IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 3011-3020
Journal homepage: http://iieta.org/journals/ts

Infrared Visible Image Fusion Algorithm Based on Illumination Prior and Attention Mechanism

Wei Zhou, Fujun Chen*

School of Integrated Circuits, Zhumadian Vocational and Technical College, Zhumadian 463000, China

Corresponding Author Email: ziei4821@163.com

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.420548

Received: 5 April 2025
Revised: 20 June 2025
Accepted: 7 September 2025
Available online: 31 October 2025

Keywords:
infrared image fusion low-light scene, deep learning, loss function

ABSTRACT

Infrared and visible image fusion (IVIF) technology integrates the all-weather perception capability of infrared thermal radiation with the fine texture information of visible images, demonstrating significant value in applications such as surveillance and navigation. To address the challenges posed by the substantial modality differences between infrared and visible images and the difficulty of information coupling under low-light conditions, this paper proposes an infrared-visible image fusion network guided by illumination prior and attention mechanisms. Firstly, a differential confocal pre-fusion module (DCPFM) is constructed, enabling bidirectional interaction and compensation between infrared thermal features and visible texture information during feature extraction, overcoming the limitation of conventional methods that perform modal interaction only at the fusion stage. Secondly, a hybrid attention fusion strategy is designed to enhance target saliency during feature fusion, effectively mitigating feature blurring caused by conflicts among multi-source information. Finally, an illumination prior regression subnet is built to accurately estimate the lighting conditions of input images, deriving their brightness levels and incorporating this information into the loss function. Experimental results demonstrate that the proposed algorithm achieves optimal performance in six out of seven objective evaluation metrics and second-best in the remaining one, comprehensively outperforming six other comparative algorithms.

1. INTRODUCTION

With the rapid development of computer vision and intelligent perception technologies, the limitations of single sensors in complex environments have become increasingly prominent, and image fusion technology that integrates multi-modal image information has been widely applied in multiple fields. Image fusion refers to the technology of combining multiple images from different sensors, different times, or different perspectives into a single image [1-5], aiming to provide richer and more comprehensive image information. IVIF, as a core technology for multi-source information collaborative analysis, has become a key means to overcome the physical limitations of single sensors. This technology fuses infrared images and visible images by integrating the thermal radiation features of infrared images with the fine texture information of visible images, which can overcome the limitations of single-modality image information representation, improve the reliability of information expression, and achieve functions such as target detection [6], target tracking [7], and semantic segmentation [8]. However, although IVIF technology can combine the all-weather detection advantages of infrared images with the rich texture information of visible images, it still faces significant challenges under low-light and complex scenarios: infrared images have low resolution and blurred details, while visible images suffer from low signal-to-noise ratio and indistinguishable target edges under weak light conditions. Moreover, dynamic interference and multi-source information conflicts in complex scenes further reduce the fusion effect, making it difficult for existing technologies to fully exploit the complementary advantages of multi-modal data. For example, in military reconnaissance night battlefield scenarios, dense smoke coverage and electromagnetic interference often reduce the signal-to-noise ratio of visible light sensors below 5 dB, while high-temperature background radiation overwhelms the target signals, significantly weakening the performance of infrared sensors. Under such harsh conditions, the performance of image fusion deteriorates, severely weakening the target detection capability of military reconnaissance systems [9-15].

According to different fusion strategies and theoretical foundations, image fusion algorithms can be divided into two main categories: traditional image fusion algorithms and deep learning-based image fusion algorithms. Traditional infrared-visible image fusion algorithms can be mainly divided into five categories: multi-scale decomposition-based fusion algorithms [16], sparse representation-based fusion algorithms [17], subspace-based fusion algorithms [18], salient region-based fusion algorithms [19], and hybrid algorithms. Although traditional methods have significant advantages in computational efficiency and resource consumption, they have

certain theoretical limitations: feature extraction lacks modality adaptability, and manually designed fusion rules lead to insufficient generalization. This provides a theoretical breakthrough point for the development of deep learning methods.

In recent years, deep learning technologies have achieved significant breakthroughs in many fields and have demonstrated excellent performance in image fusion. Deep learning-based image fusion algorithms can construct complex data relationships between source images and fused images through loss functions and network parameters obtained from multiple training iterations, thereby maximizing the retention of source image information in the fused image. Li and Wu [20] proposed the DenseFuse algorithm, training the encoder and decoder on the MS-COCO dataset. The dense block structure adopted in the encoder can retain more effective information from source images [21, 22]. In the fusion stage, DenseFuse uses addition and L1-norm strategies to fuse features. However, the fusion strategy of this method is independent of the autoencoder and does not achieve end-to-end fusion. To address the issue that the fusion strategy cannot participate in training, Li et al. [23] proposed a new learnable fusion strategy, whose core is the Residual Fusion Network (RFN). Specifically, a two-stage training is used: first training the autoencoder, then training the RFN, finally achieving end-to-end fusion of infrared and visible images. Zhang et al. [24] designed a general image fusion algorithm based on gradient and intensity ratio preservation, which extracts gradient and intensity information through a network and combines the DenseNet idea for feature reuse. Since this algorithm is a general image fusion method, when applied to infrared-visible image fusion, the loss function needs to be adjusted. Ma et al. [25] first applied Generative Adversarial Networks (GAN) in the image fusion field, realizing end-to-end infrared-visible image fusion based on GAN, avoiding manually designed fusion rules, but this method still has problems of unbalanced information extraction and information loss. To solve these problems and achieve balanced and effective extraction of information from infrared and visible images, Ma et al. [26] proposed DDcGAN, which uses dual discriminators to judge the authenticity between the generated image and the infrared and visible images respectively, achieving stable and balanced fusion of infrared and visible information. Vs et al. [27] proposed the IFT algorithm, which first applies Transformer to image fusion. This algorithm first trains an autoencoder to extract deep features at multiple scales, then uses the Spatio Transformer module to fuse features. This module, by combining Convolutional Neural Network (CNN) and Transformer, can effectively learn and obtain local and long-range features.

Deep learning-based image fusion algorithms can adaptively extract image features using multi-layer networks and achieve higher fusion quality. However, the lack of real fused images in infrared-visible image fusion makes the application of supervised learning methods challenging. Therefore, to improve the quality of fused images, research on network structures and loss functions of image fusion algorithms is particularly important.

In summary, existing infrared-visible image fusion technologies still face challenges in feature modeling, fusion strategies, and algorithm architectures. In terms of feature modeling, existing methods generally use shared-parameter encoders to perform isomorphic feature encoding of the two modality images, without fully considering the essential differences in their physical imaging mechanisms. Visible images construct spatial texture features based on the reflected light intensity of object surfaces, whose information representation is significantly constrained by ambient lighting intensity; infrared images generate temperature field features based on the thermal radiation energy distribution of targets, which are robust to illumination but have low detail resolution. This isomorphic feature encoding paradigm makes it difficult for networks to effectively capture modality-specific information, and in complex scenarios, noise interference exacerbates the confusion of cross-modal feature representations. In terms of feature fusion strategies, existing methods mostly use channel concatenation or weighted superposition as linear operations. Although computationally efficient, they do not establish nonlinear correlation mechanisms between cross-modal features, leading to fused features with high redundancy and weak complementarity.

Therefore, targeting the infrared-visible image fusion problem in low-light scenarios, this paper proposes an Illumination Prior and Attention Network (IPA-Net). The main characteristics of this algorithm include: constructing an illumination prior regression subnet to quantitatively evaluate the illumination intensity of input visible images; designing an end-to-end fusion network architecture, introducing illumination probability as adaptive weights into the loss function to dynamically adjust the weights of different modality features; and integrating a DCPFM with a hybrid attention fusion module to enhance the feature fusion effect.

2. THE ALGORITHM MODEL

2.1 Network structure

IPA-Net is constructed based on CNNs. As shown in Figure 1, the network architecture mainly consists of two parts: the infrared and visible fusion network (main network), which includes an encoder, a fusion layer, and a decoder, and the illumination regression network (subnetwork).

The fusion main network is the core component of IPA-Net. It consists of an encoder, a fusion layer, and a decoder. It takes registered infrared and visible images as input, extracts features through a CNN, and finally reconstructs a fused image that integrates the thermal radiation information of the infrared image and the detailed texture information of the visible image.

The encoder is the front-end part of the main network and is responsible for feature encoding of the input infrared and visible images. This part uses standard convolutional blocks to extract semantic features of the two modality images separately. On this basis, to enhance feature complementarity and reduce redundancy, this chapter proposes a DCPFM, which uses an interactive feature extraction mechanism to connect features of the two different modality source images during the feature extraction stage, providing a feature basis for efficient fusion of infrared and visible information in the fusion layer.

The fusion layer is the core processing part of the main network, responsible for fusing the infrared and visible features extracted by the encoder. In the fusion layer, this chapter proposes a hybrid attention fusion strategy, which first performs shallow fusion of the dual-stream features output by the encoder in a parallel fusion manner. On this basis, spatial attention and channel attention mechanisms are further

introduced. Through attention weight allocation in the spatial dimension and feature selection in the channel dimension, the fused features can focus on more important parts of the fused image, thereby enhancing feature representation ability. This fusion strategy not only preserves the thermal radiation information of infrared images but also effectively retains the detailed texture information of visible images, providing high-quality fused semantic features for subsequent image reconstruction.

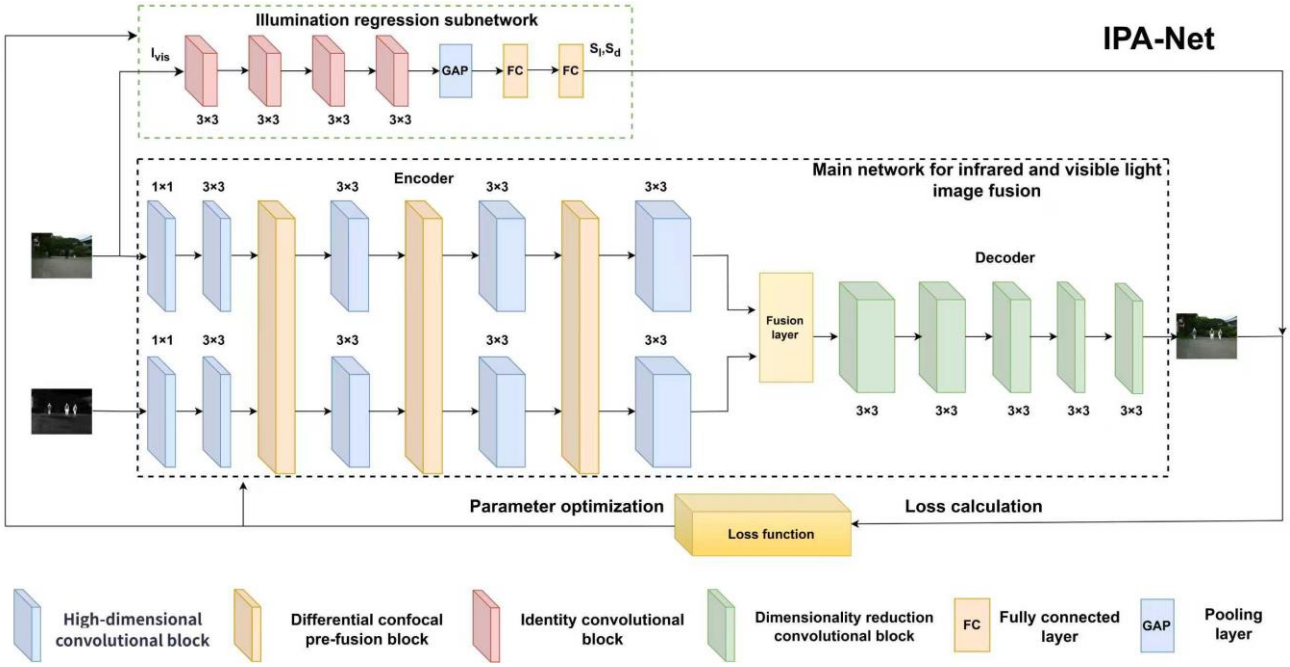

Figure 1. Overall network structure

The decoder is the back-end part of the main network and is responsible for reconstructing the fused image from the high-dimensional semantic features output by the fusion layer. This part gradually restores the spatial features of the image from the fused semantic features extracted by convolutional layers, while compressing the depth dimension of features and reconstructing channel features, outputting a fused image that visually retains both the thermal radiation characteristics of the infrared image and the detailed texture characteristics of the visible image.

The illumination regression subnetwork consists of convolutional layers, pooling layers, and fully connected layers. The convolutional layers take the visible image as input and extract illumination semantic information by gradually compressing resolution and increasing channel dimensions. The pooling layers reduce the dimensionality of the illumination semantic features extracted by the convolutional layers, reducing feature redundancy and computational load, thereby accelerating network convergence. The fully connected layers consist of an input layer, hidden layers, and an output layer. The input and hidden layers map the pooled illumination semantic feature maps into illumination semantic feature vectors, and the output layer regresses the feature vector into an illumination intensity value. The illumination regression subnetwork guides the optimization of the loss function through illumination intensity regression values, effectively enhancing image fusion performance under low-light conditions.

2.2 IVIF main network design

In the fusion main network, the input visible image I_{vis} and infrared image I_{ir} are processed by the encoder, fusion layer, and decoder to obtain the fused image I_f. The encoder network structure is shown in Figure 2, consisting of 5 convolutional layers and 3 DCPFMs. This structure can effectively extract features of infrared and visible images while reducing the differences in feature extraction between different modalities. The specific design of the fusion main network is as follows:

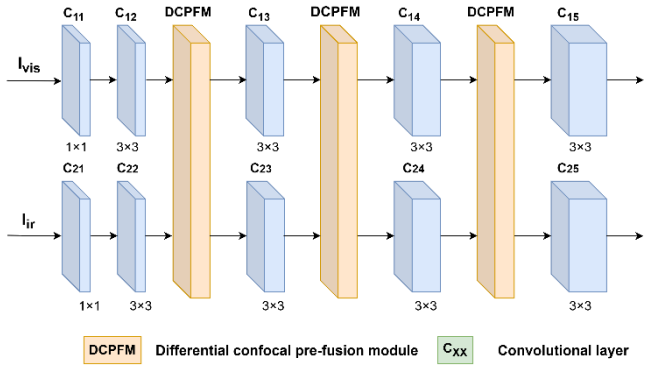

Figure 2. Encoder structure

First, a 1 × 1 convolutional layer preprocesses the input images. This convolutional layer contains two convolutional blocks c_{11} and c_{21}. Its function is to expand the number of image channels to 16 while maintaining the image resolution. This process not only separates features of different dimensions of the image but also maintains image resolution, ensuring that the number of channels of the visible image feature map is consistent with the infrared image feature map, achieving preliminary feature mapping for multi-modal images and providing a basis for subsequent cross-modal information fusion. Then, four 3 × 3 convolutional layers further extract source image features. The four convolutional layers contain eight convolutional blocks c_{12}~c_{15} and c_{22}~c_{25},

respectively increasing the channel dimensions to 16, 32, 64, and 128. The convolutional layers use the LeakyReLU activation function. As a variant of the ReLU function, LeakyReLU can effectively solve the neuron death problem in ReLU, improving network training stability and feature extraction ability. Three DCPFM modules are embedded between the four convolutional layers. This module interconnects features of the two different modality source images during feature extraction, allowing features of the two modalities to complement each other, further enhancing feature complementarity and benefiting efficient fusion in the subsequent fusion layer. The specific structure of the DCPFM module proposed in this chapter will be detailed in Section 2.3.

In the fusion stage, conventional parallel fusion only concatenates feature maps and cannot effectively fuse features of the two modalities, thereby affecting reconstruction performance. Therefore, this chapter proposes a hybrid attention fusion strategy. This strategy first performs parallel fusion of the two source image features obtained in the feature extraction stage, and then designs a fusion module based on channel and spatial attention mechanisms for deep fusion. In the channel attention module, the input feature map is first average-pooled and max-pooled along the spatial dimension, then the results are input to a two-layer fully connected network with shared parameters. After Sigmoid normalization, the obtained channel attention weights are multiplied with the original feature map to obtain the weighted feature map. In the spatial attention module, the weighted feature map is average-pooled and max-pooled along the channel dimension, and the results are concatenated and convolved. The convolutional layer has 512 input channels and 1 output channel, with a kernel size of 7 × 7 and padding set to 3. The spatial attention map obtained after convolution is multiplied with the input feature map to obtain the feature map processed by the hybrid attention module. The hybrid attention fusion strategy can enhance the saliency of fused features and further enrich the information in fused features, providing a higher quality feature basis for subsequent image reconstruction. The specific design and implementation of the hybrid attention fusion strategy will be detailed in Section 3.3.2.

The decoder structure is shown in Figure 3. The decoder consists of 5 convolutional layers. The first four convolutional layers use 3 × 3 kernels with LeakyReLU activation, which gradually restores the spatial information of the image. The last convolutional layer uses a 1 × 1 kernel with Tanh activation, mapping the feature map to a suitable pixel value range to generate the final fused image. The output channel dimensions of the five convolutional layers decrease layer by layer, being 256, 128, 64, 32, and 1, respectively. Through the convolutional layers of the decoder, the high-dimensional semantic feature map output by the fusion layer is gradually decoded and reconstructed into the final fused image.

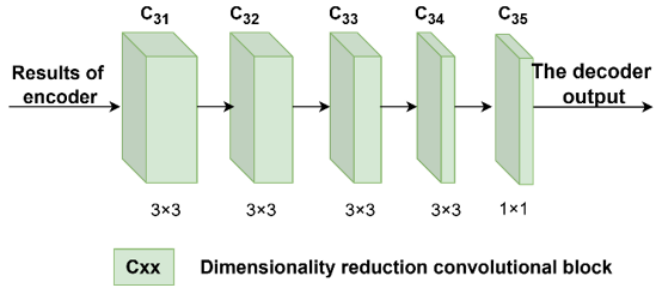

Figure 3. Decoder structure

At the same time, to prevent information loss of the feature map during multi-layer convolution and improve feature extraction accuracy, the stride of all convolutional layers in the network is set to 1. Moreover, except for the first and last layers, the padding of other convolutional layers is set to 1 to maintain the spatial size of the feature map, while the padding of the first and last layers is set to 0 to ensure effective processing of boundary information of the feature map.

2.3 DCPFM

In the field of IVIF, effectively combining the image information captured by infrared cameras and visible light cameras is crucial for improving image quality and information content. However, during the feature extraction stage, since infrared images and visible images are acquired from two different information sources, their data characteristics are significantly different. During the process of extracting feature maps using CNNs, partial feature loss is inevitable. To reduce feature loss and enhance fusion effect, this section proposes a DCPFM at the feature extraction stage. The structure of this fusion module is shown in Figure 4.

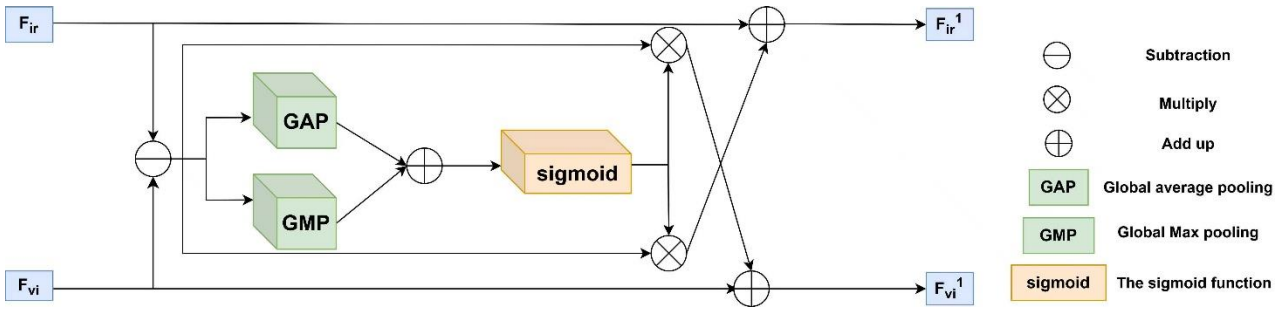

Figure 4. Structure of the differential confocal prefusion module

The DCPFM first calculates the difference between the input infrared feature map and visible feature map to obtain the differential features. Then, the differential features are subjected to Global Average Pooling (GAP) and Global Max Pooling (GMP). GAP is used to capture global statistical information of the differential features, while GMP is used to extract salient information from the differential features. The two pooled feature maps are added pixel-wise, and the differential feature vector is mapped to the range (0,1) through the Sigmoid activation function, generating attention weight vectors for the infrared and visible feature maps, respectively. Next, the generated infrared and visible attention weight

vectors are multiplied with the differential features separately. Finally, the weighted feature maps are added to the differential features to obtain pre-fused infrared and visible features containing different modality information, which are input to the next convolutional layer for further processing. In the pre-fusion module, GAP and GMP are simultaneously used to obtain global statistical information. GAP can better capture global information, while GMP can emphasize salient features. Combining these two pooling methods can fully utilize their respective advantages, reduce the dimensionality of the differences between infrared and visible features, and obtain more comprehensive and robust feature representations.

The core of the DCPFM lies in dynamically adjusting the weights of infrared and visible image features during the pre-fusion process through a differential attention mechanism. This mechanism allows the fusion network to adaptively pre-fuse information from different modalities according to the feature distribution of the input image. In the encoder, the infrared and visible features output by the previous convolutional layer are simultaneously input into the DCPFM, allowing the infrared features to contain partial visible features and the visible features to contain partial infrared features. After processing by three DCPFM modules, information loss during feature extraction can be significantly reduced, providing higher-quality input for further fusion of the two features in the fusion layer.

2.4 Hybrid attention fusion strategy

In image fusion, the fusion strategy is very important. If only simple addition or concatenation is used to fuse infrared and visible features, it may lead to imbalance of modality information and affect the reconstruction result. Therefore, this paper introduces channel and spatial attention mechanisms in the fusion stage and proposes a Hybrid Attention Mechanism (HAM). By combining these two attention mechanisms, the network can focus on more meaningful feature types during fusion, making the obtained fused features richer. The structure of the channel and spatial attention mechanisms used in the fusion part is shown in Figure 5.

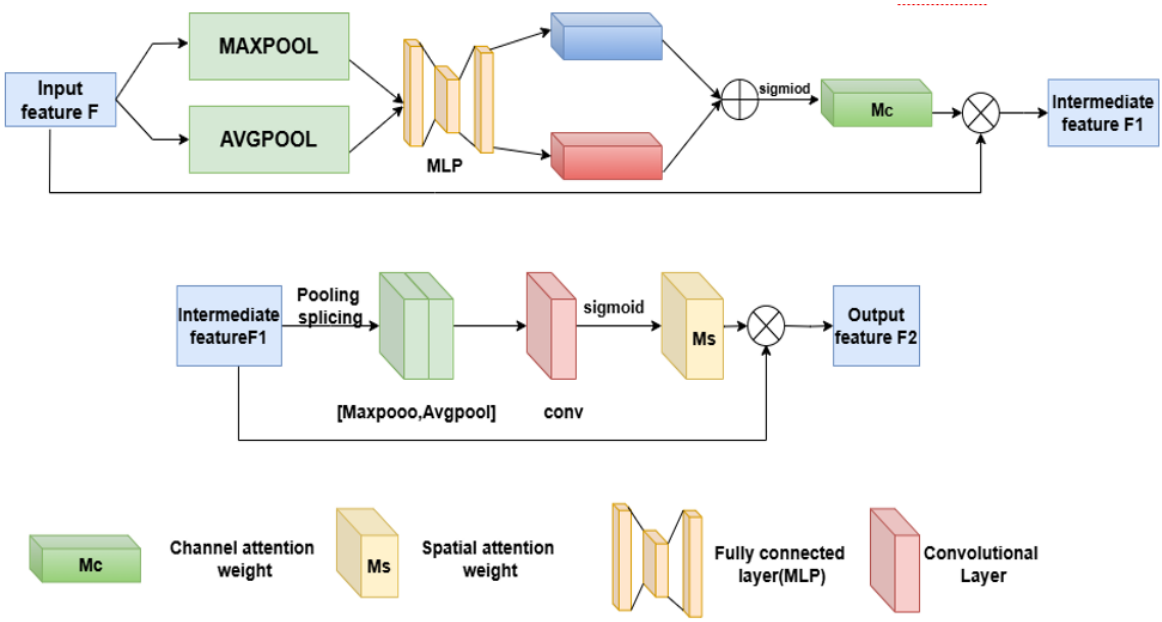

Figure 5. HAM based on channels and spatial attention mechanisms

The hybrid attention fusion strategy first concatenates the two input features and then processes them sequentially through spatial and channel attention mechanisms for deep fusion. The related calculation formulas of the channel attention module are shown in Eqs (3)-(4). The concatenated input feature F is first processed by max pooling and average pooling along each channel to compute the maximum and average feature values per channel. Then, the feature vectors after max pooling and average pooling are input into a shared fully connected layer to obtain attention weight vectors. The attention weight vectors are processed by a Sigmoid function to generate channel attention weights M_c, which are multiplied with each channel of the original input feature map to obtain the attention-weighted channel feature map F_1. After the channel attention module, the channels that are beneficial to the quality of the fused image are strengthened, while irrelevant channels are suppressed.

After parallel fusion and spatial attention mechanism fusion, the channel feature map F_1 is input into the spatial attention module. First, F_1 is max-pooled and average-pooled along the channel dimension. The obtained results are concatenated along the channel dimension and processed by a convolutional layer to generate spatial attention weights. The spatial attention weights are processed by a Sigmoid function to obtain spatial attention weights M_s in the range of 0 to 1. Finally, M_s is multiplied with F_1 to obtain the final feature map F_2 after spatial and channel attention processing. The spatial attention module helps highlight key regions in the feature map and reduces the influence of non-key regions.

2.5 Illumination prior regression subnetwork design

In the task of IVIF, illumination conditions have a significant impact on the intensity information distribution of multi-modal images. Specifically, under insufficient illumination, infrared images often contain more effective intensity information; under sufficient illumination, visible images become the primary information source. However, in

conventional deep learning-based fusion algorithms, the illumination information of visible images is often ignored, and the implicit information of illumination intensity is not fully utilized. Therefore, this chapter proposes an Illumination Prior Regression Network (IPR-Net) to evaluate the illumination condition of the input visible image and output its brightness and darkness weights. By normalizing the brightness and darkness weights and using them as weighting factors in the loss function, the fusion network can dynamically adjust the extraction ratio of intensity information from the two modalities according to the illumination condition of the input visible image. Under sufficient illumination, more visible image intensity information is extracted; under insufficient illumination, more infrared image intensity information is extracted. This illumination-adaptive mechanism significantly improves the robustness and fusion performance of the fusion network under low-light conditions. The structure of the illumination prior regression network is shown in Figure 6.

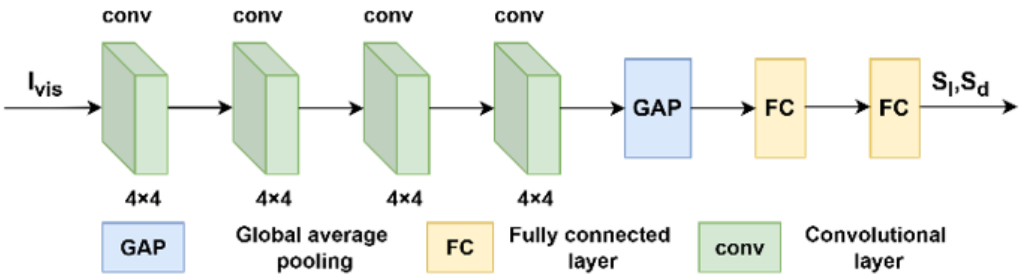

Figure 6. Illumination prior regression network

In the illumination prior regression network, the input visible image I_{vis} first passes through four 4 × 4 convolutional layers, expanding the channel dimension to 128 with LeakyReLU activation. The convolution output is then globally average pooled and input into two fully connected layers to obtain two non-negative scalars S_l and S_d. S_l and S_d represent the brightness and darkness degree of the input image, respectively.

The visible image is input into the fusion network while also being input into the illumination prior regression network. After a series of operations, the weights P_{ir} and P_{vi} for the infrared and visible images in the fusion are obtained. These weights are then used in the loss function to constrain the intensity loss of the two images.

2.6 Loss function design

The loss function plays an extremely important role in the training process of image fusion networks. A reasonable loss function can constrain the input and output images, enabling the trained network model to achieve better performance and obtain high-quality fused images. Therefore, this chapter uses intensity loss (L_i), gradient loss (L_t), and structural similarity loss (L_{SSIM}) to measure the difference between the input images and the fused image. The formula is defined as:

$$L_f = \lambda_1 \cdot L_i + \lambda_2 \cdot L_t + \lambda_3 L_{SSIM} \tag{1}$$

where, λ_1 is the weight of intensity loss, λ_2 is the weight of gradient loss, and λ_3 is the weight of structural similarity loss.

The intensity loss measures the difference between the fused image and the source images, enabling the fused image to have an optimal brightness distribution. In the intensity loss, the weights of infrared intensity loss and visible intensity loss in the loss function are assigned according to the results of the previous illumination discrimination network. The intensity loss function is expressed as:

$$L_i = P_{ir} \cdot L_{int^I} + P_{vi} \cdot L_{int^V} \tag{2}$$

where, P_{ir} and P_{vi} are the weights of infrared and visible images obtained by the illumination discrimination network, and $L_{\text{int}}^{\text{I}}$ and L_{int}^{V} are the intensity losses of infrared and visible images. The intensity losses are specifically defined as:

$$L_{\text{int}}^{\text{I}} = \frac{1}{HW} \left| I_f - I_{\text{ir}} \right|_1 \tag{3}$$

$$L_{\text{int}}^{\text{V}} = \frac{1}{HW} \left| I_f - I_{\text{vi}} \right|_1 \tag{4}$$

where, H is the height of the input image, W is the width of the input image, $|\cdot|_1$ is the L1 norm, I_{ir} and I_{vi} are the infrared and visible images, and I_f is the fused image.

To maximize the fusion of textures from infrared and visible images, gradient loss is introduced to enable the fused image to retain more detailed information from the source images. The gradient loss is formulated as:

$$L_t = \frac{1}{HW} \left| |\nabla I_f| - \max(|\nabla I_{\text{ir}}|, |\nabla I_{\text{vi}}|) \right|_1 \tag{5}$$

where, ∇ is the Sobel gradient operator, and $|\cdot|$ represents the absolute value operation.

Structural similarity loss constrains the correlation between the input images and the fused image, allowing more source image information to be preserved in the fused image. Since there are two source images, the structural similarity between the infrared image and fused image and between the visible image and fused image is calculated separately, and then weighted to obtain the final structural similarity. The SSIM (x, y) is calculated as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)} \tag{6}$$

where, μ represents the mean, σ represents the variance, σ_{xy} represents the covariance of xy, and C is a constant positive number. Additionally, $L_{SSIM_{ir}}$ represents the structural similarity between the infrared image and fused image, $L_{SSIM_{vis}}$ represents the structural similarity between the visible image and fused image, and L_{SSIM} is the total structural

similarity, defined as:

$$L_{SSIM\_ir} = 1 - SSIM(I_F, I_{ir}) \tag{7}$$

$$L_{SSIM\_vis} = 1 - SSIM(I_F, I_{vis}) \tag{8}$$

$$L_{SSIM} = \frac{1}{2} L_{SSIM\_ir} + \frac{1}{2} L_{SSIM\_vis} \tag{9}$$

The total loss function obtained by combining gradient loss, intensity loss, and structural similarity loss can retain rich textures in the fused image while dynamically preserving the intensity information of the source images according to their brightness and darkness, and ensure that the pixel distribution of the fused image is uniform, consistent with the human visual system.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental parameters and environment configuration

Experiments were conducted on the MSRS dataset [76] during the training phase. The MSRS dataset contains multiple pairs of high-quality registered infrared-visible color images. The dataset is of high quality and diverse, providing rich data support for model training. In this study, the original images were cropped using a sliding window of size 64 × 64 with a stride of 64, obtaining 49,000 images as the training set, and 15 pairs of images were selected from the MSRS dataset as the test set.

The model training parameters are shown in Table 1. The number of epochs was 30, the batch size was 4, the learning rate was set to 0.001, and the model parameters were updated using the Adam optimizer. The parameters λ_1, λ_2, and λ_3 in the loss function were set to 3, 25, and 10, respectively.

Table 1. Hyperparameter setting

Hyperparameter	Setting
Epoch	30
Batch size	4
Learning rate	0.001
Optimizer	Adam

3.2 Ablation experiment

To verify the rationality and effectiveness of each module in the proposed algorithm, three ablation experiments were conducted. The specific experimental settings are as follows: (1) In the first experiment, the attention mechanism was removed during the feature fusion phase, and only addition was used for feature fusion, i.e., without the HAM module. (2) In the second experiment, the DCPFM was removed during the feature extraction phase, and only a series of convolution operations were used for feature extraction, i.e., without the DCPFM module. (3) In the third experiment, the weights of infrared and visible images in the intensity loss were both set to 0.5, removing the effect of the prior regression network on training, i.e., without IPR-Net. Moreover, to verify the generalization performance of the proposed modules, which can adapt to both normal and low-light conditions, the ablation experiments used image data under both normal and low-light scenes.

The low-light scene experimental results are shown in Figure 7. Figures 7(a)-7(b) are the infrared and visible images to be fused. It can be seen that the infrared image contains more thermal radiation information but lacks detailed information, while the visible image contains a large amount of detail information but the main targets such as pedestrians are not prominent due to dim lighting. Figure 7(c) shows the fusion result without the HAM module, from which it can be seen that the pre-fusion module and illumination prior regression network can effectively fuse infrared and visible information, but the infrared features in the image are not obvious, and noise information is present. Figure 7(d) shows the fusion result without DCPFM, where infrared features are more obvious, but the visible image contains fewer detailed information and the texture information is blurred. Figure 7(e) shows the fusion result without IPR-Net, where the infrared and visible images are fused well, differing from the complete algorithm only in brightness.

Figure 7. The fusion results of the ablation experiment in the low-light scene. (a) Infrared images; (b) Visible images; (c) Without HAM; (d) Without DCPFM; (e) Without IPR-Net; (f) IPA-Net

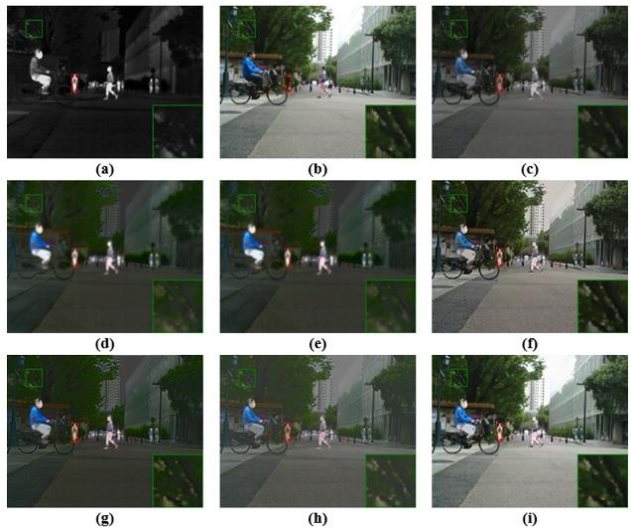

Figure 8. Compare and analyze the visualization results of the experiment 1. (a) Infrared images; (b) Visible images; (c) DenseFuse; (d) FusionGan; (e) GANMcC; (f) IFCNN; (g) SDNet; (h) U2Fusion; (i) IPA-Net

Figure 8 shows the experimental results under normal lighting conditions. Figures 8 (a)-8(b) are infrared and visible images, from which it can be seen that the proposed algorithm can well preserve the texture and details of the visible image while highlighting the infrared features of the infrared image, achieving satisfactory image fusion under normal lighting conditions.

Table 2. Quantitative analysis of ablation experiments

	EN ↑	SD ↑	SF ↑	VIF ↑	MI ↑	SCD ↑	Q_{abf} ↑
Without HAM	5.789	18.88	4.665	0.641	*4.062*	0.834	0.257
Without DCPFM	5.379	*47.359*	10.372	0.261	2.881	0.738	0.099
Without IPR-Net	*6.565*	36.509	*10.328*	*0.781*	3.429	**1.820**	*0.516*
IPA-Net	**6.648**	**47.657**	**11.788**	**0.996**	**4.675**	*1.762*	**0.665**

Table 2 shows the performance results of the ablation experiments on the seven objective evaluation metrics described in Chapter 2. The best values for each metric are highlighted in red bold, and the second-best values are marked in blue italic. It can be seen that IPA-Net achieved the best values in six metrics and the second-best in one metric. Meanwhile, this algorithm shows significant improvement in metrics such as spatial frequency, visual fidelity, and information entropy compared with other algorithms, indicating the superior performance of IPA-Net.

Combined with subjective analysis and objective evaluation metrics, it can be seen that the proposed infrared-visible image fusion algorithm can effectively fuse the features of infrared and visible images, and the attention mechanism, fusion module, and illumination perception module introduced during the fusion process can effectively improve the quality of the fusion results.

3.3 Contrast test

To further verify the performance of IPA-Net compared with other existing algorithms, this section compares IPA-Net with six representative fusion algorithms through subjective visual evaluation and objective metrics analysis.

The comparison results are shown in the figure. The compared algorithms include DenseFuse, IFCNN, U2Fusion, FusionGan, GANMcC, and SDNet. In Figures 9-11, red rectangles indicate the main infrared targets, green rectangles indicate detailed textures, and the lower right corner shows enlarged views.

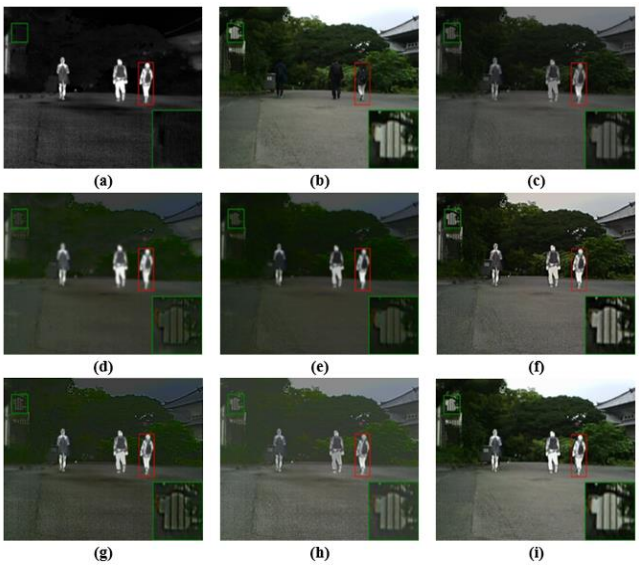

Figure 9. Compare and analyze the visualization results of the experiment 2. (a) Infrared images; (b) Visible images; (c) DenseFuse; (d) FusionGan; (e) GANMcC; (f) IFCNN; (g) SDNet; (h) U2Fusion; (i) IPA-Net

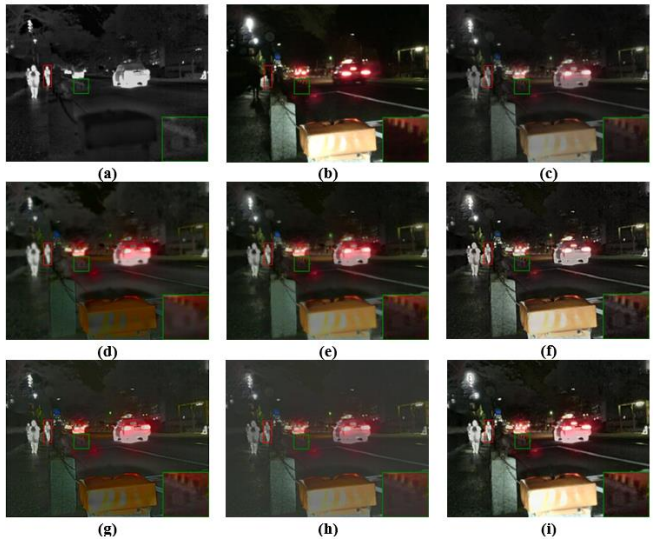

Figure 10. Compare and analyze the visualization results of the experiment 3. (a) Infrared images; (b) Visible images; (c) DenseFuse; (d) FusionGan; (e) GANMcC; (f) IFCNN; (g) SDNet; (h) U2Fusion; (i) IPA-Net

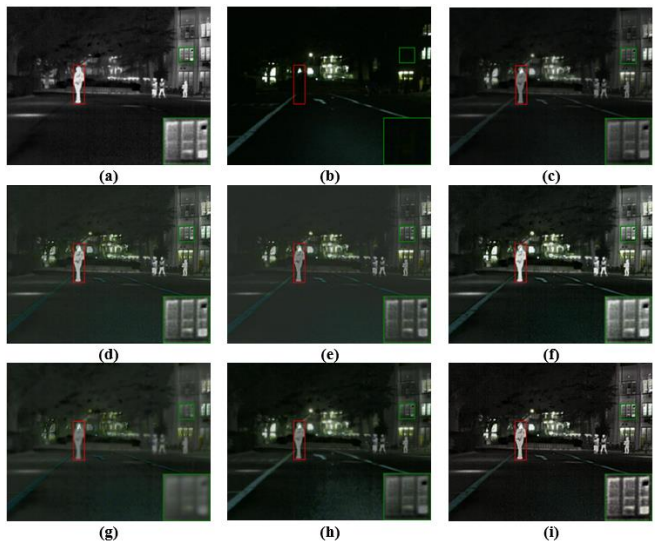

Figure 11. Compare and analyze the visualization results of the experiment 4. (a) Infrared images; (b) Visible images; (c) DenseFuse; (d) FusionGan; (e) GANMcC; (f) IFCNN; (g) SDNet; (h) U2Fusion; (i) IPA-Net

From the overall visual effect of the images, the fusion results of FusionGan and GANMcC have poor visual quality with a lot of noise information. The fusion images of DenseFuse preserve detailed textures well, but the infrared features are not obvious enough. The fusion images of SDNet have more obvious infrared features, but the image brightness is too low, and the overall image is closer to the infrared image. The fusion results of U2Fusion have strong noise interference, and the infrared features are not obvious enough. The fusion results of IFCNN and IPA-Net are relatively good, but the color of IFCNN results is obviously somewhat distorted. In comparison, IPA-Net preserves the details more consistent with the visible image, achieving the best results. In the enlarged parts of the green rectangular boxes in the figure, it can be seen that the visible image shows the details very clearly, while the infrared image is relatively blurred. In the results of DenseFuse and U2Fusion, the thermal radiation

information is not obvious, and pedestrians are not bright enough overall. In FusionGan and GANMcC, the contours and details of pedestrians are too blurred. In the images of IFCNN, SDNet, and IPA-Net, the pedestrian brightness is better, and IPA-Net retains richer detailed information.

To further evaluate the fusion performance of each algorithm, seven image fusion evaluation metrics EN, SD, SF, VIF, MI, SCD, and Qabf were used to score the experimental results of each algorithm. The statistical values of the evaluation metrics are shown in Table 3.

In Table 3, the best metrics are marked in red bold, and the second-best metrics are marked in blue italic. Specifically, in terms of EN, the value of IPA-Net is 6.648, significantly higher than other algorithms. In terms of SD, its value is 47.657, also significantly higher than other algorithms, indicating excellent performance in preserving image details. In terms of SF, the value of IPA-Net is 11.788, which is not the highest but still at a high level, indicating good performance in maintaining image structure similarity. In terms of visual VIF, IPA-Net has a value of 0.996, close to 1, indicating excellent visual information fidelity. In terms of MI, IPA-Net has a value of 4.675, significantly higher than other algorithms, indicating obvious advantages in information fusion. In terms of SCD, IPA-Net has a value of 1.762, the best among all algorithms, indicating excellent performance in maintaining image spatial consistency.

In summary, the proposed algorithm achieves the best results in quantitative metrics EN, SD, VIF, MI, SCD, and Qabf, and the second-best in SF. Moreover, the scores in SD and MI are far higher than those of other algorithms, indicating that the results of IPA-Net have better contrast and image quality than other algorithms.

Combining subjective and objective evaluation metrics, it can be concluded that the fusion images obtained by IPA-Net have overall high quality, rich detailed textures, obvious infrared features, and better fusion performance than other comparison algorithms.

Table 3. Average performance indicators of different algorithms

	EN ↑	SD ↑	SF ↑	VIF ↑	MI ↑	SCD ↑	Qabf ↑
DenseFuse	6.011	26.855	6.225	0.676	*2.784*	1.338	0.362
FusionGan	5.591	21.683	4.595	0.379	2.077	0.952	0.141
GANMcC	5.986	27.394	5.542	0.560	2.641	1.421	0.268
IFCNN	6.392	*39.563*	**12.243**	*0.800*	2.788	*1.681*	*0.606*
SDNet	*6.406*	21.624	8.629	0.506	2.069	1.037	0.384
U2Fusion	4.980	20.387	6.477	0.464	2.179	0.968	0.284
IPA-Net	**6.648**	**47.657**	*11.788*	**0.996**	**4.675**	**1.762**	**0.665**

4. CONCLUSION

This paper proposes an infrared-visible image fusion algorithm based on illumination prior and attention guidance. In image fusion tasks, illumination intensity has a non-negligible impact on fusion results. To effectively incorporate illumination factors into the fusion system, this study constructs an illumination prior regression network. Before the visible image is input into the fusion network, the illumination prior regression network outputs quantitative results regarding the brightness and darkness of the image. Subsequently, this judgment result is incorporated into the intensity loss term of the loss function to guide fine control of the fusion process, enabling the fusion network to adaptively retain more critical intensity information from infrared and visible images under different illumination conditions, thus laying a solid foundation for subsequent fusion operations.

To further improve fusion performance, a DCPFM and a hybrid attention mechanism-based fusion strategy were designed. DCPFM realizes deep information interaction between two different modal images based on the principle of image feature interaction. Embedding DCPFM in the feature extraction stage allows pre-fusion of infrared and visible features before the features are input into the fusion layer. This operation not only effectively reduces information loss during feature extraction but also promotes the network to integrate and optimize features from different modalities, improving feature quality and effectiveness.

In the hybrid attention fusion strategy, spatial attention and channel attention mechanisms are introduced. The spatial attention mechanism focuses on critical information in image spatial positions, capturing spatial distribution characteristics of targets. The channel attention mechanism focuses on important features along the image channel dimension, mining the relationships between different channels. These two attention mechanisms cooperate, allowing the network to accurately locate and extract key parts of infrared and visible features during fusion while effectively suppressing redundant information, achieving targeted and efficient feature fusion.

Through a large number of experiments, the proposed algorithm demonstrates good fusion image quality. The generated fusion images contain not only prominent infrared features but also rich texture details. Compared with six current advanced image fusion methods, IPA-Net shows obvious advantages in both subjective visual effects and objective evaluation metrics. In subjective evaluation, the fusion images perform well in target identification and detail clarity. In objective evaluation, six of the seven evaluation metrics achieve the best values, and one achieves the second-best value, significantly outperforming the other six comparison algorithms.

REFERENCES

[1] Toet, A., Van Ruyven, L.J., Valeton, J.M. (1989). Merging thermal and visual images by a contrast pyramid. Optical Engineering, 28(7): 789-792. https://doi.org/10.1117/12.7977034

[2] Chandrakanth, V., Murthy, V.S.N., Channappayya, S.S. (2022). Siamese cross-domain tracker design for seamless tracking of targets in RGB and thermal videos. IEEE Transactions on Artificial Intelligence, 4(1): 161-172. https://doi.org/10.1109/TAI.2022.3151307

[3] Lahmyed, R., El Ansari, M., Ellahyani, A. (2019). A new thermal infrared and visible spectrum images-based

pedestrian detection system. Multimedia Tools and Applications, 78(12): 15861-15885. https://doi.org/10.1007/s11042-018-6974-5

[4] Kong, S.G., Heo, J., Abidi, B.R., Paik, J., Abidi, M.A. (2005). Recent advances in visual and infrared face recognition—A review. Computer Vision and Image Understanding, 97(1): 103-135. https://doi.org/10.1016/j.cviu.2004.04.001

[5] Ariffin, S.M.Z.S.Z., Jamil, N., Rahman, P.N.M.A. (2017). Can thermal and visible image fusion improves ear recognition? In 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, pp. 780-784. https://doi.org/10.1109/ICITECH.2017.8079945

[6] Bavirisetti, D.P., Xiao, G., Zhao, J., Dhuli, R., Liu, G. (2019). Multi-scale guided image and video fusion: A fast and efficient approach. Circuits, Systems, and Signal Processing, 38(12): 5576-5605. https://doi.org/10.1007/s00034-019-01131-z

[7] Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V. (2018). An unsupervised learning model for deformable medical image registration. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 9252-9260. https://doi.org/10.1109/CVPR.2018.00964

[8] Ghosh, S., Gavaskar, R.G., Chaudhury, K.N. (2019). Saliency guided image detail enhancement. In 2019 National Conference on Communications (NCC), Bangalore, India, pp. 1-6. https://doi.org/10.1109/NCC.2019.8732250

[9] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1647-1655. https://doi.org/10.1109/CVPR.2017.179

[10] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 5967-5976. https://doi.org/10.1109/CVPR.2017.632

[11] Jaderberg, M., Simonyan, K., Zisserman, A. (2015). Spatial transformer networks. In Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal, Canada, pp. 2017-2025.

[12] Singh, R., Vatsa, M., Noore, A. (2008). Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition. Pattern Recognition, 41(3): 880-893. https://doi.org/10.1016/j.patcog.2007.06.022

[13] Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P. (2002). Color transfer between images. IEEE Computer Graphics and Applications, 21(5): 34-41. https://doi.org/10.1109/38.946629

[14] Kumar, P., Mittal, A., Kumar, P. (2006). Fusion of thermal infrared and visible spectrum video for robust surveillance. In Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006, Madurai, India, pp. 528-539. https://doi.org/10.1007/11949619_47

[15] Simone, G., Farina, A., Morabito, F.C., Serpico, S.B., Bruzzone, L. (2002). Image fusion techniques for remote sensing applications. Information Fusion, 3(1): 3-15. https://doi.org/10.1016/S1566-2535(01)00056-2

[16] Luo, Y., He, K., Xu, D., Yin, W., Liu, W. (2022). Infrared and visible image fusion based on visibility enhancement and hybrid multiscale decomposition. Optik, 258: 168914. https://doi.org/10.1016/j.ijleo.2022.168914

[17] Liu, Y., Liu, S., Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. Information Fusion, 24: 147-164. https://doi.org/10.1016/j.inffus.2014.09.004

[18] Gao, X.Q., Liu, G., Xiao, G., Bavirisetti, D.P., Shi, K. (2020). Fusion algorithm of infrared and visible images based on FPDE. Journal of Automation, 46(4): 796-804.

[19] Li, H., Wu, X.J., Kittler, J. (2020). MDLatLRR: A novel decomposition method for infrared and visible image fusion. IEEE Transactions on Image Processing, 29: 4733-4746. https://doi.org/10.1109/TIP.2020.2975984

[20] Li, H., Wu, X.J. (2018). DenseFuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing, 28(5): 2614-2623. https://doi.org/10.1109/TIP.2018.2887342

[21] Lin, T.Y., Maire, M., Belongie, S., Hays, J., et al. (2014). Microsoft coco: Common objects in context. In Computer Vision - ECCV 2014: 13th European Conference, Zurich, Switzerland, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[22] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2261-2269. https://doi.org/10.1109/CVPR.2017.243

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

[24] Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J. (2020). 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, 34(7): 12797-12804. https://doi.org/10.1609/aaai.v34i07.6975

[25] Ma, J., Yu, W., Liang, P., Li, C., Jiang, J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion, 48: 11-26. https://doi.org/10.1016/j.inffus.2018.09.004

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

[27] Vs, V., Valanarasu, J.M.J., Oza, P., Patel, V.M. (2022). Image fusion transformer. In 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, pp. 3566-3570. https://doi.org/10.1109/ICIP46576.2022.9897280