Signal, Image and Video Processing (2024) 18:295–303
https://doi.org/10.1007/s11760-023-02748-z

ORIGINAL PAPER

An end-to-end based on semantic region guidance for infrared and visible image fusion

Guijin Han[1] · Xinyuan Zhang[1] · Ya Huang[1]

Received: 9 July 2023 / Revised: 11 August 2023 / Accepted: 14 August 2023 / Published online: 5 September 2023

Abstract
The goal of infrared and visible image fusion is to fuse the dominant regions in the images of the two modalities to generate high-quality fused image. However, existing methods still suffer from some shortcomings, such as lack of effective supervision information, slow computation due to complex fusion rules, and difficult convergence of GAN-based models. In this paper, we propose an end-to-end fusion method based on semantic region guidance. Our model contains three basic parts: preprocessing module, image generation module, and semantic guided information quantity discrimination module (IQDM). Firstly, we input the infrared and visible images into the preprocessing module to achieve the preliminary fusion of the images. Subsequently, the features are fed into the image generation module for high-quality fused image generation. Finally, the training of the model was supervised by the semantic guided IQDM. In particular, we improve the image generation module based on the diffusion model, which effectively avoids the design of complex fusion rules and makes it more suitable for image fusion tasks. We conduct objective and subjective experiments on four public datasets. Compared with existing methods, the fusion results of the proposed method have better objective metrics, contain more detailed information.

Keywords Visible and infrared image · Image fusion · Diffusion models · Semantic guided · Generative network

1 Introduction

The purpose of image fusion is to combine images in different modes to generate a fusion image with the advantages of the input image. The visible image has the advantages of high resolution, high quality and rich image texture detail information. However, the image quality of the visible image is easily affected by lighting conditions such as no light or low light, and environmental factors such as object occlusion and camouflage. Infrared image has better contour information and global information of the object, which can effectively make up for the shortage of visible image. However, infrared images still suffers from low image contrast and quality, inadequate expression of texture and detail information, as well as susceptibility to noise. Therefore, the fusion of infrared and visible images can effectively overcome the limitations of single sensors, compensate for scene information, and provide richer information and stronger robustness for advanced vision tasks such as object detection, object tracking, and semantic segmentation.

In recent years, with the extensive use of residual blocks [1] and dense connections [2], DenseFuse [3] is the first work to apply deep learning to image fusion. Subsequently, RFN-Net [4] proposed a two-stage training method to achieve full learnability of the model. With the extensive application of the attention mechanism in the field of computer vision, PIA-Fuse [5] combines the cross-modal differential perception fusion module with the semi-fusion strategy, and designs an attention module based on information difference. SeaFusion [6] is the first model that uses high-level semantic information to drive image fusion, which concatenates the image segmentation task and the image fusion task, effectively enhances the fusion network ability to describe spatial details. Subsequently, DID-fuse [7] proposed a deep decomposition model, which uses foreground and background information to assist

✉ Xinyuan Zhang
xiyouzxy@stu.xupt.edu.cn

Guijin Han
hanguijin@xupt.edu.cn

Ya Huang
xiyouhy@stu.xupt.edu.cn

[1] School of Automation, Xi'an University of Posts and Telecommunications, 618 West Chang'an Avenue, Xi'an 710121, Shaanxi, China

the fusion network for the performance of fused image in subsequent vision tasks improved.

With the in-depth study of GAN [8], the GAN-based image fusion model has opened up a new method for the field of image fusion. Fusion-GAN [9] is the first one to introduce GAN in image fusion. It compensates for problems such as information loss caused by fusion strategies by generating and adversarial strategies through which the fused images are directly generated by the generator. The recent SDDGAN [10] segments the input image into foreground and background information, and completes the generation of the image through semantic information supervision. TarDAL [11] improved the fusion network of generator and dual discriminator, laying a foundation for subsequent high-level vision tasks. It is worth noting that in the recent work Diffusion [12], the diffusion model is used to realize image fusion for the first time, but the fused image still focuses on the information of the visible image and ignores the information of the infrared image. We summarize the shortcomings of current deep learning-based fusion methods as follows:

(1) Auto-encoder (AE)-based methods still suffer from numerous constraints of traditional methods by manually selecting fusion strategies.
(2) Fully convolutional neural network-based methods generally force the fused image to obtain detailed information from the visible image, while thermal radiation information in the infrared image is obtained only through content loss. It makes the fused and visible images very similar and lacks the information from the infrared image.
(3) In generative models, although VAE-based algorithms can sample quickly, the quality of generated images is low. GAN-based fusion models suffer from shortcomings, such as easy training collapse and lack of interpretability.
(4) Fusion models based on attention mechanisms have too many parameters, are computationally slow, and are difficult to perform real-time image fusion.

To solve the above problems, we propose an infrared and visible image fusion algorithm based on the diffusion model. So that, address the common issue of insufficient ground truth as supervision information in image fusion, this study employs an image segmentation model for performing semantic segmentation on both the input infrared and visible images; at the same time, an information discriminant module is designed to solve the problem of semantic level region screening, obtain the unique features of infrared image, the unique features of visible image and the common features, and realize the comparison of the semantic level information of infrared and visible images. To avoid the problems of high complexity and high computational cost of high-quality fusion rules, we choose a generative network to directly generate the fused image. Aiming at the common problems of the diffusion model and the shortcomings of the current image fusion work based on the diffusion model, the advantage of the proposed model is that the structure of the diffusion model is redesigned, which makes the training simpler and the performance more competitive.

The main contributions of this paper are as follows:

(1) We propose an image fusion method that combines semantic information with the diffusion model. The generative network is guided by the input image to directly generate the fused image, eliminating the need for complex fusion rules.
(2) To solve the problems of slow image generation and complex structure of the current diffusion model, we redesign the structure of the diffusion model. Specifically, we have designed a preprocessing module and a style attention module to shorten the training time of the model and enhance the fine-grained features of the original image.
(3) To break through the difficulty of lacking ground truth in the image fusion task, we propose an information quantity discrimination module (IQDM). The two computer vision tasks were combined, and the semantic level fusion of different modalities of information was used to constrain the model through the comprehensive consideration of multiple evaluation indicators.
(4) To measure the quantity of information contained in an image, we introduce a new evaluation index DEB, and prove that our method is superior to the existing advanced methods through a large number of experiments.

2 Proposed method

In this section, we present the prerequisites for both the partial diffusion model and the SRGFusion model framework. Firstly, a brief review of diffusion models is provided, which includes the forward and backward processes as well as a simple derivation of the loss function. Secondly, we will provide a detailed description of the proposed semantic guided information quantity discrimination module. Then, the overall model structure and the detailed design of some models will be described. Finally, we discuss the design of the loss function.

2.1 Diffusion model

The diffusion model was proposed by [13] and has been widely used for image-to-image and text-to-image generation in the recent work DDPM [14]. The model is trained by predicting the distribution of noise, and image generation is

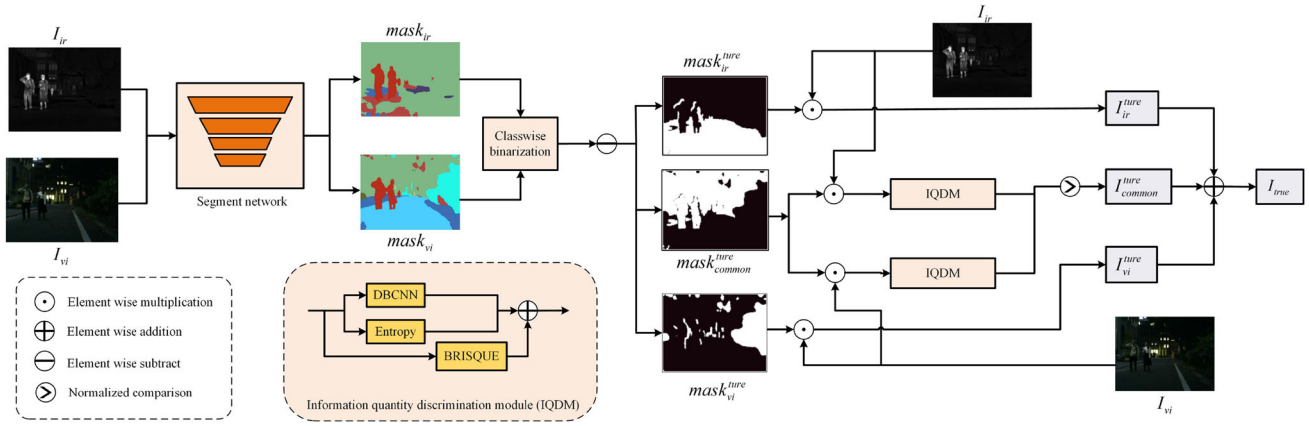

Fig. 1 Flowchart of semantic guided IQDM module

completed through randomly generated Gaussian noise. An image of size $\mathbb{R}^{H\times W\times C}$ represented by a tensor is denoted as I, and $\beta \in (0, 1)$ is a linear or sinusoidal parameter. In the forward process, the noisy image $x_t \in \mathbb{R}^{H\times W\times 3}$ is obtained after the diffusion step $t \in \{0, 1, \ldots, T-1, T\}$ the original image $x_0 \in \mathbb{R}^{H\times W\times 3}$ input to the diffusion model. In the inverse process, the noisy image x_t is input to generate the image $x'_t \in \mathbb{R}^{H\times W\times 3}$.

Forward process: The noise image x_t is generated by gradually adding noise z to the original image x_0 through the Markov chain of order T, which can be expressed as Eq. (1) using the re-parameterization technique.

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}z_t \tag{1}$$

Here, $\overline{\alpha_t} = \prod_{i=0}^{t}\alpha_i$ is a linear distribution, $\alpha_t = 1-\beta_t$ and Z_t represents random noise.

Reverse process: Predicting x_0 directly from x_T is highly unlikely, so we use the Bayesian formula to predict x_T to x_{T-1}, which leads to x_0. Write x_T predicting x_{T-1} in the form of Eq. (2):

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{2}$$

Using $\mathbb{N} \sim (\xi, \delta^2) \propto e^{-\frac{(x-\xi)^2}{2\delta^2}}$, the final relation from x_t to x_{t-1} can be obtained as shown in Eq. (3):

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{1-\overline{\alpha}_t}\varepsilon_\theta(x_t, t)\right) + \sigma_t z \tag{3}$$

Here, ε denotes the neural network and θ denotes the model parameters. In particular, the forward process is a process that does not require learning, and the corresponding β is obtained by randomly selecting the forward step size t. The inverse process generates the image by stepwise derivation.

When training the diffusion model, the model can be constrained by minimizing the difference between the predicted value of the loss function through the neural network and the true value through the forward process, which is specifically expressed in the form of Eq. (4):

$$\min_\theta L_{\text{simple}} = \left\| z_{\text{true}} - \varepsilon_\theta\left(\overline{\alpha}_t x_0 + \sqrt{1-\overline{\alpha}_t}z, t\right)\right\|^2 \tag{4}$$

2.2 Semantic guided IQDM module

We denote infrared image as $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ and visible image as $I_{vi} \in \mathbb{R}^{H\times W\times 3}$, as shown in Fig. 1.

I_{ir} and I_{vi} are input into the segmentation network to obtain the segmentation regions mask_{ir} and mask_{vi} at each semantic level. We finalize binarization process in the two masks separately to obtain $\text{mask}'_{ir} \in \{0, 1\}$ and $\text{mask}'_{vi} \in \{0, 1\}$, the infrared image private mask mask'_{ir}, the visible image private mask mask'_{vi}, and the public mask $\text{mask}_{ir}^{\text{true}}$ are obtained by element-by-element subtraction of different regions under the same category in $\text{mask}_{vi}^{\text{true}}$ and $\text{mask}_{vi}^{\text{true}}$, as shown in Eqs. (5) and (6):

$$\text{mask}'_{ir} - \text{mask}'_{vi} = \text{mask}' \tag{5}$$

$$\text{mask}' = \text{mask}_{ir}^{\text{true}} \cup \text{mask}_{\text{commen}}^{\text{true}} \cup \text{mask}_{vi}^{\text{true}} \tag{6}$$

Here, $\text{mask}_{ir}^{\text{true}}$ corresponds to the part of mask' with element values of 1, which mainly includes the parts that are not clearly visible or present in the visible image, such as occluded or disguised objects or people. The part of mask' with an element value of -1 corresponds to $\text{mask}_{vi}^{\text{true}}$, which mainly includes information such as textures and details that only exist in visible light images. Since $\text{mask}_{\text{commen}}^{\text{true}}$ represents features present in both infrared and visible images, its corresponding value in mask' should be 0. To facilitate subsequent calculations, the three masks are reprocessed so that their internal element values are all 0 or 1.

It is worth mentioning that to make the semantic region guidance approach more general, we have tested it on two image segmentation networks, BisNet [15] and DeeplabV3 [16]. Among them, BisNet is a lightweight image segmentation network with high computational efficiency and short computation time, but poor segmentation accuracy can affect the quality of the final model generated images. DeepLabV3 focuses more on the improvement of segmentation accuracy and provides better supervised information needed by the model, but has a larger computational complexity and a longer computation time.

To effectively distinguish the image quality of common region in infrared and visible image and fuse high-quality features, we proposed an information quantity discrimination module. The $\text{mask}^{\text{true}}_{\text{commen}}$ is calculated by Eqs. (7) and (8) to obtain the public area of I_{ir} and I_{vi} respectively. The I^{common}_{ir} and I^{common}_{vi} are fed into the information volume discrimination module and the final common area features are obtained by comparing the normalized scores.

$$I^{\text{common}}_{ir} = I_{ir} \odot \text{mask}^{\text{true}}_{\text{common}} \tag{7}$$

$$I^{\text{common}}_{vi} = I_{vi} \odot \text{mask}^{\text{true}}_{\text{common}} \tag{8}$$

Here, $\odot$ represents the multiplication of each element. Specifically,I^{common}_{ir} and I^{common}_{vi} are input into IQDM to calculate the three scores, and the larger values in I_{ir} and I_{vi} under the same index are used as the upper bound of normalization, as shown in Eq. (9). By comparing the sum of the scores, the optimal common feature region $I^{\text{true}}_{\text{common}}$ is obtained, which consists of highly informative regions from different semantic regions $I^{\text{true}}_{\text{common}}$.

$$\text{score}_a = \frac{M(I^{\text{common}}_a)}{M(I_a)} \tag{9}$$

Here, $a \in \{ir, vi\}$ and $M()$ represent the scores for calculating the three indicators.

It is worth noting that in the design of the IQDM, we introduce three no-reference image quality assessment methods, namely DB-CNN [17], Entropy, and BRISQUE [18]. The influence of the three methods on the final image quality and whether they are suitable for the evaluation index of the fusion image will be discussed with some details in Sect. 3.3.1

Finally, after the calculation based on the IQDM, we multiply $\text{mask}^{\text{true}}_{vi}$ and $\text{mask}^{\text{true}}_{vi}$ by the corresponding I_{ir} and I_{vi}, respectively, to obtain the corresponding supervised features I^{true}_{ir}, I^{true}_{vi}, and $I^{\text{true}}_{\text{common}}$, and the real value I_{true} for supervised learning can be obtained by combining the three supervised features. For the single channel image region, we replicate the tensor in the channel dimension to achieve visible image computation with three channels.

2.3 General framework

In the image fusion task, we put image I_{ir} and I_{vi} concatenated on the channel dimension, and then input the preprocessing module to obtain x_0, and then realize the image fusion through the forward and reverse process. Among them, I_{ir} and I_{vi} are preliminarily fused through the preprocessing module, and then the preliminary fusion features are synchronously input into the style attention module and the diffusion model to generate the fusion image $I_f \in \mathbb{R}^{H\times W\times 3}$. Secondly, semantic region guidance is used to generate I_{true} for supervised training. Finally, the loss function is used to constrain the training of the network, as shown in Fig. 2.

The preprocessing module and the style attention module are both designed to adapt the diffusion model to the fusion task of infrared and visible images. Among them, the preprocessing module performs a preliminary fusion of input images to shorten the training time for the diffusion model. The style attention module incorporates the features into each layer of the diffusion model, thereby constraining the diffusion model to produce high-quality fused image. In particular, the noise prediction network of the diffusion model is a network structure similar to the U-Net, and its encoder and decoder are in the exact corresponding structure. We input the preprocessed image features into the style attention module to force the constrained diffusion model to generate the fused image, which is conducive to the enhancement of the two different types of features.

2.4 Loss function

We design a loss function based on semantic guidance to better use the existing knowledge to constrain the fusion image, and train the network by minimizing the loss between the input image and the output image, as shown in Eq. (10):

$$L_{\text{total}} = \alpha L_{\text{mse}} + \beta L_{\text{ssim}} + \gamma L_{\text{color}} \tag{10}$$

Here α, β, and γ are all hyperparameters used to balance the three classes of loss functions.For the values of the three hyperparameters, we used the GridSearch method for the search. Specifically, the parameters are tuned sequentially by step size over the specified parameter range. The adjusted parameters are used to test the network and the parameter with the highest accuracy on the validation set is found from all the parameters.

L_{mse} can guide the network to fit each pixel in the image to minimize the difference between the generated image and the true value, which is specifically expressed as Eq. (11):

$$L_{\text{mse}} = \sqrt{\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}(I_f(x,y) - I_{\text{true}}(x,y))} \tag{11}$$

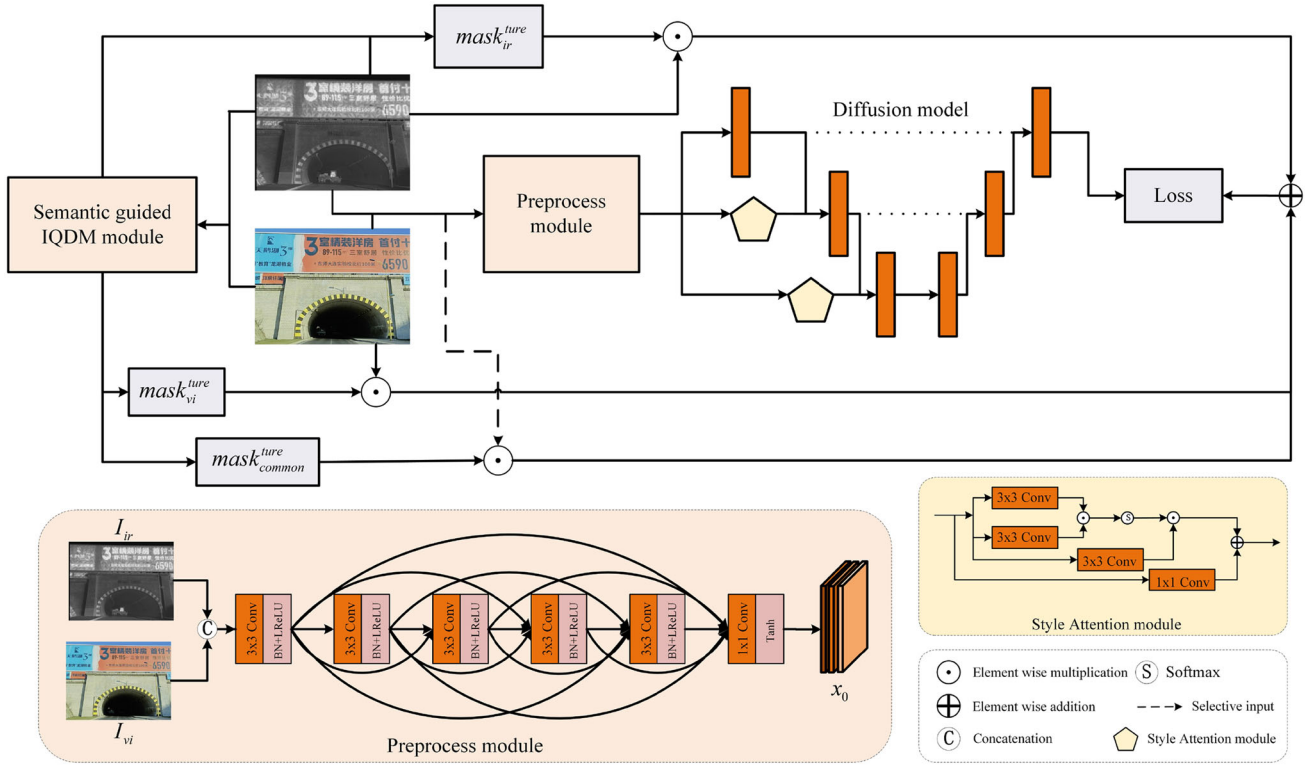

Fig. 2 The framework of the proposed SRGFusion

In order to keep the structure between the supervised image and the generated image as complete as possible, L_{ssim} as shown in Eq. (12) is used to constrain.

$$L_{\text{ssim}} = 1 - \text{SSIM}(I_f, I_{\text{true}}) \tag{12}$$

Since the model uses RGB three-channel visible image to directly generate the fused image, the color similarity loss L_{color} can enhance the color preservation of the fused image. The specific form is shown as Eq. (13):

$$L_{\text{color}} = \frac{1}{HWC} \sum_{i \in \eta} \sum_{k=1}^{K} \angle(I_{vi}^{i}, I_{f}^{i}), \eta \in \{R, G, B\} \tag{13}$$

Here, C represents the number of channels and K represents the number of pixels. $\angle(\cdot, \cdot)$ illustrates the pixel-wise calculation of the discrete cosine similarity between the fused image and the original visible image in the RGB channels. Using L_{color} can better reduce the chroma distortion of the fused image and also capture more scene information.

3 Experiment

In this section, we first introduce the experimental setup, which includes the dataset selection and model training details. Secondly, we present the model's training method and experimental results for each stage. Thirdly, we compare test results and visual fusion images of related advanced algorithms under various evaluation indicators. During the ablation experimental phase, we reveal the effectiveness of each module in our proposed model.

3.1 Experiment details

1) Datasets: We evaluate the model using infrared and visible images contained in the LLVIP [19], M3FD [11], Road Scene [20], and TNO [21] datasets. The model is trained on the LLVIP dataset, where the original dataset consists of 12025 sets of infrared and color visible image pairs and the test dataset consists of 3463 sets of image pairs. It is worth mentioning that in order to prevent gradient explosion all images are resized to size and pixel values are normalized to [0,1] before feeding into the network.

2) Training details: This model is implemented based on the PyTorch framework, using Intel Xeon(R) CPU E5-2620 v4 @ 2.10GHz 32 processors, running on Ubuntu 20.04.2 LTS 64-bit operating system. We performed model training in two stages on four NVIDIA Corporation GP102 GeForce GTX 1080 Ti graphics cards, setting the batch size of a single card to 12 and training the model for 300 epochs. When training the network, Set the number of diffusion step to 200,

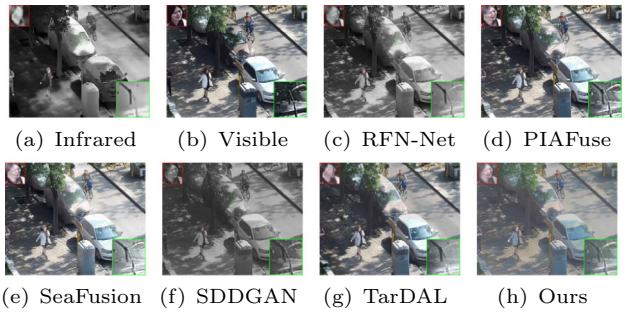

Fig. 3 Fusion results of #260001 in the LLVIP dataset

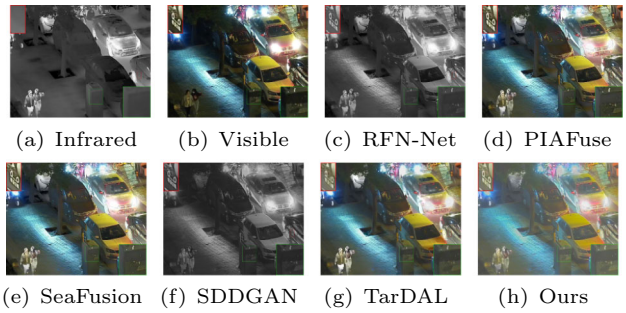

Fig. 4 Fusion results of #230070 in the LLVIP dataset

we have chosen the DeepLabV3 segmentation network as the basis for the semantic region guidance. The Adam optimizer is used to minimize the loss, and the initial learning rate is set to 0.001. In the GridSearch method, we set the interval of hyperparameters to [0, 1], with a step size of 0.1. Finally, the values of α, β, and γ are 0.9, 0.5, and 0.2.

3.2 Performance analysis of fusion

To demonstrate the advantages of our proposed method, we conducted a comprehensive evaluation of fusion performance on four datasets and compared it with the five most recent state-of-the-art methods.

3.2.1 Qualitative results analysis

The LLVIP and M3DF datasets include two types of image pairs: daytime and nighttime. Among them, infrared image highlight extreme heat radiation targets in the scene, while visible image contain further texture information, detailed features, and color information. To more intuitively compare the advantages of our method in preserving source image information, highlighting detail information, and color fidelity, we selected four sets of infrared and visible image pairs from the LLVIP and M3FD datasets for day and night scenes for visual analysis.

Of the five compared methods, RFN-Net is based on the encoder-decoder structure, PIAFuse applies an attention mechanism and illumination guidance module to image fusion, and SeaFusion introduces high-level vision task as supervision information, SDDGAN and TarDAL based on generative networks and their variants.

For demonstrate the advantages of our method more intuitively, in Fig. 3, we show the fusion result for infrared and visible images of #260001 in the LLVIP dataset. In the daytime visible and infrared image fusion, the visible image contains a large amount of information, how to effectively preserve the texture features and common features in the visible image is a research difficulty.

The original texture and color features of human faces cannot be maintained in RFN-Net and TarDAL. Although PIAFuse and Seafusion can retain the detailed information about human faces, they are badly blurred and the fused images are not clearly sufficient. In contrast, our method effectively preserves information and color information. In addition, the prominent feature of the wiper in the green area in the infrared image is the head of the wiper, and the prominent feature in the visible image is the wiper rod. Only our method and SDDGAN can effectively retain the details information of the wiper and its surroundings, and our method retains additional color information.

In Fig. 4, #230070 in the LLVIP dataset are selected to show the fusion results of nighttime images. In Fig. 4, the door handle in the green area is the private feature of the infrared image, and the license plate in the red area is the private feature of the visible image.

In Fig. 4, the door handle in the green area is the private feature of the infrared image, and the license plate in the red area is the private feature of the visible image. In the experimental results, only TarDAL is fuzzy for the red region, while only SDDGAN gives poor results for the green region.

3.2.2 Quantitative results analysis

In order to make a fair comparison with other works, we use six evaluation metrics in our quantitative evaluation. Mutual information (MI) is used to evaluate the aggregation quality of the information of the original image pair in the fused image, visual information fidelity (VIF) is used to evaluate the fidelity of the information in the fused image, spatial frequency (SF) is used to evaluate the spatial frequency related information in the combined data, and Qabf is used to quantify the edge information of the source image. The evaluation metric standard deviation SD is used to evaluate the contrast of fused image, and the metric MS-SSIM is used to evaluate multi-scale structural similarity.

We introduce an evaluation index DEB for quantifying the overall information content of the fused image, to better verify the quality of the image and lay the foundation for

the subsequent advanced vision tasks, which are composed of DB-CNN, Entropy, and BRISQUE. The higher the DEB score, the more information is perceived. Since the image fusion task is a computer vision task lacking effective prior knowledge, and DEB is the image evaluation index under no reference, it can be more effective to verify the quality of the fused image.

We selected 40 sets of infrared and visible image pairs in each of the four datasets for comparison, and show the test results in Table 1.

Our method performs prominently on the LLVIP dataset and achieves the optimum in all indicators. On the remaining two types of color datasets, our method has a large difference in the DEB index compared with alternative methods, which is attributed to the fact that the learning method based on semantic information guidance better retains the feature information in the two types of images. In the TNO dataset, our method performs nicely and has a tiny difference in DEB values compared with other methods, which is limited by the fact that the input images are gray images with low resolution.

3.3 Ablation study

3.3.1 Information quantity discrimination module

We select DB-CNN, Entropy, and BRISQUE to judge the information quantity of the input image, respectively. Among them, DB-CNN is primarily used to judge image contrast stretch, image quantization with color jitters, over- and under-exposure issues, and will have higher scores for high-quality sharp image. Entropy is used to represent the amount of information about an image. A large amount of information represents a small range of data dispersion, and more image details are preserved. BRISQUE extracts the mean subtracted contrast normalized (MSCN) coefficients of natural image to determine the possible image artifacts and distortions in the rich texture regions.

We determine the final amount of IQDM used by testing the effect of different amounts of IQDM on the quality of the fused image. Specifically, we adopted the strategy of controlling variables to conduct experiments in the LLVIP dataset, and tested the influence of each module on the final evaluation index without changing the network structure. According to Table 2, compared with the single method, the combination of the two information content judgment methods improves the image quality obviously, but the optimal index is still composed of the combination of the three information content modules. Therefore, we believe that the three methods of judging the amount of information are all helpful to the improvement of the final index, and the combination of multiple methods is more obvious for the improvement of the index.

In particular, the above three modules are better suited to judge the amount of information under this vision task, and the relevant evaluation metrics can be changed in the construction of the model to suit different vision tasks.

It is worth mentioning that the main function of IQDM in the model is to supervise the selection of information. Therefore, in order to truly compare the effect of each evaluation index, the values in Table 2 are the final results of retraining the model to convergence.

Table 1 Performance of SRGFusion and related methods in four datasets

Method	LLVIP database						M3FD database					
	MI	VIF	Qabf	SD	MS-SSIM	DEB	MI	VIF	Qabf	SD	MS-SSIM	DEB
RFN-NET	1.98	0.54	0.15	8.37	0.68	39.57	2.83	0.87	0.48	9.38	0.72	37.16
PIAFuse	3.97	1.86	0.63	8.84	0.89	68.43	4.21	**1.16**	0.64	8.80	0.93	66.99
SDDGAN	3.16	0.89	0.30	9.01	0.64	45.71	3.07	0.71	0.31	**9.52**	0.65	46.38
SeaFusion	4.11	1.87	0.64	8.41	0.81	64.59	4.02	1.02	**0.66**	8.41	0.83	68.87
TarDAL	3.42	0.59	0.40	8.54	0.72	52.77	3.37	0.80	0.43	9.26	0.92	54.39
Ours	**4.76**	**1.92**	**0.65**	**9.61**	**0.96**	**87.61**	**4.21**	1.10	0.63	9.37	**0.93**	**74.83**
Method	Road scene database						TNO database					
	MI	VIF	Qabf	SD	MS-SSIM	DEB	MI	VIF	Qabf	SD	MS-SSIM	DEB
RFN-NET	1.64	0.56	0.36	8.26	0.72	42.29	2.97	0.82	0.65	9.72	0.71	40.74
PIAFuse	4.42	1.14	0.61	8.13	0.84	68.16	**4.74**	**1.14**	0.66	8.95	0.92	67.17
SDDGAN	3.94	0.69	0.42	8.57	0.74	51.14	3.26	0.72	0.39	8.86	0.63	52.24
SeaFusion	**4.98**	1.10	**0.64**	8.54	0.69	62.74	4.21	1.22	**0.71**	8.35	0.94	64.98
TarDAL	3.81	0.76	0.42	8.27	0.79	54.15	3.82	0.87	0.49	9.36	0.91	58.84
Ours	4.56	**1.15**	0.62	**8.83**	**0.91**	**77.62**	4.41	1.41	0.62	**9.44**	**0.94**	**69.27**

Bold values indicate better results than other filtering methods

Table 2 Impact of different information modules on performance (DB represents DB-CNN, EN represents Entropy, and BR represents BRISQUE)

Modules			MI	VIF	Qabf	SD	MS-SSIM
DB	EN	BR					
✓			4.17	0.88	0.46	8.87	0.62
	✓		4.02	0.62	0.46	8.79	0.64
		✓	3.98	0.77	0.47	8.81	0.62
✓	✓		4.33	1.09	0.59	9.26	0.79
✓		✓	4.35	1.16	0.51	9.31	0.75
	✓	✓	4.26	1.13	0.54	9.28	0.83
✓	✓	✓	**4.76**	**1.15**	**0.65**	**9.61**	**0.96**

Bold values indicate better results than other filtering methods

Table 3 Performance comparison of models with or without diffusion (N/SRGFusion stands for removing diffusion process)

	MI	VIF	Qabf	SD	MS	DEB
Dataset	LLVIP database					
N/SRGFusion	4.13	1.57	0.52	9.25	0.87	78.24
SRGFusion	**4.76**	**1.92**	**0.65**	**9.61**	**0.96**	**87.61**
Dataset	M3FD database					
N/SRGFusion	4.08	0.92	0.58	9.31	0.84	69.34
SRGFusion	**4.21**	**1.10**	**0.63**	**9.37**	**0.93**	**74.83**
Dataset	Road Scene database					
N/SRGFusion	3.97	0.98	0.58	8.76	0.79	70.88
SRGFusion	**4.56**	**1.15**	**0.62**	**8.83**	**0.91**	**77.62**
Dataset	TNO database					
N/SRGFusion	3.85	1.27	0.63	**9.57**	0.86	62.41
SRGFusion	**4.41**	**1.41**	**0.62**	9.44	**0.94**	**69.27**

Bold values indicate better results than other filtering methods

3.3.2 Use diffusion process or not

To demonstrate the effectiveness of the diffusion model, we perform ablation experiments on the diffusion model. Specifically, we retain the U-Net structure used by the original diffusion model as well as our proposed style attention module, but cancel the diffusion process. We summarize the experimental results in Table 3.

On the LLVIP, M3FD, and Road Scene datasets, it performs well in six categories of metrics: MI, VIF, Qabf, SD, MS, and DEB. In the TNO dataset, only the SD index is slightly lower than the model structure under the removing diffusion process, which proves that the use of the diffusion model is extremely beneficial for the generation of high-quality fused image.

In the diffusion model, the steps of diffusion plays a decisive role in the quality of the final image generated by the model. In Table 4, we show the comparison of model scores and parameter numbers for different diffusion step sizes.

Table 4 Diffusion steps and parameters

Diffusion step	MI	VIF	Qabf	SD	MS	Parameters
50	0.35	0.21	0.13	1.62	0.09	33 M
100	2.27	1.18	0.48	4.11	0.51	68 M
200	4.76	**1.92**	**0.65**	9.61	0.96	138 M
300	**4.78**	1.91	0.64	**9.82**	**0.97**	206 M

Bold values indicate better results than other filtering methods

We selected a total of 40 fused images to compute the average index. As can be seen from Table, the smaller the size of the diffusion step, the worse the quality of the generated images. Then, in order to balance the quality of the fused images and the computational efficiency, we choose 200 as the final diffusion steps.

4 Conclusion

In this paper, we propose a semantic information guided image fusion network based on diffusion model for infrared and visible image fusion, called SRGFusion. Firstly, the preprocessing module is used to pre-fuse the infrared visible images to shorten the model training time. Then, the style attention mechanism and diffusion model are used to generate high-quality fused image. Finally, IQDM is used to generate supervision information and compute the loss to ensure model training. In summary, we investigate a diffusion model-based image fusion framework and attempt to bypass complex fusion rules to directly generate high-quality fused image for applications to high-level vision tasks. In the future, we may explore additional lightweight network structures to meet the needs of real-time image fusion.

Author Contributions GH wrote the manuscript and performed data analysis; XZ performed funding acquisition and validation; YH curated and visualized the data.

Funding This work was supported by the Key Research and Development Plan General Project of Shaanxi Provincial Science and Technology Department under Grants (No. 2023-YBGY-032).

Data availability The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
2. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
3. Prabhakar, K.R., Srikar, V.S., Babu, R.V.: Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4724–4732 (2017). https://doi.org/10.1109/ICCV.2017.505
4. Li, H., Wu, X.-J., Kittler, J.: Rfn-nest: an end-to-end residual fusion network for infrared and visible images. Inf. Fusion **73**, 72–86 (2021). https://doi.org/10.1016/j.inffus.2021.02.023
5. Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J.: Piafusion: a progressive infrared and visible image fusion network based on illumination aware. Inf. Fusion **83–84**, 79–92 (2022). https://doi.org/10.1016/j.inffus.2022.03.007
6. Tang, L., Yuan, J., Ma, J.: Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. Inf. Fusion **82**, 28–42 (2022). https://doi.org/10.1016/j.inffus.2021.12.004
7. Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., Zhang, J.: DIDFuse: deep image decomposition for infrared and visible image fusion (2020)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
9. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: a generative adversarial network for infrared and visible image fusion. Inf. Fusion **48**, 11–26 (2019). https://doi.org/10.1016/j.inffus.2018.09.004
10. Zhou, H., Wu, W., Zhang, Y., Ma, J., Ling, H.: Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. IEEE Trans. Multimed. **25**, 635–648 (2023). https://doi.org/10.1109/TMM.2021.3129609
11. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5792–5801 (2022). https://doi.org/10.1109/CVPR52688.2022.00571
12. Yue, J., Fang, L., Xia, S., Deng, Y., Ma, J.: Dif-fusion: towards high color fidelity in infrared and visible image fusion with diffusion models (2023)
13. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics (2015)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
15. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision–ECCV 2018, pp. 334–349. Springer, Cham (2018)
16. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017)
17. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circuits Syst. Video Technol. **30**(1), 36–47 (2020). https://doi.org/10.1109/TCSVT.2018.2886771
18. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
19. Jia, X., Zhu, C., Li, M., Tang, W., Liu, S., Zhou, W.: LLVIP: a visible-infrared paired dataset for low-light vision (2023)
20. Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X.: Fusiondn: a unified densely connected network for image fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12484–12491 (2020). https://doi.org/10.1609/aaai.v34i07.6936
21. Toet, A.: The TNO multiband image data collection. Data Brief **15**, 249–251 (2017). https://doi.org/10.1016/j.dib.2017.09.038