DPGAN: Detail-Fidelity and Perception-Optimized Generative Adversarial Network for Infrared Image Super-Resolution

Jujian Lv, Xin Liu, Zhijie Xie, Kaihan Lin, Wu Li, Zhiqian Wang, Hangyu Zhou, Rongjun Chen, and Jiawen Li, *Member, IAENG*

***Abstract*—Current infrared image super-resolution algorithms based on generative adversarial network (GAN) generally suffer from two significant limitations: the distortion of high-frequency details and the inconsistency between subjective visual perception and objective evaluation metrics. To address these challenges, this paper proposes a novel infrared image super-resolution reconstruction algorithm incorporating detail fidelity constraints and perceptual optimization strategies into a unified framework, i.e., DPGAN. This algorithm constructs a hybrid multi-feature fusion generator, which is jointly propelled by channel-spatial attention mechanisms and is designed to enhance the generator's ability to extract high-frequency detail features from infrared images. In addition, a discriminator network employing the exponential linear unit (ELU) activation function is developed to strengthen its nonlinear representational capacity and improve its discrimination of texture realism in the generated images. A collaborative optimization strategy is further introduced, incorporating perceptual loss, total variation regularization, and structural similarity constraints to build a loss function to align subjective perception with objective evaluation. Comparative experiments conducted on the CVC09 and CVC14 datasets demonstrate that compared to existing methods such as enhanced super-resolution generative adversarial network (ESRGAN) and swin transformer for image restoration (SWINIR), the proposed DPGAN achieves a 2.95 dB improvement in peak signal-to-noise ratio (PSNR) and an average increase of 0.10 in structural similarity index (SSIM). Moreover, the mean opinion score (MOS) reaches 4.32 out of 5, significantly surpassing the baseline methods. Such results indicate that DPGAN can reconstruct infrared images with more realistic and richly detailed textures, which are more consistent with human subjective visual perception.**

***Index Terms*—Infrared images; Super-resolution reconstruction; Generative adversarial networks; Channel spatial attention mechanisms; Hybrid multi-feature fusion; Human subjective visual perception**

I. Introduction

INFRARED imaging technology relies on the principle of thermal radiation. It uses an infrared detector to sense the difference in thermal radiation between the target and the background, facilitating image formation. Unlike traditional visible light imaging, infrared technology offers advantages such as immunity to visible light interference, all-weather imaging, and long-range detection. Thus, it has been widely applied in target detection [1-2], medical imaging [3–4], video surveillance, and remote sensing [5-6]. Nonetheless, environmental and hardware limitations limit the quality of infrared images. From the ecological perspective, atmospheric moisture attenuates energy in the infrared band [7]. Meanwhile, from the hardware standpoint, due to the physical constraints of the detector, it is challenging to increase the density of the focal plane array units further [8]. These factors result in everyday issues such as insufficient detail resolution and loss of fine texture information in infrared images, limiting its broader application in real-world engineering systems.

Improving infrared image quality can be accomplished through two primary approaches. One is hardware optimization through the refinement of optical components and sensor structures. However, due to the physical limitations of focal plane arrays, this approach suffers from high implementation costs and limited scalability due to physical constraints. Another one is the algorithmic approach, which adopts super-resolution techniques to restore fine details and suppress noise to recover high-frequency details from low-resolution

Manuscript received April 19, 2025; revised June 22, 2025. This work was supported in part by the Guangzhou Science and Technology Plan Project under Grants 2024B03J1361 and 2023B03J1327, in part by the University-Industry Collaborative Education Program of Ministry of Education under Grants 241003632084003 and 241005211144020, in part by the Key Discipline Improvement Project of Guangdong Province under Grants 2022ZDJS015 and 2021ZDJS025, in part by the Guangdong Province Ordinary Colleges and Universities Young Innovative Talents Project under Grant 2023KQNCX036, in part by the Guangdong Provincial Higher Vocational Education Teaching Quality and Teaching Reform Project under Grant 2023JG296, in part by the Guangdong Provincial Higher Education Teaching Research and Reform Project under Grant 202430803, in part by the Scientific Research Capacity Improvement Project of the Doctoral Program Construction Unit of Guangdong Polytechnic Normal University under Grant 22GPNUZDJS17, in part by the Graduate Education Demonstration Base Project of Guangdong Polytechnic Normal University under Grant 2023YJSY04002, in part by the Graduate Education Innovation Program of Guangdong Polytechnic Normal University under Grant 2025XJANLK008, and in part by the Research Fund of Guangdong Polytechnic Normal University under Grant 2022SDKYA015.

Jujian Lv is an assistant professor of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: jujianlv@gpnu.edu.cn).

Xin Liu is a postgraduate student of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: asanjin@stu.gpnu.edu.cn).

Zhijie Xie is a postgraduate student of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: xiezhijie@stu.gpnu.edu.cn).

Kaihan Lin is an assistant professor of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (co-corresponding author, phone:+86-020-38256730; fax: +86-020-38257901; e-mail: kaihanlin@gpnu.edu.cn).

Wu Li is an undergraduate student of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: liwu1357@163.com).

Zhiqian Wang is an undergraduate student of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: robot_code2025@163.com).

Hangyu Zhou is an undergraduate student of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: Daydreamer3570@outlook.com).

Rongjun Chen is a professor of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China (e-mail: chenrongjun@gpnu.edu.cn).

Jiawen Li is an assistant professor of School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665 China. (co-corresponding author, phone:+86-020-38256730; fax: +86-020-38257901; e-mail: lijiawen@gpnu.edu.cn).

images and reduce noise. This manner offers practical benefits without requiring additional hardware modifications, cost-effectiveness, and real-time processing capabilities. As a result, it has found wide applications in fields such as medical imaging [9] and satellite remote sensing [10].

The goal of image super-resolution reconstruction [11] is to generate a high-quality, high-resolution image from single or multiple low-resolution inputs. Existing methods are typically classified into three ways: interpolation-based [12], reconstruction-based [13], and learning-based [14]. Interpolation-based methods, such as bilinear, bicubic, and nearest-neighbor interpolation, estimate pixel values based on spatial continuity assumptions. They are computationally efficient but fail to recover high-frequency details lost during degradation, incurring blurred textures in the reconstructed images. Reconstruction-based methods create degradation models to inversely derive high-resolution images by combining low-resolution features with structural constraints, such as Markov random fields and regularization techniques. While these methods usually perform better than interpolation, they heavily depend on a priori knowledge and are often less robust in handling complex scenes and reconstructing complex structures. Learning-based methods use machine learning algorithms to learn implicit priors from data from a large set of training images. This characteristic enables the system to learn image features, which beneficially reconstruct the missing high-frequency details in low-resolution images.

The generative adversarial network (GAN) is presented to enhance the realism of generated images, establishing the adversarial training paradigm of the generator-discriminator paradigm [15]. Its generator produces super-resolution images using residual networks, while the discriminator distinguishes between generated and authentic images based on adversarial loss. This approach is denoted as the super-resolution generative adversarial network (SRGAN), which introduces perceptual loss derived from visual geometry group (VGG) feature space alongside adversarial loss to alleviate the limitations of mean squared error (MSE) loss and enhance perceptual quality, significantly improving the fidelity of texture details. However, it has limitations in handling complex noise scenarios. To address this issue, Wang et al. [16] proposed an enhanced super-resolution generative adversarial network (ESRGAN), which enhances generalization ability through residual scaling and a relativistic discriminator. Moreover, a downsampling module is incorporated to approximate the real-world degradation model better, making the reconstructed network more adaptive to actual noise interference. Previous findings show that ESRGAN improves peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) compared to SRGAN, effectively balancing visual quality and objective metrics.

GANs perform well in visible image super-resolution tasks. However, when directly applied to infrared images, these methods are less effective due to the unique characteristics of infrared imaging, such as dependence on thermal radiation, high noise, low contrast, and low texture contrast or sparsity. On the one hand, thermal radiation affects the infrared images, resulting in sparse high-frequency details and non-uniform noise distributions [17]. Simple single-branch feature extraction networks struggle to separate high-frequency signals from noise, incurring blurred details or amplified noise in the reconstructed images. On the other hand, the existing algorithms using PSNR-based pixel-wise loss functions overly focus on global pixel alignment while disregarding the human visual system's sensitivity to local texture realism [18]. Consequently, while these methods perform well in objective metrics like PSNR, they produce unnaturally smooth images, causing discrepancies between subjective and objective evaluations.

To address the above issues, this paper proposes a detail-fidelity and perception-optimized generative adversarial network, i.e., DPGAN, for infrared image super-resolution. First, constructing a hybrid generator with a multi-level feature fusion network guided by channel spatial attention realizes multi-level feature extraction to strengthen the network's ability to extract high-frequency detailed features. Second, the channel spatial attention mechanism is applied, which can effectively allocate the attention weights of different regions in the low-resolution infrared image so that the network can focus on the high-frequency detail regions of the infrared image and suppress the background noise interference. Next, the exponential linear unit (ELU) activation function enhances the network's nonlinear representational capacity in the discriminator design. It improves the discriminator's ability to assess the realism of texture representations in the generator's infrared images and reduces the perceptual artifacts caused by hallucinated details in human subjective perception. Finally, by including a synergistic optimization mechanism of visual perception loss, total variance regularization, and structural similarity constraints, a loss function system oriented to the consistency of subjective and objective evaluations is constructed, which further reduces the discrepancy between the generated infrared images and the subjective visual perception of human beings accordingly.

II. Proposed Methodology

In this paper, the design of DPGAN, a super-resolution generative adversarial network for infrared images with detail fidelity and perceptual optimization, is carried out based on ESRGAN from the practical problems existing in infrared images. The overall flowchart of the DPGAN is illustrated in Fig. 1. In this section, the main components of the network (generator network and discriminator network), the core modules (improved channel spatial attention module and hybrid multi-feature extraction fusion module), and the loss function are described step by step.

A. Generator Network and Discriminator Network

The generator network continuously optimizes the quality of the generated image, making it asymptotically close to the real image through the adversarial training mechanism. It comprises three main components: a shallow feature extractor, a deep feature extractor, and an upsampling-based reconstruction module.

The shallow feature extractor is a convolutional layer used to extract shallow features from low-resolution infrared images, as expressed:

$$F_{shallow} = F_C(I_{LR}) \tag{1}$$

where I_{LR} denotes the input low-resolution infrared image, F_C is the convolution operation, and $F_{shallow}$ means the extracted shallow features.

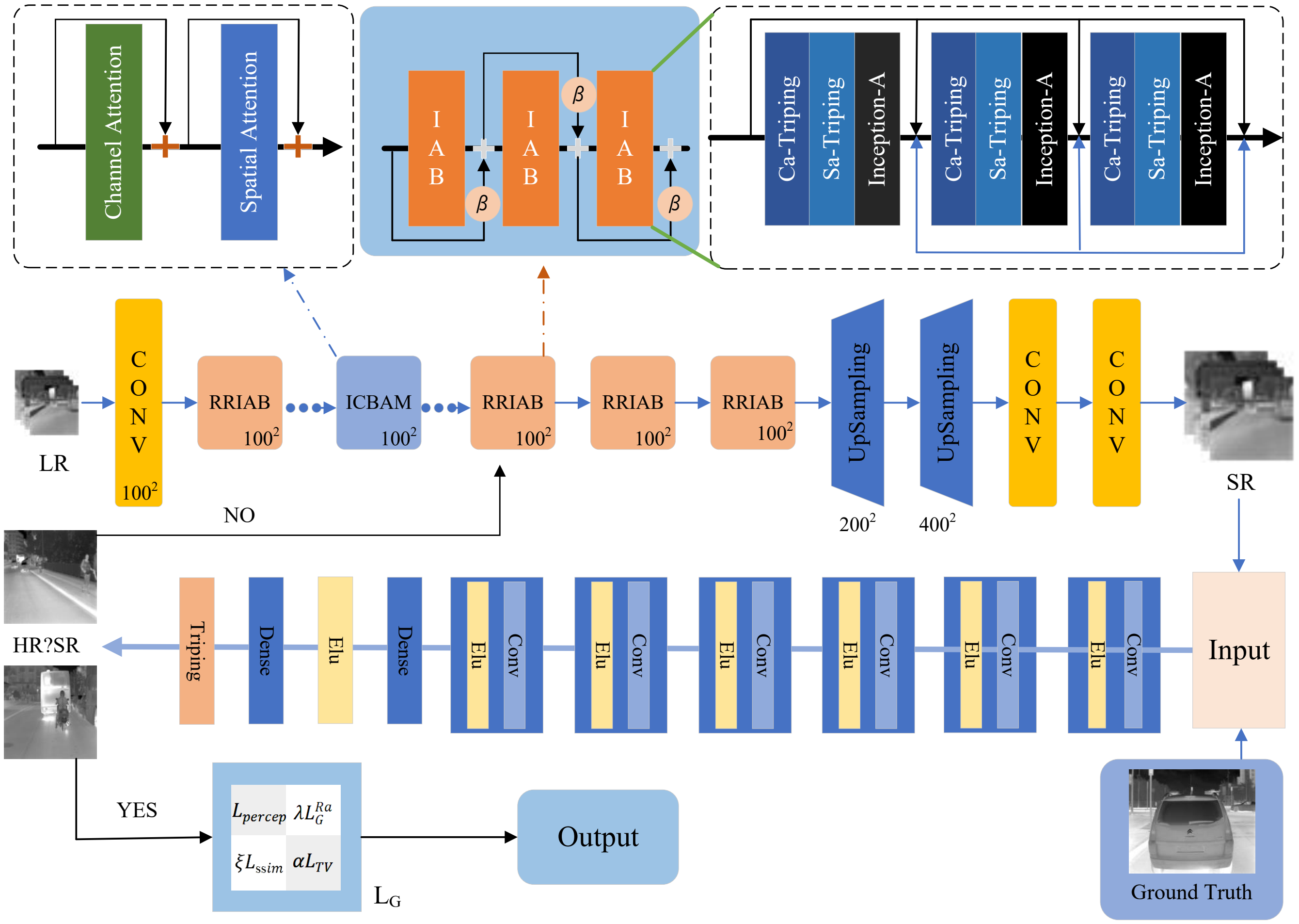

Fig. 1. Overall flowchart of the proposed DPGAN.

After extracting the shallow features, the residual in the residual inception attention block (RRIAB) and improved convolutional block attention module (ICBAM) extract deep features from the infrared images. The deep features are then merged with the shallow features through a long-link operation to obtain the fused features, as expressed:

$$F_{deep} = F^{N}_{RRIAB}(F^{M}_{ICBAM}(F^{N}_{RRIAB}(F_{shallow}))) \quad (2)$$

$$F_{merged} = F_{concat}(F_{shallow}, F_{deep}) \quad (3)$$

where F^{N}_{RRIAB} and F^{M}_{ICBAM} are the N-RRIAB and M-ICBAM blocks employed for extracting the deep features, respectively, F_{deep} denotes the extracted deep features, F_{concat} represents the merging of shallow and deep features, and F_{merged} is the fused features obtained after merging.

The final fused features are processed through two sub-pixel convolutional up-sampling operations. Moreover, two convolutional layers are employed to refine the detailed features of the infrared image, reconstructing a high-quality super-resolution infrared image accordingly.

$$I_{SR} = F^{2}_{C}(F^{2}_{up}(F_{merged})) \quad (4)$$

where F^{2}_{up} represents the two upsampling operations, F^{2}_{C} means the two convolution operations, and I_{SR} denotes the reconstructed super-resolution infrared image.

Subsequently, regarding the discriminator network, its core function effectively discriminates between synthesized and ground-truth images with high precision, which can be regarded as a classifier in this study. Specifically, it receives the generated and corresponding authentic infrared images as input, extracting features from these images followed by ELU activations.

These extracted features are then passed through two dense layers and a final tripping function to obtain the probability of sample classification. In this paper, the discriminator network is based on the adaption from the original ESR-GAN discriminator, in which ELU is used in place of the original LeakyReLU function. Compared to the LeakyReLU, the ELU function enhances the nonlinear representational capacity of the discriminator network, improving the discriminator's ability to capture subtle differences between real and synthetic data distributions. The ELU is as follows:

$$y = \begin{cases} \alpha(e^{x} - 1) & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \quad (5)$$

B. Improved Channel Spatial Attention Module

Preserving high-frequency details remains a core challenge in infrared image super-resolution, leading to blurred textures and degraded perceptual fidelity. Therefore, this paper considers a dual-channel attention module (channel attention and spatial attention) in the generator network based on the theoretical framework of the convolutional block attention module (CBAM) [19], which facilitates the network focuses on high-frequency detail regions in the low-resolution infrared images while ignoring irrelevant areas. In addition, this paper improves the traditional channel and

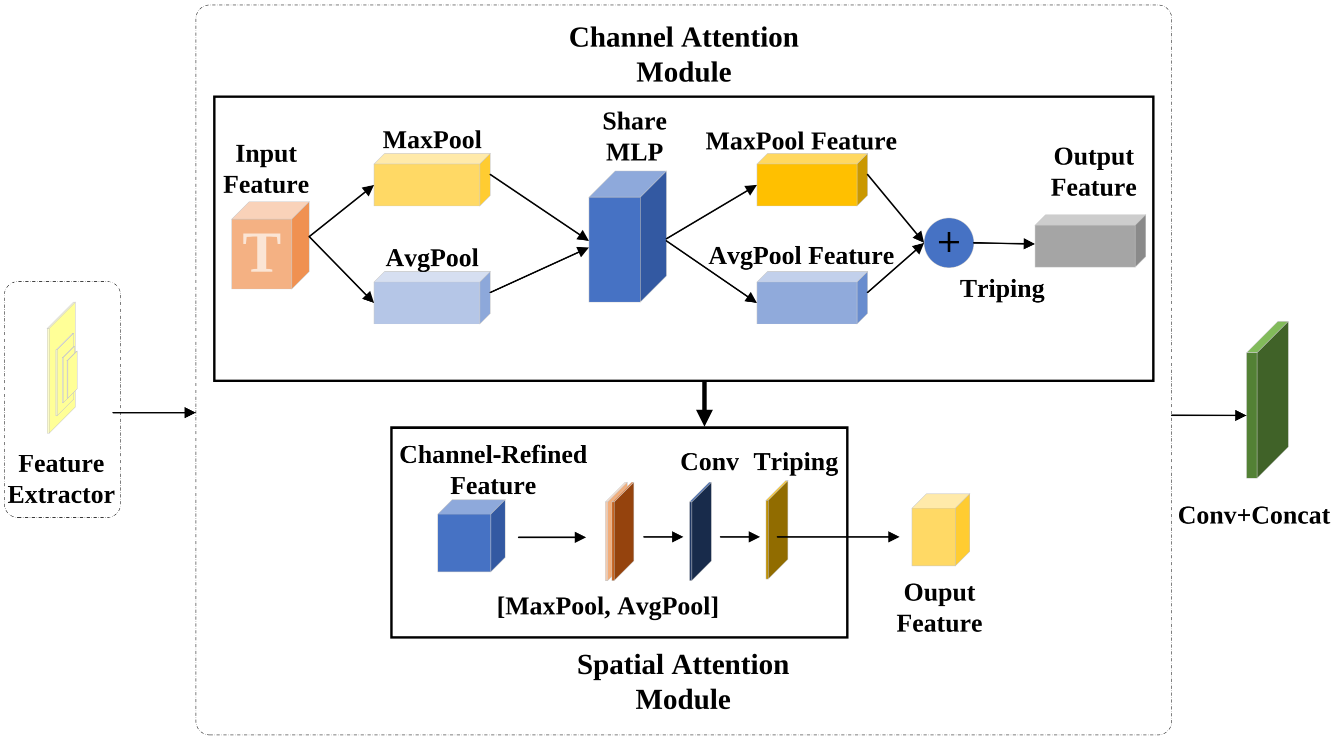

Fig. 2. Improved dual-channel attention module.

spatial attention modules by replacing the Sigmoid activation function in the dual-channel attention module with the Triping function. This results in the new channel attention module (CA_Triping) and spatial attention module (SA_Triping), as shown in Fig. 2.

Here, the proposed improvements to the attention module are guided by two design principles: first, the non-saturation property of the Triping function reduces the risk of gradient vanishing and enhances the convergence speed of the model; second, the sparse activation property of the Triping function effectively inhibits the activation of redundant features, strengthening the network's ability to focus on high-frequency details in infrared image features.

C. Hybrid Multi-feature Extraction Fusion Module

Next, this paper presents a multi-branch feature extraction fusion module, Inception-A [20], into the generator network to address the challenge of improving the network's ability to separate the coupled features of high-frequency signals and noise in infrared images using a relatively straightforward, single-branch network architecture. The Inception-A module has four branches, AvgPooling+1×1Conv, 1×1Conv, 1×1Conv+3×3Conv, and 1×1Conv+3×3Conv+3×3Conv, as depicted in Fig. 3. After the input feature map $T_1(H \times W \times C)$ is processed in parallel through four branches, four feature matrices $k_i(i = 1, 2, 3, 4)$ are obtained, each capturing distinct levels of feature information. These feature matrices are subsequently concatenated through a matrix concatenation operation to generate a new feature map $T_2(H \times W \times C)$.

Inspired by the ESRGAN residual-in-residual structure, this paper innovatively constructs the residual-in-residual hybrid multi-feature extraction fusion module (RRIAB). Specifically, the channel attention module (CA_Triping), spatial attention module (SA_Triping), and Inception-A modules are cascaded to form a hybrid multi-feature extraction fusion module (IAB). Based on that, the IAB module is embedded into the residual-in-residual framework for constructing the RRIAB.

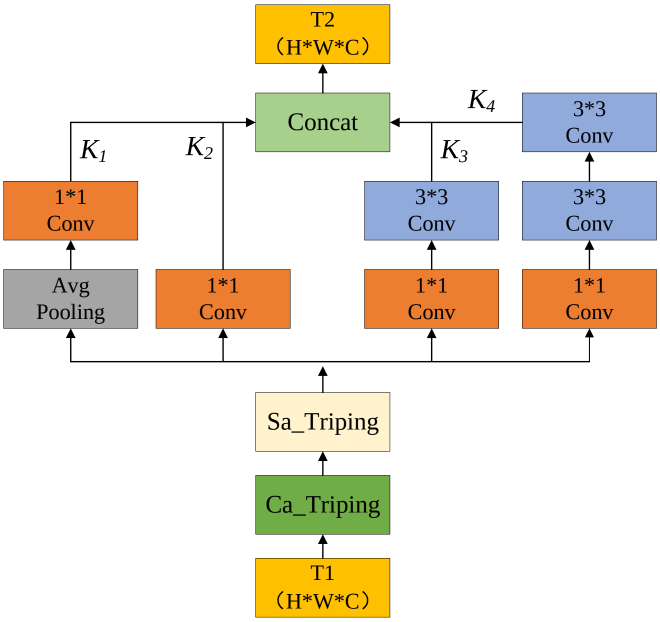

Fig. 3. Hybrid multi-feature extraction fusion module.

D. Loss Function

During the training process of the generative adversarial network, the generator G and the discriminator D are trained against each other through a minimax game. The total loss function of the overall network is:

$$L_{total} = \min_G \max_D (L_G + L_D) \tag{6}$$

where L_{total} is the total loss function, L_G means the generator loss, L_D represents the discriminator loss, $\min_G$ denotes minimization (continuously optimizing the quality of the IR image generated by the generator), and $\max_D$ refers to maximization (continuously improving the discriminator's ability to correctly distinguish between the real IR image and the generated IR image).

The loss function of the generator is:

$$L_G = L_{percept} + \lambda L_G^{Ra} + \xi L_{ssim} + \alpha L_{TV} \tag{7}$$

where $L_{percept}$ denotes the visual perception loss, L_G^{Ra} means the generator confrontation loss, L_{ssim} refers to the structural similarity loss, and L_{TV} represents the total variance loss. Moreover, λ is the coefficient of L_G^{Ra} , which takes the value of 8×10^{-2} in this study. ξ is the coefficient of L_{ssim}. To align the effect of the generated infrared image with subjective visual evaluation, it is set to 1. α is the coefficient of L_{TV}. In the experiment, it was found that when is α greater than 2×10^{-8}, the generated result exhibits a significant color difference compared to the original image. On the other hand, when α is less than 2×10^{-8}, the generated infrared image shows slight speckling. Thus, α is set to 2×10^{-8}.

The visual perceptual loss $L_{percept}$ is utilized to measure the difference between the generated image and the target image in the visual feature space [21]. It evaluates image quality by comparing the features extracted at different layers of a pre-trained network, ensuring that the generated image visually resembles the target image:

$$L_{percept} = \sum_t \frac{1}{N_t} \parallel \phi_l(I_{generated}) - \phi_l(I_{target})\|^2 \tag{8}$$

where $I_{generated}$ refers to the generated image, I_{target} denotes the target image, ϕ_l represents the features extracted by the pre-trained network at layer l , and N_t is the dimensionality of the layer features.

Traditional super-resolution reconstruction algorithms for infrared images usually utilize a content loss L_{pixel} based on pixel-level metrics within the generator loss. Although content loss L_{pixel} ensures pixel-level alignment between the generated and real images, it frequently causes excessive smoothing, degrading the image's subjective quality perceived by humans. To mitigate this issue, this paper introduces the structural similarity loss L_{ssim}, as expressed:

$$L_{ssim}(I_{SR}, I_{HR}) = 1 - SSIM(I_{SR}, I_{HR}) \tag{9}$$

where I_{SR} means the generated super-resolution infrared image and I_{HR} denotes the real high-resolution infrared image.

The total variation loss L_{TV} minimizes the differences between adjacent pixels in an image, reducing artifacts and enhancing image smoothness [22], as expressed:

$$L_{TV} = \int_{D_w} \sqrt{w_x^2 + w_y^2}\, dxdy \tag{10}$$

where D_w refers to the domain of definition of the function w, and w_x and w_y are the partial derivatives of the function w with respect to x and y, respectively.

The generator loss design incorporates a combination of visual perception loss, structural similarity loss, and total variation loss to establish a loss function system that aligns with subjective and objective evaluation criteria. This system effectively minimizes the discrepancy between the generated infrared images and human subjective visual perception.

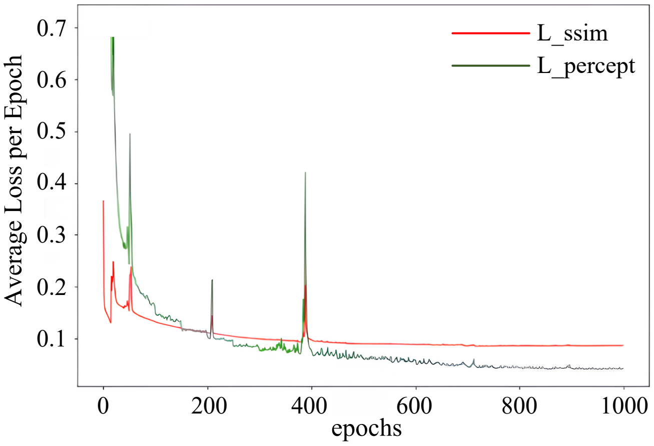

Fig. 4. The changes in training process L_{ssim} and $L_{percept}$.

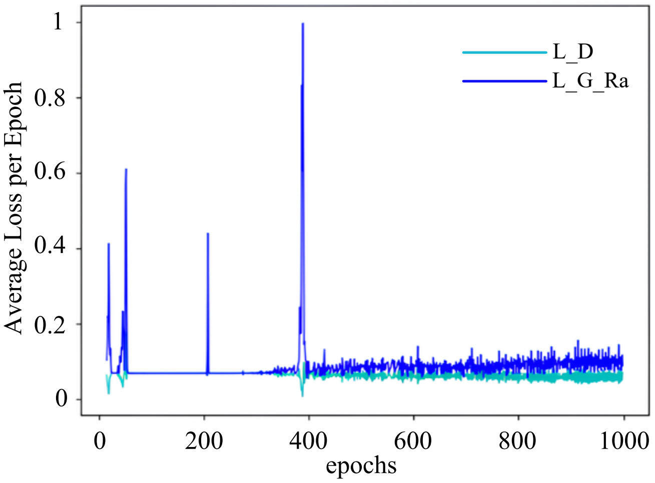

Fig. 5. The changes in training process L_D and L_G^{Ra}.

Generator adversarial loss L_G^{Ra} represents a key component in GAN, realized through adversarial training between the generator and the discriminator, as shown in (11) and (12):

$$L_G^{Ra} = -E_{X_r}[\log(1 - D_{Ra}(X_r, X_f))] - E_{X_f}[\log(D_{Ra}(X_f, X_r))] \tag{11}$$

$$D_{Ra}(x_r, x_f) = \sigma(C(X_r) - E_{x_f}[C(x_f)]) \tag{12}$$

where $x_f = G(x_i)$and x_i denote the input low-resolution infrared images, σ is the activation function operation, and $C(X_r)$ means the untransformed output of the discriminator.

Finally, the loss function of the discriminator is presented in (13):

$$L_D = L_D^{Ra} = -E_{X_r}[\log(D_{Ra}(X_r, X_f))] - E_{X_f}[\log(1 - D_{Ra}(X_f, X_r))] \tag{13}$$

where E_{x_f} represents the averaging operation over all pseudo-data in a mini-batch.

III. Experiment

A. Experimental Details

The datasets used in this paper are the publicly available CVC09 and CVC14 datasets provided by FIR. They consist

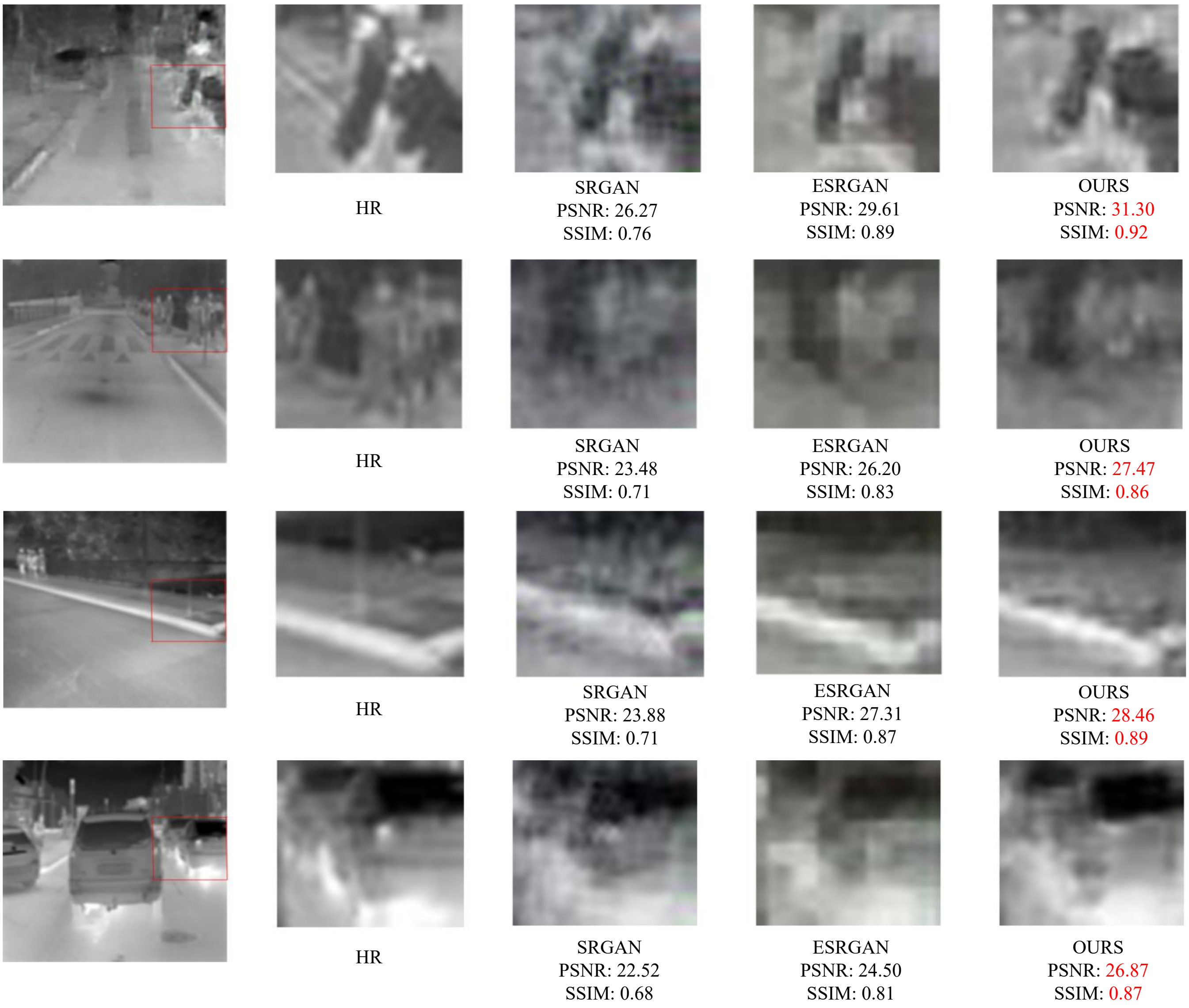

Fig. 6. Comparison of the reconstruction effect using various algorithms.

of two sets of single-channel infrared images: the day set and the night set. These infrared images mainly feature street scenes, covering common targets such as pedestrians, cars, buildings, and roads, making them suitable for studying the super-resolution reconstruction task of infrared images.

The deep learning framework used in this experiment is PyTorch. The experiments were conducted on two NVIDIA GeForce GTX 4090D GPUs. All training images were uniformly cropped to 100×100 pixels for efficient batch processing. Low-resolution infrared images were generated using bicubic downsampling with a fourfold degradation. Before the experiment, 1000 infrared images were randomly selected from the training sets of the CVC09 and CVC14 datasets, respectively, and then shuffled to form a new training set, train-CVC, containing 2000 infrared images. Additionally, 1000 infrared images were randomly selected from the test sets of each dataset to create two new test sets, test-CVC09, and test-CVC14, which were used for performance evaluation in this study. Adam is chosen as the optimizer for both the generator network and the discriminator network, batch_size is set to 32, epochs are set to 1000, and the initial learning rate is set to 0.0004.

B. Evaluation Metrics

In this study, ten RRIAB layers are used within the generator network to enhance deep feature extraction. During training, a learning rate decay strategy is adopted, where the learning rate is multiplied by a decay coefficient of 0.8 every 50 epochs, promoting stable convergence and improved model performance over time. During the training process, for the first 15 epochs, only the L_{ssim} is trained. Starting from the 16th epoch, both the generator loss L_G and discriminator lossare L_D optimized together, enabling the adversarial training between the generator and discriminator to improve the overall performance of the model.

PSNR is an objective metric used to evaluate image quality [23]. Generally, a higher PSNR value indicates less image distortion and better reconstruction. The PSNR is as follows:

$$PSNR = 10 \cdot \log_{10}(\frac{MAX^2}{MSE}) \tag{14}$$

$$MSE = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i,j) - K(i,j))^2 \tag{15}$$

where MAX refers to the maximum possible pixel value of the image, MSE denotes the mean squared error,$I(i,j)$

TABLE I
COMPARISON OF DIFFERENT ALGORITHMS ON OBJECTIVE INDICATORS

Dataset	Evaluation Metrics	SRCNN	ESPCN	SRGAN	ESRGAN	SWINIR	Lite-SRGAN	Ours
test-CVC09	SSIM	0.72	0.74	0.75	0.88	0.88	0.87	**0.91**
	PSNR	23.25	24.19	25.04	27.95	29.98	29.78	**30.02**
test-CVC14	SSIM	0.71	0.70	0.72	0.85	**0.88**	**0.88**	**0.88**
	PSNR	22.93	23.16	23.39	25.89	27.61	27.23	**27.67**

means the pixel value of the target image, $K(i,j)$ is the pixel value of the generated image, and m and n represent the width and height of the image.

SSIM measures the similarity between two images regarding luminance, contrast, and structure [24]. Unlike PSNR, SSIM aligns more closely with human subjective visual perception. The SSIM is as follows:

$$SSIM(x,y) = \frac{(2u_x u_y + C_1)(2\sigma_{xy} + C_2)}{(u_x^2 + u_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (16)$$

where u_x and u_y represent the average luminance of the two images, σ_x^2 and σ_y^2 represent the contrast of the two images, σ_{xy} refers to the covariance between the two images, which measures their structural similarity, and C_1 and C_2 refer to the constants added to avoid a zero denominator.

The mean opinion score is a commonly used method for subjective image quality evaluation [25]. It involves professionally trained evaluators rating image samples on a scale from 1 to 5 (with 1 indicating blurriness and 5 indicating high clarity). The MOS value is calculated as the arithmetic mean of all evaluators' scores. This subjective quality assessment can be further categorized into absolute and relative evaluation methods. In this paper, we use the relative evaluation criteria. As a result, in this paper, PSNR and SSIM are used as objective evaluation metrics, while MOS is used as the subjective evaluation metric for the generated image results.

C. Comparison Results

This paper tests and compares PSNR and SSIM metrics for objective evaluation on the test-CVC09 and test-CVC14 datasets. The performance of SRCNN, ESPCN [26], SRGAN, ESRGAN, SWINIR [27], Lite-SRGAN [28], and the proposed DPGAN algorithm is assessed. As shown in Table I (with bolded values representing the best results), the proposed approach significantly improves both evaluation metrics, particularly in SSIM, which is more aligned with human vision. Compared to SRCNN, ESPCN, and SRGAN, the results indicate that the proposed approach achieves better super-resolution reconstruction of infrared images, as evidenced by the data.

To further validate the proposed DPGAN's effectiveness, two representative images from the test-CVC09 and test-CVC14 datasets are selected for visual comparison with SRGAN and ESRGAN. The results of these experiments are presented in Fig. 6.

TABLE II
COMPARISON OF ALGORITHMS ON SUBJECTIVE EVALUATION METRICS

Algorithms	Evaluation Metrics	Score (in 5-point scale)
SRGAN	MOS	4.18
ESRGAN	MOS	4.24
DPGAN	MOS	**4.32**

The SRGAN algorithm exhibits the lowest overall clarity, with significant distortion and poor visualization. While ESRGAN performs better than SRGAN, it still has blurred edges, suggesting a loss of high-frequency information during reconstruction and a more significant color difference from the original high-resolution infrared image. In contrast, the proposed DPGAN algorithm produces more detailed results, showing effective denoising during reconstruction and better texture and feature details recovery. Such improvements highlight that the algorithm in this paper enhances both objective evaluation metrics and human subjective visual perception.

In the subjective evaluation comparison, SRGAN, ESRGAN, and the proposed DPGAN algorithm are evaluated based on MOS scores for the generated results under the test-CVC09 dataset. The objective evaluations of the infrared images generated by these three algorithms were collected from 50 volunteers, followed by testing multiple discrete samples. The results are presented in Table II (with bolded values representing the best results).

Table II shows that the proposed DPGAN algorithm achieves the highest score of 4.32, followed by ESRGAN with a score of 4.24 and SRGAN with a score of 4.18. Such performance indicates that compared to algorithms like ESRGAN, the infrared images generated by DPGAN are better aligned with human subjective visual perception.

D. Ablation Experiment

To demonstrate the effectiveness of incorporating the ELU activation function and structural similarity loss function in enhancing the model's performance, we remove these improvements from the DPGAN algorithm and revert to the original configurations. The experimental results are then compared between the improved and the original models. The ELU activation function in the DPGAN discriminator is replaced with the original LeakyReLU function to obtain the DPGAN-LeakyReLU model. Similarly, the DPGAN-Pixel model is created by replacing the L_{ssim} in the generator loss function L_G of DPGAN with the original pixel-based loss L_{pixel}. Each model was run under the previously experimental conditions to obtain the evaluation metrics, as listed in Table III.

As shown in Table III (with bolded values representing the

TABLE III
ABLATION EXPERIMENTS FOR ACTIVATION FUNCTION ELU AND LOSS FUNCTION L_{ssim}

Model	test-CVC09		test-CVC14	
	SSIM	PSNR	SSIM	PSNR
DPGAN-LeakyRelu	0.9	28.65	0.87	26.39
DPGAN-Pixel	0.86	27.94	0.82	25.88
DPGAN	**0.91**	**30.02**	**0.88**	**27.67**

TABLE IV
COMPARISON OF ATTENTION MODULE ABLATION EXPERIMENTS

Model	test-CVC09		test-CVC14	
	SSIM	PSNR	SSIM	PSNR
DPGAN-Sigmoid (5 layers)	0.91	29.94	0.88	27.53
DPGAN-Sigmoid (10 layers)	0.91	30.06	0.88	27.67
DPGAN-Triping (5 layers)	0.91	30.02	0.88	27.67
DPGAN-Triping (10 layers)	**0.92**	**30.24**	**0.89**	**27.81**

best results), after removing the improvements of the ELU activation function and loss function L_{ssim}, the PSNR and SSIM values decrease significantly. This proves that both the ELU activation function and loss function L_{ssim} effectively enhance the infrared image reconstruction performance using the proposed DPGAN.

Furthermore, the improvement of the channel attention module and spatial attention module is vital to the algorithm proposed in this paper. The following ablation experiments were designed to verify their effectiveness: The model DPGAN-Sigmoid was obtained by replacing the Triping activation function in CA_Triping and SA_Triping within RRIAB with a Sigmoid function, respectively. The below models were then run under the previously mentioned experimental conditions:

1) DPGAN-Sigmoid with RRIAB layers of 5;
2) DPGAN-Sigmoid with RRIAB layers of 10;
3) DPGAN-Triping with RRIAB layers of 5;
4) DPGAN-Triping with RRIAB layers of 10;

The experimental results are shown in Table IV. As analyzed in Table IV (with bolded values representing the optimal results), DPGAN-Triping and DPGAN-Sigmoid achieve equal SSIM scores using the RRIAB with five layers. However, DPGAN-Triping outperforms DPGAN-Sigmoid in PSNR metrics, indicating that DPGAN-Triping reconstructs infrared images with more realistic textures. At an RRIAB of 10 layers, DPGAN-Sigmoid maintained identical SSIM scores and showed a slight improvement in PSNR, while DPGAN-Triping improved both PSNR and SSIM.

In summary, the improved attention module in this paper demonstrates stronger high-frequency information extraction and more effective feature detail mining compared to traditional attention mechanisms, leading to a better reconstruction of infrared images.

IV. CONCLUSION

This paper proposes a DPGAN, which addresses the issues of high-frequency detail distortion and inconsistency between subjective and objective evaluations in infrared image super-resolution reconstruction tasks. The proposed DPGAN primarily consists of a generator network and a discriminator network. The generator synthesizes high-quality infrared images, while the discriminator progressively guides the generator to produce images that visually approximate real infrared images. Furthermore, to enhance the realism and richness of the texture details in the generated infrared images, a hybrid multi-feature fusion network structure is constructed, guided by channel-spatial attention mechanisms. Lastly, a loss function framework oriented towards the consistency between subjective and objective evaluations is designed, ensuring that the generated infrared images are better aligned with human visual perception. Experimental results demonstrate that the proposed algorithm outperforms comparative methods in objective evaluation metrics and human subjective visual perception. Future research can explore extending the network's depth and width and further optimizing the attention module.

REFERENCES

[1] J. Zhang and Y. Zhang, "Infrared small target detection with UAV based on convolutional neural networks," *Engineering Letters*, vol. 33, no. 5, pp. 1505–1512, 2025.

[2] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 87–119, 2022.

[3] Y. Chai and X. Zhang, "Enhanced chest CT detection of pulmonary nodules based on YOLOv8," *Engineering Letters*, vol. 32, no. 12, pp. 2221–2231, 2024.

[4] J. C. Torres-Galvan, E. Guevara, E. S. Kolosovas-Machuca, A. Oceguera-Villanueva, J. L. Flores, and F. J. Gonzalez, "Deep convolutional neural networks for classifying breast cancer using infrared thermography," *Quantitative InfraRed Thermography Journal*, vol. 19, no. 4, pp. 283–294, 2022.

[5] D. Hou, Y. Zhang, and J. Ren, "A lightweight object detection algorithm for remote sensing images," *Engineering Letters*, vol. 33, no. 3, pp. 704–711, 2025.

[6] L. Li, L. Jiang, J. Zhang, S. Wang, and F. Chen, "A complete YOLO-based ship detection method for thermal infrared remote sensing images under complex backgrounds," *Remote Sensing*, vol. 14, no. 7, p. 1534, 2022.

[7] S. Stoyanov and C. Bailey, "Modeling insights into the assembly challenges of focal plane arrays," *IEEE Access*, vol. 11, pp. 35 207–35 219, 2023.

[8] Y. Zhang, Y. Xu, and J. Zhai, "Diabetic retinopathy image segmentation method based on fusion densenet and u-net network," *Infrared Physics & Technology*, vol. 33, no. 2, pp. 418–428, 2025.

[9] P. Nandal, S. Pahal, A. Khanna, and P. R. Pinheiro, "Super-resolution of medical images using real esrgan," *IEEE Access*, 2024.

[10] C. Wang, X. Zhang, W. Yang, G. Wang, X. Li, J. Wang, and B. Lu, "Mswagan: multi-spectral remote sensing

image super resolution based on multi-scale window attention transformer," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[11] J. Li, Z. Pei, W. Li, G. Gao, L. Wang, Y. Wang, and T. Zeng, "A systematic survey of deep learning-based single-image super-resolution," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–40, 2024.

[12] D. Han, "Comparison of commonly used image interpolation methods," in *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*. Atlantis Press, 2013, pp. 1556–1559.

[13] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4544–4556, 2012.

[14] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[16] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[17] L. Wu and J. Wang, "Features of infrared thermal image and radiation," *Science in China Series D: Earth Sciences*, vol. 41, no. 2, pp. 158–164, 1998.

[18] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *arXiv preprint arXiv:1406.7799*, 2014.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[20] M. A. S. Al Husaini, M. H. Habaebi, T. S. Gunawan, M. R. Islam, E. A. Elsheikh, and F. Suliman, "Thermal-based early breast cancer detection using inception v3, inception v4 and modified inception mv4," *Neural Computing and Applications*, vol. 34, no. 1, pp. 333–348, 2022.

[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[22] H. Nguyen-Truong, K. N. Nguyen, and S. Cao, "SRGAN with total variation loss in face super-resolution," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 2020, pp. 292–297.

[23] A. Tanchenko, "Visual-psnr measure of image quality," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 874–878, 2014.

[24] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2366–2369.

[25] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[26] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SWINIR: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[28] H. S. El-Assiouti, H. El-Saadawy, M. N. Al-Berry, and M. F. Tolba, "Lite-srgan and lite-unet: toward fast and accurate image super-resolution, segmentation, and localization for plant leaf diseases," *IEEE Access*, vol. 11, pp. 67 498–67 517, 2023.