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

Guijin Han
Xi'an University of Posts and Telecommunications
Xinyuan Zhang (xiyouzxy@stu.xupt.edu.cn)
Xi'an University of Posts and Telecommunications
Ya Huang
Xi'an University of Posts and Telecommunications

Research Article

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

Posted Date: July 17th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3154119/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Signal, Image and Video Processing on September 5th, 2023. See the published version at https://doi.org/10.1007/s11760-023-02748-z.

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

Guijin Han, Xinyuan Zhang*, Ya Huang

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

Contributing authors: hanguijin@xupt.edu.cn; xiyouzxy@stu.xupt.edu.cn; xiyouhy@stu.xupt.edu.cn;

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 (SRGFusion). Our model contains three basic parts: preprocessing module, image generation module, and information content discrimination module. 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 information quantity discrimination module (IAQM). 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, and are more suitable for subsequent vision tasks.

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 [1], object tracking [2], and semantic segmentation [3].

We divide the image fusion algorithms into traditional methods and deep learning-based methods, and next we will introduce some representative works, respectively. In the beginning, researchers commonly used traditional methods to complete the image fusion task. The performance of traditional image fusion methods mainly depends on the ability of the model to extract features, and handcrafted feature extractors are used to extract image features followed by manual selection of fusion strategies. Among

them, the multi-scale transform method [4] [5] uses specific rules to decompose different scale features, and various types of fusion strategies are employed to fuse the decomposition results. Sparse representation (SR) [6] and low-rank representation-based methods (LRR) [7] learns a complete model from a high-quality image and utilized to enhance the representation of the source image. Although the above methods achieve good results, they still suffer from the following drawbacks: the fusion performance highly depends on the hand-designed feature extractor, the time is long and the quality is low when dealing with complex image.

In recent years, with the extensive use of residual blocks [8] and dense connections [9], DenseFuse [10] is the first work to apply deep learning to image fusion. Subsequently, RFN-Net [11] 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, PIAFuse [12] combines the cross-modal differential perception fusion module with the semi-fusion strategy, and designs an attention module based on information difference. SeaFusion [13] 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 [14] proposed a deep decomposition model, which uses foreground and background information to assist the fusion network for the performance of fused image in subsequent vision tasks improved.

With the in-depth study of GAN [15], the GAN based image fusion model has opened up a new method for the field of image fusion. Fusion-GAN [16] 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. DDcGAN [17] designed the structure of two discriminators and a generator, so as to retain the information of the two types of images to a greater extent. This has been extensively cited in subsequent work [18–20]. The recent SDDGAN [21] segments the input image into foreground and background information, and completes the generation of the image through semantic information supervision. TarDAL [22] 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 [23], 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. Since the GAN model still has problems such as difficult training and convergence, we innovatively introduce the diffusion model as the fusion image generator. 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 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 [24] and has been widely used for image-to-image and text-to-image generation in the recent work DDPM [25]. The specific steps are shown in Fig. 1. The model is trained by predicting the distribution of noise, and image generation is 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 $t \in \{0,1,\cdots,T-1,T\}$ 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 $x'_t \in \mathbb{R}^{H\times W\times 3}$.

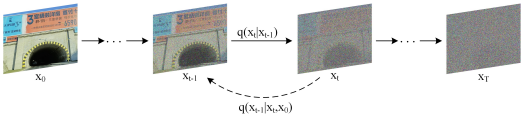


Fig. 1 Diffusion model forward and reverse process.

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):

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

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):

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

2.2 Information Quantity Discrimination Module

We denote the infrared image as $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ and the visible image as $I_{vi} \in \mathbb{R}^{H\times W\times 3}$. As shown in Fig. 2, I_{ir} and I_{vi} are input into the segmentation network to

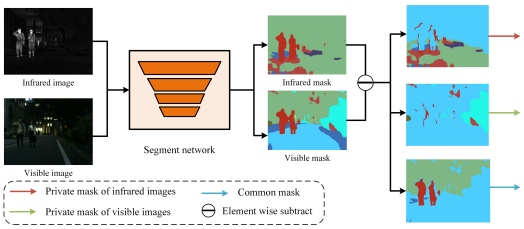


Fig. 2 Three kinds of region masks are obtained by the segmentation model

obtain each semantic level segmentation region, and three common regions are calculated for each region through normalization, as shown in Eq.(5):

$$Seg(I_{ir}) - Seg(I_{vi}) = mask_{ir}^{true}, mask_{commen}^{true}, mask_{vi}^{true} \tag{5}$$

Within this group, $mask \in \{0,1\}$ distinguishes three types of features by positive, negative, and zero values. $mask_{ir}^{true}$ represents the private mask of infrared image, which mainly includes the unclear or nonexistent parts of visible image such as occluded or disguised objects or people. $mask_{commen}^{true}$ corresponds to the common mask of two types of images, mainly including features that appear simultaneously in infrared and visible images. $mask_{vi}^{true}$ corresponds to the private mask of the visible image, which mainly includes texture information, detail features and other information.

To address the issue of insufficient supervision information in fused images and effectively distinguish feature quality, we propose an Information Quantity Discrimination Module (IQDM), as shown in Fig. 3. This module

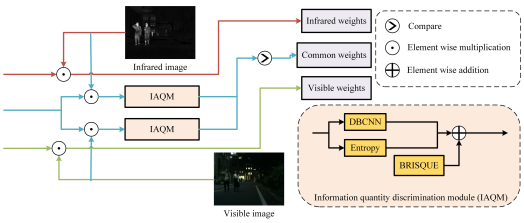


Fig. 3 IQDM obtaining supervision information

obtains high-quality regional image features by evaluating the image quality of each region. Specifically, $mask_{commen}^{true}$ is used to calculate the common areas of I_{ir} and I_{vi} through Eqs.(6) and (7), and the final common area features are obtained through the IQDM.

$$I_{ir}^{common} = I_{ir} \times mask_{common}^{true} \tag{6}$$

$$I_{vi}^{common} = I_{vi} \times mask_{common}^{true} \tag{7}$$

In particular, $mask_{commen}^{true}$ is covered on I_{ir} and upper I_{vi} to obtain each semantic region under common features, and the corresponding information quantity is obtained by

calculating each semantic region, and the optimal common feature region I_{common}^{true} is obtained after comparison. It is worth noting that in the design of the IQDM, we introduce three no-reference image quality assessment methods, namely DB-CNN [26], Entropy, and BRISQUE [27]. 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 Section 3.3.1.

Finally, after calculation by the information module guided by semantic information, we will get I_{ir}^{true}, I_{vi}^{true} and I_{common}^{true}. The three features are combined to obtain the true value I_{true} for supervised learning.

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 $I_f \in \mathbb{R}^{H\times W\times 3}$. Finally, the loss function is used to constrain the training of the network, as shown in Fig. 4.

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.(8):

$$L_{total} = \alpha L_{mse} + \beta L_{ssim} + \gamma L_{\text{color}} \tag{8}$$

Here α, β, and γ are all hyperparameters used to balance the three classes of loss functions.

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.(9):

$$L_{mse} = \sqrt{\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}(I_f(x,y) - I_{true}(x,y))} \tag{9}$$

In order to keep the structure between the fused image and the generated image as complete as possible, L_{ssim} as shown in Eq.(10) is used to constrain.

$$L_{ssim} = 1 - SSIM(I_f, I_{true}) \tag{10}$$

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.(11):

$$L_{color} = \frac{1}{HWC}\sum_{i\in\eta}\sum_{k=1}^{K}\angle(I_{vi}^{i}, I_{f}^{i}), \eta \in \{R,G,B\} \tag{11}$$

Where, 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. In addition, during the ablation experimental phase, we reveal the effectiveness of each module in our proposed model. Finally, the effect of the fusion results in the segmentation model is tested to prove that the proposed method is suitable for high-level vision tasks.

3.1 Experiment Details

1) Datasets: We evaluate the model using infrared and visible images contained in the LLVIP [28], M3FD [29], Road Scene [30], and TNO [31] 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, the Adam optimizer is used to minimize the loss, and the initial learning rate is set to 0.001. 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.

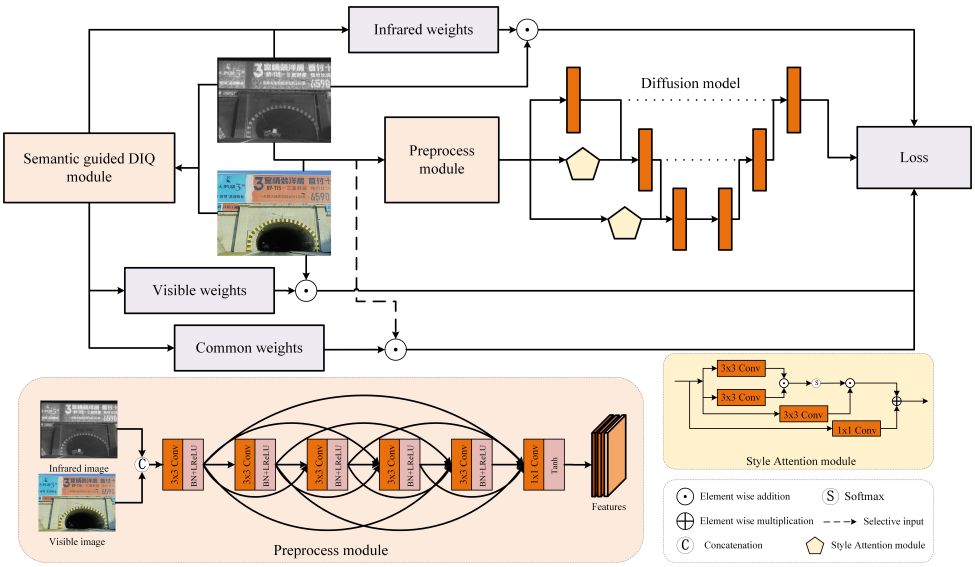


Fig. 4 The framework of the proposed SRGFusion.

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.5 and Fig.6, we show the fusion results for infrared and visible images of #260001 in the LLVIP dataset and #M01442 in the M3DF dataset, respectively. 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.

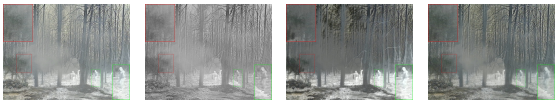
(a) Infrared (b) Visible (c) RFN-Net (d) PIAFuse

(e) SeaFusion (f) SDDGAN (g) TarDAL (h) Ours

Fig. 5 Fusion results of #260001 in the LLVIP dataset

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.

(a) Infrared (b) Visible (c) RFN-Net (d) PIAFuse



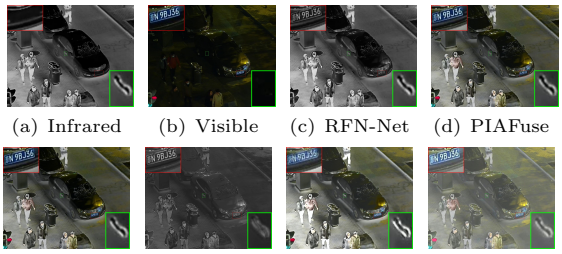
(e) SeaFusion (f) SDDGAN (g) TarDAL (h) Ours

Fig. 6 Fusion results of #M01442 in the M3DF dataset

Fig.6. shows the image fusion results in the wild, and there is a relatively clear smoke edge in the green box area of the visible image. In RFN-Net, Seafusion, and TarDAL, the edge of smoke is blurred, and only PIAFuse and SDDGAN are fused with our method more obviously. However, the trees partially obscured by smoke in the red box region are also challenging for the model to accurately distinguish the texture features. In comparison, our method can better complete the distinction and fusion of texture under the premise of ensuring adequate information.

In Fig.7. and Fig.8., #230070 and #190015 in the LLVIP dataset are selected to show the fusion results of nighttime images. In Fig.7, 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.

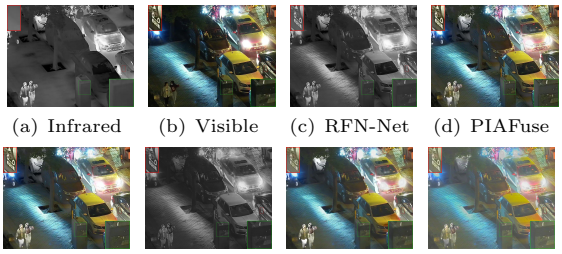


(a) Infrared (b) Visible (c) RFN-Net (d) PIAFuse

(e) SeaFusion (f) SDDGAN (g) TarDAL (h) Ours

Fig. 7 Fusion results of #230070 in the LLVIP dataset

In Fig.7, 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.



(a) Infrared (b) Visible (c) RFN-Net (d) PIAFuse

(e) SeaFusion (f) SDDGAN (g) TarDAL (h) Ours

Fig. 8 Fusion results of #190015 in the LLVIP dataset

In Fig.8, the red area is the advertising sign of the car door, and some environmental information brought by illumination changes is also included around it. It can be seen that various methods are able to fuse such features efficiently, except for the TarDAL method, which has a slight drawback in texture. However, there is no clear texture feature in the infrared image for the transformer box part in the green area, and the dark rectangular band extending to the side is not clear in the visible image. In this case, most methods ignore the extension effect of the dark band, and only our method and TarDAL distinctly retain this feature.

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 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.

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 original network structure but remove the diffusion process, and 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.

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**

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**

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**

Table 4 Performance of the fused image in the segmentation model

Method	RFN-NET	PIAFuse	SDDGAN
PA(%)	10.34	28.42	37.88
mIoU(%)	19.36	34.58	43.69
Method	SeaFusion	TarDAL	ours
PA(%)	35.61	29.34	**38.43**
mIoU(%)	42.64	32.56	**43.77**

3.4 Performance on High-level Vision Tasks

In order for prove that the proposed method is more suitable for subsequent high-level vision tasks, we tested the performance of fused image on DeepLab v3+ [32]. To be fair, the segmentation model uses officially provided pre-trained weights and we do not retrain on the LLVIP dataset.

Table 4. shows the results for image segmentation, which we test on a testset consisting of 40 images, and compare several of the above methods in terms of the common image segmentation metrics pixel accuracy (PA) and mean intersection over union (mIoU). Our fused image perform well in both PA and mIoU, indicating that better segmentation performance can be achieved by assigning different weights to different semantic regions to ensure that the fused image still have their features. Therefore, we believe that supervised learning with semantic information is an approach that can be further investigated and is beneficial for subsequent advanced vision applications.

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.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- 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).

- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
  The authors declare that they have no competing interests.
- Ethics approval
  Not applicable.
- Consent to participate
  Not applicable.
- Consent for publication
  Not applicable.
- Availability of data and materials
  The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.
- Code availability
  Not applicable.
- Authors' contributions
  Guijin Han wrote the manuscript and performed data analysis; Xinyuan Zhang performed funding acquisition and validation; Ya Huang curated and visualized the data.

References

[1] Zhang, H., Wang, S.: Detection of surgical instruments based on gaussian kernel. Signal Image and Video Processing **17**, 3221–3227 (2023) https://doi.org/10.1007/s11760-023-02548-5

[2] Li, C., Zhu, C., Huang, Y., Tang, J., Wang, L.: Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

[3] Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5108–5115 (2017). https://doi.org/10.1109/IROS.2017.8206396

[4] Li, C.-x., Bin, Z.: Using optfr-multiwavelet to improve the fusion of infrared and visible images. In: 2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics, pp. 597–601 (2009). https://doi.org/10.1109/CADCG.2009.5246831

[5] Naidu, V.P.S.: Image fusion technique using multi-resolution singular value decomposition. Defence Science Journal **61**, 479–484 (2011)

[6] Zhang, Q., Fu, Y., Li, H., Zou, J.: Dictionary learning method for joint sparse representation-based image fusion. Optical Engineering **52** (2013)

[7] Li, H., Wu, X.-J.: Multi-focus image fusion using dictionary learning and low-rank representation. In: Image and Graphics, pp. 675–686 (2017)

[8] 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

[9] 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

[10] 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

[11] Li, H., Wu, X.-J., Kittler, J.: Rfn-nest: An end-to-end residual fusion network for infrared and visible images. Information Fusion **73**, 72–86 (2021) https://doi.org/10.1016/j.inffus.2021.02.023

[12] Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J.: Piafusion: A progressive infrared and visible image fusion network based on illumination aware. Information Fusion **83-84**, 79–92 (2022) https://doi.org/10.1016/j.inffus.2022.03.007

[13] 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. Information Fusion **82**, 28–42 (2022) https://doi.org/10.1016/j.inffus.2021.12.004

[14] Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., Zhang, J.: DIDFuse: Deep Image Decomposition for Infrared and Visible Image Fusion (2020)

[15] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks (2014)

[16] Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan:

A generative adversarial network for infrared and visible image fusion. Information Fusion **48**, 11–26 (2019) https://doi.org/10.1016/j.inffus.2018.09.004

[17] Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.-P.: Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing **29**, 4980–4995 (2020) https://doi.org/10.1109/TIP.2020.2977573

[18] Guo, L., Tang, D.: Infrared and visible image fusion using a generative adversarial network with a dual-branch generator and matched dense blocks. Signal Image and Video Processing **17**, 1811–1819 (2023) https://doi.org/10.1007/s11760-022-02392-z

[19] Li, J., Huo, H., Li, C., Wang, R., Feng, Q.: Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. IEEE Transactions on Multimedia **23**, 1383–1396 (2021) https://doi.org/10.1109/TMM.2020.2997127

[20] Yang, Y., Liu, J., Huang, S., Wan, W., Wen, W., Guan, J.: Infrared and visible image fusion via texture conditional generative adversarial network. IEEE Transactions on Circuits and Systems for Video Technology **31**(12), 4771–4783 (2021) https://doi.org/10.1109/TCSVT.2021.3054584

[21] 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 Transactions on Multimedia **25**, 635–648 (2023) https://doi.org/10.1109/TMM.2021.3129609

[22] 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

[23] 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)

[24] Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics (2015)

[25] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models (2020)

[26] Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2020) https://doi.org/10.1109/TCSVT.2018.2886771

[27] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012) https://doi.org/10.1109/TIP.2012.2214050

[28] Jia, X., Zhu, C., Li, M., Tang, W., Liu, S., Zhou, W.: LLVIP: A Visible-infrared Paired Dataset for Low-light Vision (2023)

[29] 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

[30] Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X.: Fusiondn: A unified densely connected network for image fusion. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 12484–12491 (2020) https://doi.org/10.1609/aaai.v34i07.6936

[31] Toet, A.: The tno multiband image data collection. Data in Brief **15**, 249–251 (2017) https://doi.org/10.1016/j.dib.2017.09.038

[32] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (2018)