatial and channel attention fusion strategies instead of average weighting operations for better fusion of deep

features. In 2020, Zhang et al. [13] proposed an approach named PMGI, it established gradient as well as intensity paths based on the input ratio difference of the multimodal images, while incorporating an information interaction module directly into these paths to minimize information loss. In 2022, Xu et al. [3] used an information measurement strategy in an image fusion model to obtain the adaptive information retention of multimodal images, which performed well in several image fusion tasks. The above deep learning-based model extracts the local details of the source image through convolutional operations, but it is not sufficient for the perception of global information.

In 2021, V. S. et al. [14] proposed IFT model to utilize Transformer for image fusion task to solve the issue that convolution is not able to notice global information. Qu et al. [15] proposed TransMEF model in 2022 to perform multi-task learning in a self-supervised manner depending on the source image characteristics to optimize the feature extraction ability of their method. In 2022, Wang et al. [16] used pure Transformer to process infrared-visible image fusion task, they employed an AE-based architecture including global feature extraction, L1-Norm based fusion strategy, and feature reconstruction. Moreover, Rao et al. [17] utilized Transformer to construct GAN network, their TGFuse is based on optical weights and adversarial learning to fuse infrared-visible images. Most of the fusion methods using Transformer are affected by the computational complexity of Transformer, resulting in lower efficiency, and they are unable to effectively receive global information while also utilizing multi-scale information to its fullest extent.

3. Method

3.1. Overall framework

To make the fused images present both detailed and global information better and to get a better balance between fusion quality and efficiency, we devise a global-local feature aggregation and enhancement multimodal image fusion method called GLAFusion. Our algorithm has three key components: global feature extraction, fully connected attention network, and FAEM. The total flow of our GLAFusion is shown in Fig. 1. In our GLAFusion, two images I_1 and I_2 with different modalities are concatenated channel-by-channel as input, and infrared-visible images are taken as examples in Fig. 1. The initial shallow features are first obtained by a 3×3 convolutional layer, and then to acquire the global information and local traits of the input images. GLAFusion is divided into two branches: global feature extraction and fully connected attention network. We select the Restormer block named RB in Fig. 1 to extract global features quickly and efficiently. Restormer was proposed by Zamir et al. [7] for high-resolution image restoration, who used multi-Dconv head transposition attention instead of multi-head attention in the Transformer without local windowing to more quickly capturing global

features from the channel aspect. The purpose of our fully connected attention network is to acquire the local detailed features adequately. After the multi-scale decomposition, the detailed feature extraction module DFEM can better perceive the important information and general information from feature maps of different scales respectively and assign different weights. The full-scale skip connection can also minimize information loss. In addition, to achieve global and local information fusion and enhancement, we design the feature aggregation and enhancement module FAEM, and finally we can get the output image I_f through a 1×1 convolutional layer.

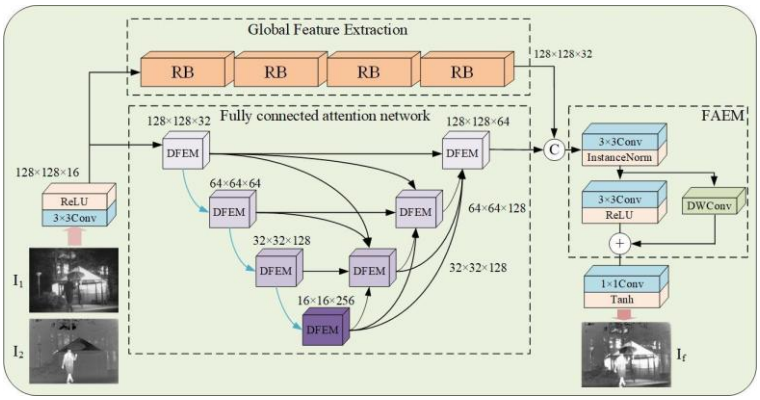

Fig. 1. The overall structure of GLAFusion

3.2. Global feature extraction path

In the exploration process, concatenating more RB does not significantly enhance the fusion effect of the algorithm, but will have a negative impact on the fusion efficiency due to the load. In order to balance the quality of fusion results and time consumption, we chained four RBs in GLAFusion. Fig. 2 illustrates the RB's particular design. The multi-Dconv head transposed attention and gated-Dconv feed-forward network make up the majority of the RB. The former is in charge of detecting global information across channels, while the latter is in charge of refining useful information.

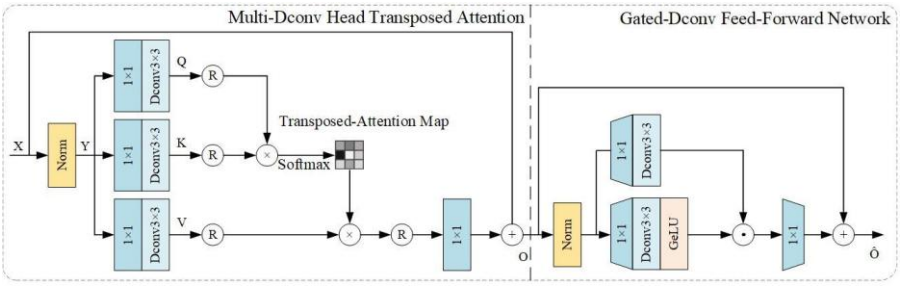

Fig. 2. The structure of RB

3.3. Fully connected attention network

To roundly capture the local traits in multimodal images, we design a multiscale fully connected attention network, which mainly consists of skip connections and attention mechanism, its structure can be understood in Fig. 1. The structure of the DFEM is shown in Fig. 3.

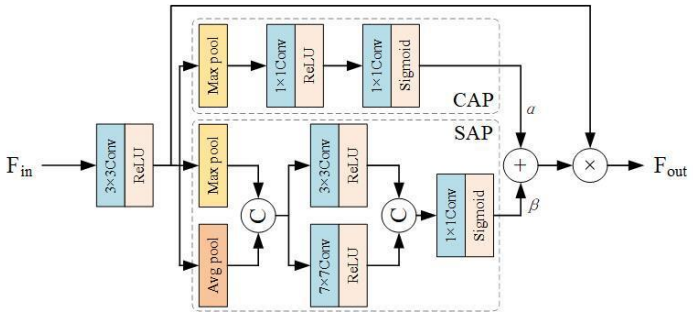

Fig. 3. The structure of DFEM

DFEM starts with a 3×3 convolutional layer. In this part, we employ ReLU activation function. Then we first obtain four feature maps with different sizes by downsampling three times. After each downsampling, the feature maps' width and height both decrease to half of the processed counterparts, and then the local features are perceived by the designed attention mechanism. The multiscale decomposed feature maps are used as inputs to the attention mechanism, which is split into a spatial attention path SAP and a channel attention path CAP, and these two have a parallel structure. The attention weights of channels and regions are obtained respectively, and ultimately multiplies the input features pixel-by-pixel by a residual connection to obtain the multiscale attention feature maps. In particular, to extract local information more comprehensively during the processing of features at different scales, we set two different sizes of convolution kernels in SAP, 7×7 and 3×3, respectively. α and β are both set to 0.5 in this paper, and they are weighting parameters of the channel and spatial attention.

3.4. Feature aggregation and enhancement module

In order to fuse the extracted global and local information in more detail, we design a feature aggregation and enhancement module FAEM, and the detailed structure can be seen from Fig. 1. FAEM is mainly composed of two layers of ordinary 3×3 convolution and a layer of depth-wise convolution with skip connection. After the first convolution layer, we adopt Instancenorm operation, which only normalizes each channel of each sample. Compared with Batchnorm, it is computationally simple and facilitates more effective retention of feature information. Depth-wise convolutional skip connections are used to emphasize feature details.

3.5. Loss functions

The total loss of GLAFusion can be expressed as:

$$L = \mu L_{\mathrm{i}} + \nu L_{\mathrm{g}} + L_{\mathrm{s}} \tag{1}$$

$$L_{\mathrm{i}} = \left\| I_{\mathrm{f}} - \mathrm{M}\left(I_1, I_2\right) \right\|_1 / HW \tag{2}$$

$$L_{\mathrm{g}} = \left\| \left|\nabla I_{\mathrm{f}}\right| - \max\left(\left|\nabla I_1\right|, \left|\nabla I_2\right|\right) \right\|_1 / HW \tag{3}$$

$$L_{\mathrm{s}} = \theta\left(1 - \mathrm{SSIM}\left(I_{\mathrm{f}}, I_1\right)\right) + \sigma\left(1 - \mathrm{SSIM}\left(I_{\mathrm{f}}, I_2\right)\right) \tag{4}$$

Three components comprise total loss L, strength loss L_{i}, structure loss L_{s} and gradient loss L_{g}. The weighting parameters that control L_{i} and L_{g} are μ and ν. $\|\cdot\|_1$ denotes L1-norm. M(·) is a pixel-by-pixel operation strategy of two modal images, we adopt the max strategy for the infrared-visible image fusion task, because both images may contain extensive intensity information in different scenarios. And we adopt the mean-max strategy for the multi-exposure image fusion task because most of the luminance information exists in the overexposed images. ∇ denotes the gradient operator, the height and width of the input image are denoted by H and W, SSIM(·) is the structural similarity metric, θ and σ are parameters that control the percentage share of each.

4. Experimental details and analysis of results

4.1. Experimental setup and comparison methods

41 pairs of images from the TNO [18] dataset were used to train the infrared-visible image fusion task during the training phase. The multi-exposure image fusion task was trained using 32 pairs of images from the MEFB [19] dataset. To get enough training data, the chosen training images were cropped into 128×128 image segments with a stride of 24. This resulted in the acquisition of 17002 pairs of multi-exposure image segments and 13366 pairs of infrared-visible image segments. 16 epochs and a batch size of 40 are specified. Using an exponential decay method and Adam optimizer, we optimized our GLAFusion by setting the initial learning rate and decay rate to 0.0004 and 0.9, respectively. The weighting parameters μ and ν are 2 and 5, respectively, and both θ and σ are 0.5. The experimental hardware facilities are NVIDIA GeForce RTX3090 GPU and Intel(R) Xeon(R) W-2245 CPU, the deep learning framework is TensorFlow.

In the comparison experiments, we take six advanced algorithms, NestFuse [1], PMGI [11], SDNet [2], SwinFusion [20], U2Fusion [3] and YDTR [21], as our comparison algorithms. We take standard deviation (SD), average gradient (AG), information entropy (EN), visual information fidelity (VIF), and spatial frequency (SF) as objective evaluation metrics.

4.2. Comparative experiment of infrared-visible image fusion

20 pairs of randomly chosen images from the TNO dataset serve as the test data in this section. The results of the quantitative comparisons made using the TNO dataset, where the data are the average scores of each method, are displayed in Table 1. It is evident that GLAFusion gets the best results across four assessment metrics: SD, AG, EN, and SF. In two metrics, SD and SF, our GLAFusion performs much better than the other algorithms. In VIF, SwinFusion obtains the suboptimal value whereas NestFuse obtains the optimum value, GLAFusion shows a flaw, but also got a score of 0.7179, just after SwinFusion.

Table 1

Quantitative comparison experiments on the TNO dataset

	NestFuse	PMGI	SDNet	SwinFusion	U2Fusion	YDTR	Ours
SD	41.6492	38.3224	33.1786	39.6015	36.1611	28.2380	57.7209
AG	3.5720	3.2136	4.1168	3.7013	4.4322	2.5831	4.7753
EN	7.0879	7.0328	6.6993	6.8916	6.9743	6.4909	7.4490
SF	9.1424	7.8971	10.3806	9.4779	10.3999	6.9900	12.5749
VIF	0.8967	0.6443	0.5852	0.7770	0.6236	0.6327	0.7179

Fig. 4 displays the qualitative comparison findings of GLAFusion and other algorithms on the TNO dataset.

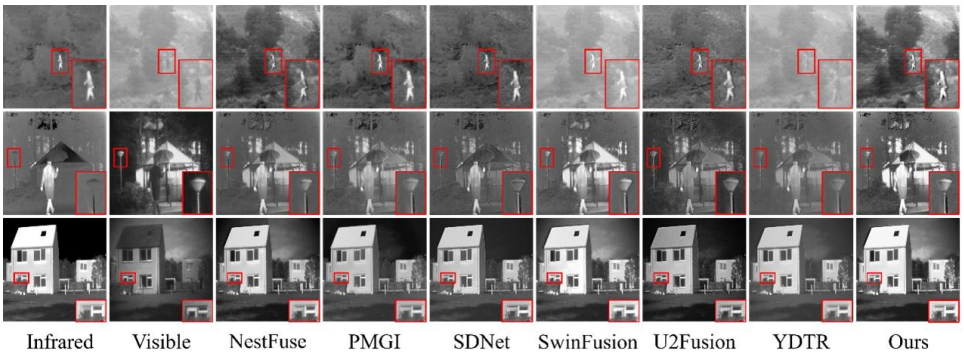

Fig. 4. Qualitative comparison on the TNO dataset

Both SwinFusion and YDTR effectively preserve the structural information, although some scenes lack contrast details. PMGI retains certain contrast features of the infrared image but fails to provide clear target edges. NestFuse, SDNet, U2Fusion, and our proposed GLAFusion all show the capacity for extracting texture and contrast information in multimodal images. Furthermore, NestFuse and GLAFusion exhibit superior scene representation capabilities.

4.3. Comparative experiment of multi-exposure image fusion

The quantitative comparison experiments carried out on the MEFB dataset are displayed in Table 2. GLAFusion consistently gets the highest value on the metrics of SD, AG, EN, and SF. Notably, in terms of SF, GLAFusion outperforms SwinFusion by a margin of 30.36%. Regarding VIF, our algorithm obtains the

suboptimal value of 1.2381 while NestFuse remains at the top with a score of 1.3779.

Table 2

Quantitative comparison experiments on the MEFB dataset

	NestFuse	PMGI	SDNet	SwinFusion	U2Fusion	YDTR	Ours
SD	67.8103	57.1727	55.5664	67.7289	59.4382	59.4509	70.3737
AG	4.4711	4.7873	4.9596	5.5490	4.7279	3.8850	6.5639
EN	7.5038	7.3944	7.4339	7.3435	7.3952	7.2271	7.6519
SF	15.5704	15.4876	16.7919	17.5602	14.8902	14.1629	22.8929
VIF	1.3779	0.9746	1.1915	1.1865	1.2149	1.2030	1.2381

We select three representative fused images from the test data in the MEFB dataset to demonstrate qualitative comparisons. Fig. 5 reveals that the overexposed image's brightness information is efficiently preserved by NestFuse and YDTR, but the underexposed image's textural details are less noticeable. For instance, the lake surface features in the second row of Fig. 5 are not sufficiently displayed by the fusion findings of NestFuse and YDTR.

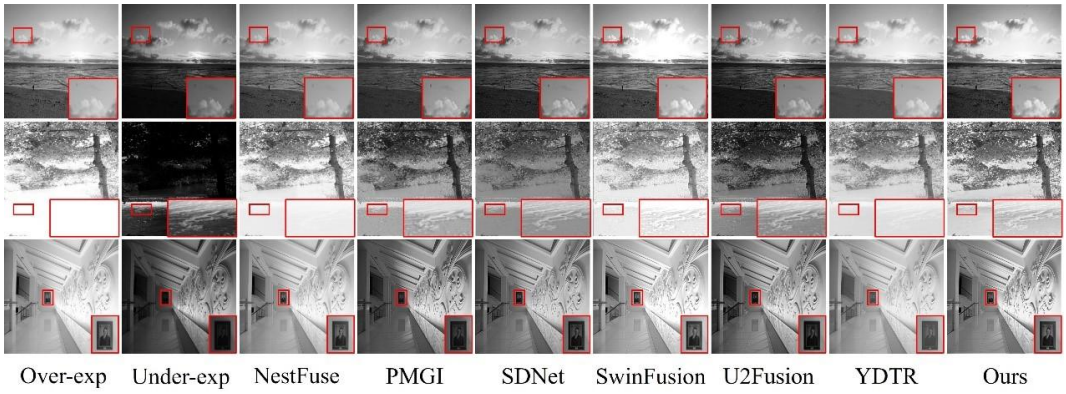

Fig. 5. Qualitative comparison on the MEFB dataset

Other algorithms successfully retain a significant amount of edge information. However, SwinFusion exhibits a slight lack of contrast in its fusion result, and U2Fusion fails to capture sufficient brightness information from the overexposed image, resulting in a slightly darker fusion effect. PMGI, SDNet, and GLAFusion appropriately preserve both brightness information and rich texture details.

4.4. Ablation analyses

GLAFusion is mainly composed of a global feature extraction path, a fully connected attention network, and FAEM, which enables our model to better perceive and fuse global and local information. To verify the role of each part, we design four different model structures and conduct comparative experiments on the TNO dataset and MEFB dataset: (1) There is no FAEM in the model. (2) There is

no global feature extraction path in the model. (3) There is no fully connected attention network in the model. (4) Ours.

Table 3

Quantitative comparison of methods with diverse network structures

	TNO				MEFB			
	(1)	(2)	(3)	Ours	(1)	(2)	(3)	Ours
SD	44.9376	47.8127	58.6625	57.7209	57.0489	60.6821	62.2851	70.3737
AG	3.6841	3.9630	4.7502	4.7753	5.2787	5.3473	6.2517	6.5639
EN	7.1892	7.2550	7.3527	7.4490	7.4761	7.5352	7.5640	7.6519
SF	9.8753	10.5106	12.2652	12.5749	18.1982	18.4532	21.5420	22.8929
VIF	0.7773	0.7594	0.6830	0.7179	1.3080	1.2481	1.0610	1.2381

The findings of the quantitative comparison of the TNO and MEFB datasets are displayed in Table 3. GLAFusion obtains the suboptimal value on SD and the optimal performance on AG, EN, and SF on the TNO dataset. GLAFusion gets the best outcomes on the SD, AG, EN, and SF metrics on the MEFB dataset. From the perspective of visual effects, these models have achieved good effects, but there are still differences in the details. We can see that in the first row of fusion results in Fig. 6, the contrast of the person in our GLAFusion is more prominent.

Fig. 6. Qualitative comparison of methods with different network structures

We use different intensity loss functions L_i for the infrared-visible and the multi-exposure image fusion tasks, for the former we adopt the pixel maximum value preservation strategy(max) and for the latter we adopt the joint action strategy of pixel maximum preservation and pixel averaging(mean-max). We designed three different sets of intensity loss function strategies to verify why we do this: (a): mean-max, (b): mean, (c): max. In Table 4, (c) algorithm using the max strategy yields the optimal performance on SD, VIF and EN on the TNO dataset. However, (a) algorithm using the mean-max strategy obtains the optimal value on the three metrics EN, SF and VIF on the multi-exposure dataset MEFB.

Table 4

Quantitative comparison of methods with different L_i

	TNO			MEFB		
	mean-max	mean	max	max	mean	mean-max
SD	39.5179	14.5535	57.7209	78.6036	20.7455	70.3737

AG	4.8957	5.3499	4.7753	5.8443	6.7421	6.5639
EN	7.0240	5.7582	7.4490	7.1321	6.2118	7.6519
SF	12.7957	14.3901	12.5749	21.2367	22.7846	22.8929
VIF	0.6246	0.4683	0.7179	1.0844	0.7567	1.2381

When L_i with max strategy is used to train the model for infrared-visible image fusion (refer to the first row of Fig. 7), the fusion results can better emphasize the target information, however the results obtained with mean strategy lose too much contrast information. For multi-exposure image fusion (see the second line of Fig. 7), (a) algorithm using max strategy results in excessive brightness information and loses part of the texture features. The results of multi-exposure image fusion using mean strategy yields the same results as infrared-visible image fusion, too much contrast information is lost, and the target cannot be highlighted. The results obtained using the mean-max strategy can properly preserve texture and brightness information. The above ablation analysis validates the criticality of the components of GLAFusion. And we can conclude that L_i using max strategy is more appropriately used for training infrared-visible image fusion tasks, while L_i using mean-max strategy is more appropriately used for training multi-exposure image fusion tasks.

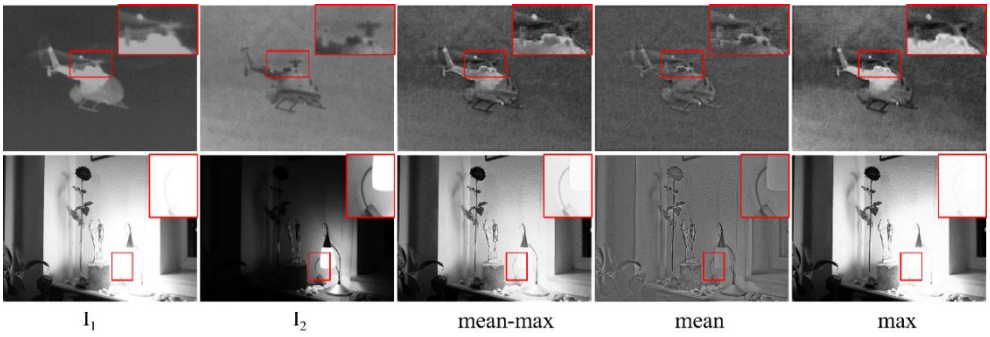

Fig. 7. Qualitative comparison of methods with different L_i

In GLAFusion, both parameters θ and σ of SSIM were set to 0.5 and we performed ablation experiments on the TNO dataset. We set up three sets of experiments, case1: $\theta = 0.3$, $\sigma = 0.7$**,** case2: $\theta = 0.7$, $\sigma = 0.3$**,** Ours: $\theta = 0.5$, $\sigma = 0.5$. The fusion analysis results of the algorithms for the three cases are shown in Table 5.

Table 5

Ablation experiment on parameter setting of SSIM loss function

	Case1	Case2	Ours
SD	55.8469	55.4886	57.7209
AG	5.2320	4.2974	4.7753
EN	7.3630	7.3933	7.4490
SF	14.0443	11.1831	12.5749
VIF	0.7164	0.7039	0.7179

From Table 5, it can be seen that when both θ and σ are set to 0.5, the fusion result achieves the optimal values in all the three metrics of SD, EN, and VIF, and the overall performance is the best.

4.5. Efficiency analysis

We compared the running efficiency of GLAFusion with other advanced algorithms on infrared-visible image datasets and multi-exposure image datasets. All algorithms were run on CPU. The fusion time-consuming analysis is presented in Table 6. We can see that the running efficiency of SDNet dominates on any of the datasets. Compared to other algorithms, SwinFusion runs worse in terms of efficiency. In terms of computational efficiency, our algorithm also does not perform well. However, experiments prove that our GLAFusion performs better in balancing fusion quality and efficiency.

Table 6

Efficiency comparison of our GLAFusion with other advanced algorithms (unit: s)

	NestFuse	PMGI	SDNet	SwinFusion	U2Fusion	YDTR	GLAFusion
TNO	5.0768	0.6543	0.1403	20.9490	1.1673	2.4646	3.8846
MEFB	8.6980	1.1622	0.2477	40.0986	2.1806	13.9933	6.0192

5. Conclusions

We propose a global-local feature aggregation and enhancement method for multimodal image fusion in this paper, and it can be applied to handle infrared-visible image fusion tasks and multi-exposure image fusion tasks. We construct global feature extraction paths through an efficient Restormer module to perceive global information at the channel level. Then we utilize a fully connected attention network to extract local detailed features from the multi-scale feature maps. Moreover, in aiming at the fusion and enhancement of global and local information, we design a simple and effective feature aggregation and enhancement module (FAEM). In quantitative and qualitative comparison experiments with six other advanced algorithms on multiple datasets, our algorithm demonstrates certain advantages and generalization capabilities. Network structure ablation experiments validate the potency of each module. Loss function ablation analyses validate that the intensity loss function using the max strategy in our approach is more appropriately used for training infrared-visible image fusion tasks, while the intensity loss function using the mean-max loss function strategy is more appropriately used for training multi-exposure image fusion tasks. Although our approach has made some progress in balancing fusion quality and efficiency, in future work, we will conduct further research on how to improve the efficiency of the algorithm.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (62073239).

REFERENCES

[1]. *H. Li, X. J. Wu, T. Durrani*. "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models", IEEE Transactions on Instrumentation and Measurement, **vol. 69**, no. 12, 2020, pp. 9645-9656.
[2]. *H. Zhang, J. Ma*. "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion", International Journal of Computer Vision, **vol. 129**, no. 10, 2021, pp. 2761-2785.
[3]. *H. Xu, J. Ma, J. Jiang, et al*. "U2Fusion: A unified unsupervised image fusion network", IEEE Transactions on Pattern Analysis and Machine Intelligence, **vol. 44**, no. 1, 2020, pp. 502-518.
[4]. *H. Li, K. Ma, H. Yong, et al*. Fast multi-scale structural patch decomposition for multi-exposure image fusion. IEEE Transactions on Image Processing, **vol. 29**, 2020, pp. 5805-5816.
[5]. *L. Jian, R. Rayhana, L. Ma, et al*. "Infrared and visible image fusion based on deep decomposition network and saliency analysis", IEEE Transactions on Multimedia, **vol. 24**, 2021, pp. 3314-3326.
[6]. *A. Vaswani, N. Shazeer, N. Parmar, et al*. "Attention is all you need", Advances in Neural Information Processing Systems, 2017, 30.
[7]. *S. W. Zamir, A. Arora, S. Khan, et al*. "Restormer: Efficient transformer for high-resolution image restoration", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022, pp. 5728-5739.
[8]. *S. Woo, J. Park, J. Y. Lee, et al*. "Cbam: Convolutional block attention module", Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.
[9]. *Y. Liu, X. Chen, J. Cheng, et al*. "Infrared and visible image fusion with convolutional neural networks", International Journal of Wavelets, Multiresolution and Information Processing, **vol. 16**, no. 03, 2018, pp. 1850018.
[10]. *J. Ma, W. Yu*, *P. Liang, et al*. "FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion", Information Fusion, **vol. 48**, 2019, pp. 11-26.
[11]. *J. Ma, H. Xu, J. Jiang, et al*. "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion", IEEE Transactions on Image Processing, **vol. 29**, 2020, pp. 4980-4995.
[12]. *H. Li, X. Wu*. "DenseFuse: A fusion approach to infrared and visible images", IEEE Transactions on Image Processing, **vol. 28**, no. 5, 2018, pp. 2614-2623.
[13]. *H. Zhang, H. Xu, Y. Xiao, et al*. "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity", Proceedings of the AAAI conference on artificial intelligence, **vol. 34**, no. 07, 2020, pp. 12797-12804.
[14]. *V. Vs, J. M. Jose Valanarasu, P. Oza, et al*. "Image fusion transformer", 2022 IEEE International conference on image processing (ICIP). IEEE, 2022, pp. 3566-3570.
[15]. *L. Qu, S. Liu, M. Wang, et al*. "Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning", Proceedings of the AAAI conference on artificial intelligence, **vol. 36**, no. 2, 2022, pp. 2126-2134.
[16]. *Z. Wang, Y. Chen, W. Shao, et al*. "SwinFuse: A residual swin transformer fusion network for infrared and visible images", IEEE Transactions on Instrumentation and Measurement, **vol. 71**, 2022, pp. 1-12.
[17]. *D. Rao, T. Xu, X. Wu*. "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network", IEEE Transactions on Image Processing, 2023.
[18]. *A. Toet*. "The TNO multiband image data collection", Data in brief, **vol. 15**, 2017, pp. 249-251.
[19]. *X. Zhang*, "Benchmarking and comparing multi-exposure image fusion algorithms", Information Fusion, **vol. 74**, 2021, pp. 111-131.
[20]. *J. Ma, L. Tang, F. Fan, et al*. "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer", IEEE/CAA Journal of Automatica Sinica, **vol. 9**, no. 7, 2022, pp. 1200-1217.
[21]. *W. Tang, F. He, Y. Liu*. "YDTR: Infrared and visible image fusion via Y-shape dynamic transformer", IEEE Transactions on Multimedia, **vol. 25**, 2022, pp. 5413-5428.