Learning to Follow Infrared Prior Repersentation for Image Dehazing

Yiquan Du[1], Ming Yang[1], Haiyang Huo[1], Jiaqi Shi[1], Jiangang Ding[1], Lili Pei[2,*]
[1]School of Information Engineering, Chang'an University, Xi'an, Shaanxi 710064, China
[2]School of Data Science and Artificial Intelligence, Chang'an University, Xi'an, Shaanxi 710064, China
{dyq, ym, 2023124014, 2023124015, djg, peilili}@chd.edu.cn

***Abstract*—The infrared image can distinguish the targets from the background based on radiation differences, providing more significant target visibility under dense haze. Fusion of visible haze images with infrared prior representations can generate high-quality fused images for high-level tasks. Consequently, we propose a novel dual-modal fusion network structure that makes full use of infrared prior representations for dehazing. Specifically, we emphasize a Multi-modal Feature Extraction Network (MMFE) to extract deep multi-scale features. Meanwhile, we introduce a Multi-scale Feature Extraction module (MSFE), integrating an Efficient Dual Attenion block (EDAB) to efficiently explore more spatial and marginal information. Additionally, we propose a new feature fusion strategy, which calculates feature fusion weights based on an adaptive multi-head self-attention. Therefore, IPRDehazeNet achieves better dehazing results through dual-modal fusion. Experimental results indicate that IPRDehazeNet outperforms various advanced methods.**

***Index Terms*—Multi-modal, Multi-scale, Visible-Infrared fusion, Image dehazing**

I. Introduction

In traffic monitoring and autonomous driving, dense hazy weather degrades image quality, leading to loss of target and impacting downstream tasks [1], [2]. Currently, single image dehazing methods mainly include the following aspects: 1) Traditional dehazing algorithms based on dark channel prior and color attenuation prior [3] are limited in dense fog conditions. 2) Deep learning methods are still impacted by the limited information of a single modal, including CNN-based [4], GAN-based [5], and Transformer-based [6] approaches. Infrared technology relies on thermal radiation imaging principles, which possesses strong penetration capabilities and excellent resistance to environmental interference.

Currently, many studies have leveraged cross-modal information to enhance high-level tasks [7], [8], [9], [10]. Combining visible light and infrared modalities could improve target prominence and texture in fused images [11], [12]. Therefore, researchers have been continuously improving infrared and visible image fusion methods, such as based on autoencoders (AE) [13], convolutional neural networks (CNN) [14], generative adversarial networks (GANs) [15], and transformers [16]. Although these methods have made great progress [17], [18], images restored in dense hazy conditions still don't meet expectations. The issues are pixel-level alignment and deep information extraction [19], [20], [21], as shown in Fig. 1 (c). Our inspiration comes from: a) In the feature extraction stage, dual-modal images can leverage attention mechanisms and high-low frequency signal processing to extract deep, multi-scale features. Additionally, infrared prior representation is utilized to enhance the quality of the extracted features. b) In the feature fusion stage, we can design a novel strategy to adjust the fusion weights, which emphasizes more reliable and consistent information. Recently, the Spatial and Channel Reconstruction Convolution (SCConv) [22], [23] has emerged as an attention mechanism that enhances the network's sensitivity to key features. Based on this, we propose a Multi-Scale Feature Extraction module that incorporates an Efficient Dual Attention Block (EDAB) to explore more spatial and marginal information within the feature maps. The EDAB utilizes a Spatial Reconstruction Unit (SRU), a Channel Reconstruction Unit (CRU), and efficient convolutional blocks to achieve enhanced performance. Additionally, adaptative multi-head self-attention is a core component of the Transformer model.By calculating global attention weights and integrating diverse information [24], it enhances the model's ability to capture long-range dependencies and enrich the representation of elements within the sequence. Consequently, we propose the Adaptive Feature

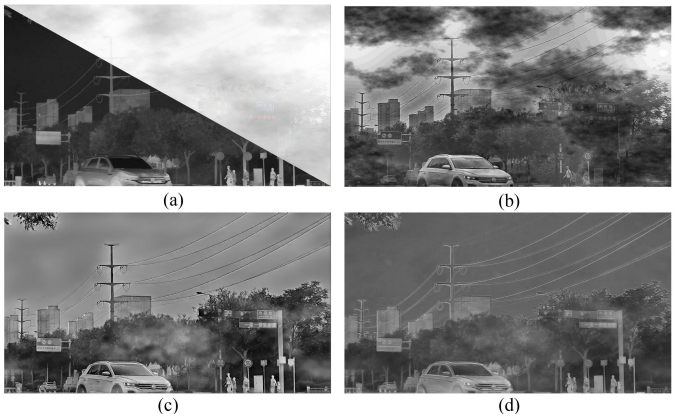

Fig. 1. (a) shows the original visible image and infrared image. (b) illustrates the dehazing effect using the traditional dark channel prior method. (c) presents the dehazing results of the deep learning method LDFusion. (d) demonstrates the dehazing effect of our infrared and visible fusion method.

* Corresponding author. This work was supported in part by the National Natural Science Foundation of China (No. 51978071), the Fundamental Research Funds for the Central Universities, CHD (No. 300102244714) and the Scientific Innovation Practice Project of Postgraduates of Chang'an University (No. 300103724054).

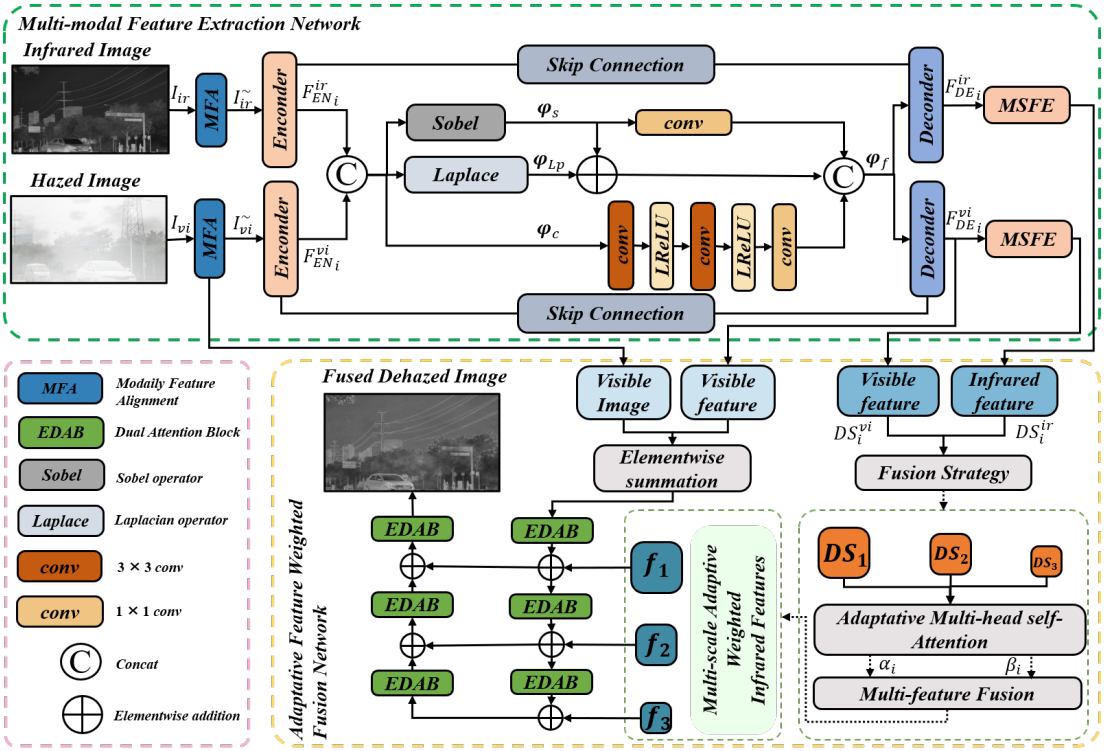

Fig. 2. The overview of our workflow. MMFE is used to extract deep multi-scale features from the images. The AFWF performs adaptive weighted fusion of deep features at different scales, followed by image reconstruction.

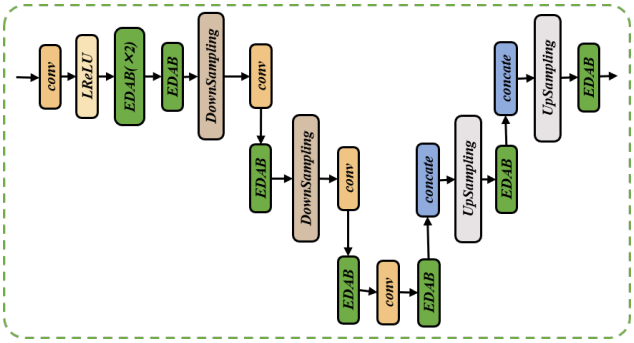

Fig. 3. The architecture of Encoder-Decoder.

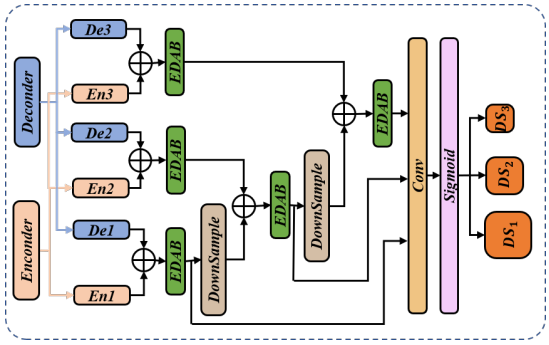

Fig. 4. The architecture of MSFE module.

Weighted Fusion (AFWF) network, which introduces a novel strategy for adaptively assigning feature fusion weights. This approach leads to improved dehazing results during the image reconstruction phase. The main contributions are as follows:

- We propose a novel dual modal fusion network IPRDehazeNet to achieve image dehazing.
- We present the Multi-scale Feature Extraction (MSFE) module which incorporates an Efficient Dual Attention block (EDAB) to utilize infrared prior representation.
- The adaptative multi-head self-attention dynamically adjusts the fusion weights between visible and infared features.

II. Methodology

A. Multi-modal Feature Extraction Network

In a binocular vision system, the resolution disparity between visible and infrared images necessitates precise alignment before processing. To address this, we introduce a Modality Feature Alignment (MFA) module [25].

The network architecture uses an encoder-decoder framework with skip connections to extract multi-scale features and integrate them into the upsampling process, as shown in Fig. 3. It also uses EDAB to improve feature extraction effect. The encoder extracts multi-scale features through convolution and downsampling, while the decoder restores spatial resolution through upsampling and concatenation, leveraging multi-scale information from the encoder.

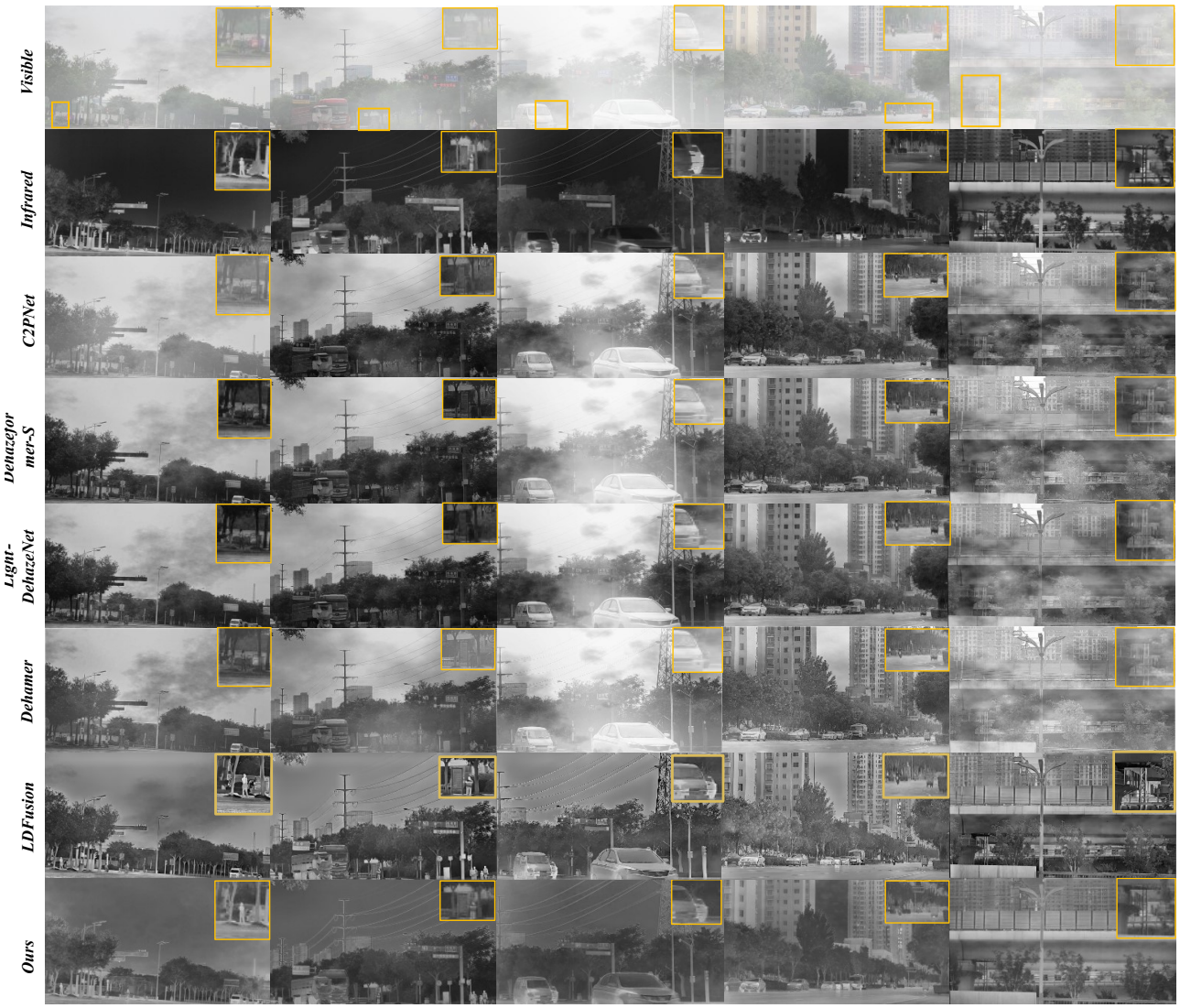

Fig. 5. Results of different methods. Zoom in to see the details.

The process of feature extraction uses 3×3 convolutional layers with LReLU and a 1×1 convolutional layer. The first residual stream incorporates the Sobel operator to preserve strong textures, while the 1×1 layer eliminates channel dimension differences. The second residual stream introduces the Laplace operator to extract weak textures. The final filtered feature ϕ_{filter} is obtained by concatenating the outputs of two operations, as illustrated in Eq.1.

$$\phi_{filter} = \text{concat}\left(\text{conv}\left(\phi_{Lp} \oplus \nabla^2 \phi_{Lp}\right), \text{conv}\left(\nabla \phi_S\right)\right) \quad (1)$$

B. Multi-scale Feature Extraction Module

The primary goal of this module is to enhance the interaction between contextual and multi-scale spatial features, as shown in Fig. 4. By using convolutional layers as encoders F_{ENi} and decoders F_{DEi}, it extracts multi-scale deep structural features, integrating more perceptual information. The process can be summarized as follows: EDAB adjusts the weights of each channel, considering the mutual influence between multi-scale features. By connecting high-level feature maps generated by downsampling with low-level feature maps, it helps maintain high spatial resolution information. EDAB is then applied again to account for dependencies between feature channels, optimizing channel weights and enhancing feature representation. Finally, deep structural features at three scales are derived through convolutional layers and Sigmoid activation function layers. The calculation of deep structural features at the *i*-th scale DS_i can be expressed as:

$$\begin{cases} \text{DS}_1 = \sigma\left(\mathbb{E}\left(F_{ED1}\right)\right) \\ \text{DS}_2 = \sigma\left(\mathbb{E}\left(\mathbb{E}\left(\downarrow\left(F_{ED1}\right) \oplus F_{ED2}\right)\right)\right) \\ \text{DS}_3 = \sigma\left(\mathbb{E}\left(\mathbb{E}\left(\downarrow\left(\mathbb{E}\left(\downarrow\left(F_{ED1}\right) \oplus F_{ED2}\right)\right) \oplus F_{ED3}\right)\right)\right) \end{cases} \quad (2)$$

where σ denotes the combination of the final convolution layer and Sigmoid function layer, and $\mathbb{E}$ represents the EDAB. Besides, the downsampling process is symbolized as $\downarrow$. The concatenated feature map F_{EDi} combines F_{ENi} and F_{DEi}.

C. Adaptive Feature Weighted Fusion Network

To fully leverage the depth structure features at various scales, we propose a strategy based on the adaptive multi-head self-attention to generate dynamic weights. This method captures the dependencies between features at different scales and generates appropriate weights for multi-scale and multi-modal fusion. This stage produces multi-scale and multi-modal fusion features. The formulas are as follows:

$$\alpha_i, \beta_i = \text{GlobalPool}\left(\text{Concat}\left(\text{head}_1, \text{head}_2, \text{head}_3\right) \mathbf{W}_o\right) \quad (3)$$

$$f_i = \left(\alpha_i \, \text{DS}_i^{vi} \cdot \text{DS}_i^{in} + \beta_i \overline{\text{DS}_i^{vi}} \cdot \overline{\text{DS}_i^{in}}\right) \otimes DS_i^{in} \quad (4)$$

where α_i and β_i represent the corresponding weight of each items. $\overline{\text{DS}_i^{vi}}$ and $\overline{\text{DS}_i^{in}}$ mean the inverse operation of the deep structure feature. f_i is the adaptive weighted structure feature.

D. Loss Function

We designed a composite loss function consisting of two key components: 1) Mean Squared Error (MSE), also known as L1 loss, which is widely used in image dehazing tasks for its simplicity and effectiveness. 2) Dice Loss ensures the fused image accurately reflects the stronger edges and texture details of the input or target image, enhancing overall image quality.

$$\begin{aligned} \mathcal{L}_{\text{global}} &= \lambda_1 \mathcal{L}_{\text{L1}} + \lambda_2 \mathcal{L}_{\text{Dice}} \\ &= \frac{\lambda_1}{HW} \left\| |I_f| - \max\left(|I_{vi}|, |I_{ir}|\right) \right\|_1 \\ &\quad + \frac{\lambda_2}{HW} \left\| |\nabla I_f| - \max\left(|\nabla I_{vi}|, |\nabla I_{ir}|\right) \right\|_1 \end{aligned} \quad (5)$$

where λ_1 , λ_2 are the corresponding coefficients. I_f , I_{vi} and I_{ir} respectively representing the fused image, the infrared image, and the dehazed image. ∇ represents their gradients.

III. Experiments

A. Dataset

There is no publicly available cross-modal dataset for fog removal in visible and infrared image fusion. To address this, we used a binocular camera to collect a dataset of 10,900 pairs of images, each containing an RGB image (2688×1520) and a thermal infrared image (1280×1024). We registered these images using calibration algorithms and applied varying fog levels using the imgaug library in Python.

B. Implementation Details

Our graphics card is an Nvidia RTX 4060 Ti 16 GB. We are using PyTorch version 1.8.2, and Torchvision version 0.9.2. The CUDA version installed is 11.1. The batch size is 8, and the number of epochs is 100. The initial learning rate is 0.0001.

TABLE I
COMPARISON RESULTS OF DIFFERENT METHODS FOR INFRARED IMAGE FUSION AND DEHAZING EXPERIMENTS.THE SYMBOLS ↑ AND ↓ DENOTE A PREFERENCE FOR HIGHER AND LOWER VALUES. THE FIRST THREE RESULTS ARE IN THE ORDER OF RED, BLUE, AND GREEN.

Method	PSNR↑	SSIM↑	MSE↓	FID↓	Params(M)↓	Time(s)↓
Dehamer [26]	11.05	0.758	0.5106	53.26	29.44	0.115
Dehazeformer-S [27]	13.42	0.728	0.2959	45.68	1.28	0.013
Light-DehazeNet [28]	15.08	0.783	0.2019	46.17	2.60	0.720
C2PNet [29]	17.44	0.850	0.1172	32.17	7.17	0.231
IPRDehazeNet (Ours)	22.37	0.894	0.0398	30.64	6.89	0.105

TABLE II
COMPONENT ABLATION RESULTS. THE SYMBOLS ↑ AND ↓ DENOTE A PREFERENCE FOR HIGHER AND LOWER VALUES, RESPECTIVELY. THE FIRST THREE RESULTS ARE IN THE ORDER OF RED, BLUE, AND GREEN.

Model	Base	MSFE	Fusion Strategy	PSNR↑	SSIM↑
Model1	✓			16.62	0.818
Model2	✓	✓		19.14	0.853
Model3	✓		✓	20.71	0.873
Model4	✓	✓	✓	22.37	0.894

C. Comparison with SOTA Methods

To validate the effectiveness of our proposed method, we compared it with several state-of-the-art approaches. The experimental results report that our IPRDehazeNet achieved the best performance, as shown in Table I. We provide a visual comparison in Fig. 5. Additionally, we only compare with dehazing methods and not with fusion methods. Because although the fusion method looks good, there are actually serious structural distortions and ghosting, as shown in Fig. 5 in LDFusion section. The parameter count and inference time per image are at relatively low levels. This demonstrates that our method not only ensures dehazing capability but also considers efficiency.

D. Component Ablation

Table II reports the contribution of each component. The experimental results confirm the effectiveness of each independent component we introduced. The results show that all the proposed modules are effective, and the absence of the depth feature extraction module or the adaptive weight fusion strategy would lead to a performance decrease.

IV. Conclusion

In this work, we introduce IPRDehazeNet which designed to fuse infrared and visible images to generate high-quality dehazed images. It is achieved through a multi-modal deep feature extraction network and an adaptive feature weight fusion strategy based on multi-head self-attention. Experimental results demonstrate that our network achieves superior dehazing and fusion performance.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers for their constructive comments.

REFERENCES

[1] Liu J, Fan X, Huang Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modal benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5802-5811.

[2] Ding J, Li W, Pei L, et al. Sw-YoloX: An anchor-free detector based transformer for sea surface object detection[J]. Expert Systems with Applications, 2023, 217: 119560.

[3] He K, Sun J, Tang X. Single image haze removal using dark channel prior[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(12): 2341-2353.

[4] Li B, Peng X, Wang Z, et al. Aod-net: All-in-one dehazing network[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4770-4778.

[5] Ren W, Zhou L, Chen J. Unsupervised single image dehazing with generative adversarial network[J]. Multimedia Systems, 2023, 29(5): 2923-2933.

[6] Valanarasu J M J, Yasarla R, Patel V M. Transweather: Transformer-based restoration of images degraded by adverse weather conditions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2353-2363.

[7] Ding J, Li W, Pei L, et al. Novel Pipeline Integrating Cross-modal and Motion Model for Nearshore Multi-Object Tracking in Optical Video Surveillance[J]. IEEE Transactions on Intelligent Transportation Systems, 2024.

[8] Ding J, Li W, Yang M, et al. SeaTrack: Rethinking Observation-Centric SORT for Robust Nearshore Multiple Object Tracking[J]. Pattern Recognition, 2025, 159: 111091.

[9] Zhao Y, Li W, Ding J, et al. Nearshore optical video object detector based on temporal branch and spatial feature enhancement[J]. Engineering Applications of Artificial Intelligence, 2024, 138: 109387.

[10] Zhao Y, Li W, Ding J, et al. Crack instance segmentation using splittable transformer and position coordinates[J]. Automation in Construction, 2024, 168: 105838.

[11] Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: A survey[J]. Information fusion, 2019, 45: 153-178.

[12] Zhang X, Demiris Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10535-10554.

[13] Wang Y, Miao L, Zhou Z, et al. Infrared and visible Image Fusion with Language-driven Loss in CLIP Embedding Space[J]. arxiv preprint arxiv:2402.16267, 2024.

[14] 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[J]. Information Fusion, 2022, 82: 28-42.

[15] Ma J, Yu W, Liang P, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information fusion, 2019, 48: 11-26.

[16] Chen J, Li X, Luo L, et al. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition[J]. Information Sciences, 2020, 508: 64-78.

[17] Yu M, Cui T, Lu H, et al. VIFNet: An end-to-end visible-infrared fusion network for image dehazing[J]. Neurocomputing, 2024: 128105.

[18] Bijelic M, Gruber T, Mannan F, et al. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11682-11692.

[19] Cui G, Feng H, Xu Z, et al. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition[J]. Optics Communications, 2015, 341: 199-209.

[20] Tang L, **ang X, Zhang H, et al. DIVFusion: Darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91: 477-493.

[21] Li H, Wu X J, Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. Information Fusion, 2021, 73: 72-86.

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[23] Li J, Wen Y, He L. Scconv: Spatial and channel reconstruction convolution for feature redundancy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 6153-6162.

[24] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[25] Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modal imbalance problems[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer International Publishing, 2020: 787-803.

[26] Guo C L, Yan Q, Anwar S, et al. Image dehazing transformer with transmission-aware 3d position embedding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5812-5820.

[27] Hoang T, Zhang H, Yazdani A, et al. Transer: Hybrid model and ensemble-based sequential learning for non-homogenous dehazing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1670-1679.

[28] Ullah, Hayat, et al. "Light-DehazeNet: a novel lightweight CNN architecture for single image dehazing." IEEE transactions on image processing 30 (2021): 8968-8982.

[29] Zheng Y, Zhan J, He S, et al. Curricular contrastive regularization for physics-aware single image dehazing[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5785-5794.