scientific reports

Check for updates

OPEN

Dual decoding generative adversarial networks for infrared image enhancement

Yang Yu[1], Lin Jiang[1], Qijun Hu[2]✉, Qijie Cai[3], Qiang Zeng[1] & Xin Sun[1]

Infrared imaging technology is vital for security monitoring, industrial detection, and medical diagnosis. However, atmospheric thermal radiation degrades its quality, causing contrast reduction, texture blurring, and non-uniform noise. To address these challenges, this paper introduces a novel infrared image enhancement method using a dual decoding generative adversarial network (2D-GAN). First, internal and external skip connections are designed to enhance high-frequency detail transmission and mitigate gradient vanishing in deep networks, with local details being preserved as a result. Second, a cross-layer attention mechanism is proposed to adaptively adjust feature map weights spatially and across channels, with information loss during encoding-decoding being minimized and texture clarity and structural coherence being improved. Finally, a joint loss function is designed to integrate pixel-level accuracy, semantic consistency, and global structural coherence, with image realism and perceptual quality being enhanced consequently. Experiments demonstrate superior performance over existing methods in comparative and ablation studies on public datasets, confirming excellent enhancement capabilities and generalization.

Keywords 2D-GAN, Image enhancement methods, Infrared image, Deep learning, Encoder–decoder network

In recent years, with the widespread application of infrared images in fields such assecurity monitoring, industrial detection and medical diagnosis, their significance has become increasingly prominent. Infrared images are formed by infrared scanners that receive and record the thermal radiation emitted by target objects. As a passive imaging technology, infrared imaging relies on the thermal radiation characteristics of the target object. Compared with conventional optical images, infrared images exhibit a smaller grayscale variation range and are not linearly related to the target's reflective properties. They generally have lower resolution and contrast. Additionally, images captured by infrared sensors often contain strong clutter noise, resulting in a low signal-to-noise ratio (SNR) and degraded visual quality. During the image acquisition process, factors such as atmospheric interference in thermal radiation transmission, differences in the detection range of imaging devices, and variations in the temperature range of target objects[1–4] contribute to common issues in infrared images, including insufficient brightness, low SNR, and blurred details[5,6]. These challenges significantly impact subsequent processing tasks such as feature extraction, image segmentation, and object detection. As a result, infrared images typically require enhancement to improve image quality and visualization effects, ensuring more accurate and effective image analysis. Image enhancement technology aims to adjust image intensity values to improve visual quality, enhance contrast, and reduce noise. However, image enhancement is inherently subjective: if an image has low contrast, its intensity must be increased; conversely, if the image has high contrast, its intensity should be reduced to achieve proper visualization. In general, infrared images tend to have low contrast, making intensity enhancement a necessary step for improving their visual representation[7].

In recent years, deep learning-based methods have achieved remarkable progress in infrared image enhancement, with Generative Adversarial Networks being widely adopted[8]. Through adversarial training, GANs can learn the distribution of infrared image features and generate enhanced images with higher quality. However, existing infrared image enhancement methods still face several challenges, such as loss of fine details, over-enhancement, and increased noise[9–13]. For instance, GAN-HA[14] introduces a heterogeneous dual-discriminator network to simultaneously learn the thermal radiation information of infrared images and the texture details of visible images. However, this method relies on large-scale training data, making it less

[1]School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, Sichuan, China. [2]School of Civil Engineering, Southwest Jiaotong University, Chengdu 611756, Sichuan, China. [3]School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu 610500, Sichuan, China. ✉email: huqijunswpu@163.com

effective in small-sample scenarios. Additionally, the lightweight GAN-based image fusion algorithm proposed by Wu et al.[15] has made notable progress in computational efficiency, enabling more effective deployment of infrared image enhancement in embedded systems. Nonetheless, its generalization ability remains limited when dealing with complex backgrounds and varying illumination conditions. Another lightweight adversarial-based infrared image enhancement method[16] enables efficient enhancement through edge-side deployment. However, its performance in enhancing target details in high-brightness regions is still suboptimal, leading to a loss of fine details in some bright areas, ultimately affecting image quality.

To address the limitations of existing approaches, this paper proposes a novel Dual Decoding Generative Adversarial Network (2D-GAN) to improve infrared image enhancement quality. The proposed method integrates internal and external skip connections, a cross-level attention mechanism, and a joint loss function to enhance feature transmission and optimize perceptual quality. The dual decoding structure enables the network to better capture both global information and local details of infrared images, facilitating a more refined enhancement process. The internal and external skip connections effectively mitigate information loss during encoding-decoding by preserving multi-scale feature consistency. Furthermore, the cross-level attention mechanism adaptively allocates positional weights across feature maps, optimizing enhancement in different regions and ensuring a more natural appearance of the generated images while preventing over-enhancement or detail loss. To further improve perceptual quality, the method incorporates a joint loss function that balances feature representation at multiple scales, ensuring both fine-grained texture details and overall structural coherence. Overall, this work makes four key contributions to infrared image enhancement:

- A double decoding generative adversarial network is introduced to enhance infrared image features across multiple scales, leveraging internal and external skip connections for improved detail preservation.
- A cross-layer attention mechanism is developed to adaptively weight spatial features in the feature map, facilitating region-specific enhancement while enhancing texture clarity and structural coherence.
- A joint loss function integrating WGAN-GP, MSE, VGG and Efficient Net is formulated to enhance image realism and perceptual quality through pixel-level accuracy, semantic consistency, and global structural coherence.
- A cross-scenario validation framework is established, with comparative experiments and ablation studies on public datasets confirming superior performance and generalization over existing methods.

The remaining sections are organized as follows: In section "Related work", we review the relevant literature and analyze the limitations of current infrared image enhancement techniques. In section "Methods", we introduce the implementation of 2D-GAN in detail, including the network architecture and loss function design. In section "Experimental results and analysis", we describe the experimental dataset and experimental environment, and evaluate the performance of 2D-GAN through a series of experiments, while comparing it with existing methods. In section "Discussion", we discuss the challenges encountered during the experiment, analyze the limitations of this method, and explore future research directions. Finally, in section "Conclusion", we conclude this paper.

Related work

Overview of methods

Traditional infrared image enhancement methods mainly rely on histogram equalization (HE), which improves image contrast by adjusting pixel value distribution. Although this method enhances image visibility to some extent, its global adjustment strategy often leads to detail loss, noise amplification, and over-enhancement. To overcome these limitations, researchers have proposed various improvements in recent years.

For example, Wang et al.[17] proposed an infrared image enhancement method based on multi-scale multi-rectangular fusion. This method extracts image features at different scales through multi-scale mapping, enriching texture details and improving scene adaptability. Rivera-Aguilar[18] adopted differential evolution to optimize histogram equalization, achieving adaptive pixel distribution adjustment and effectively avoiding oversaturation. Lu[19] combined normal histogram matching with sharpening processing to improve brightness, contrast, and sharpness; however, its enhancement effect is limited by the selection of the reference histogram, resulting in low adaptability. Although these methods alleviate some of the shortcomings of HE, they still rely on global transformation strategies, making it difficult to achieve fine-grained local adaptive adjustments in complex infrared scenes, which may lead to the loss of local details or excessive enhancement.

Regarding detail enhancement in infrared images, Branchitta et al.[20] proposed a dynamic range segmentation algorithm based on bilateral filtering (BF) in 2009, laying the foundation for subsequent layered filtering enhancement methods. In recent years, researchers have further optimized infrared image enhancement by integrating adaptive filtering, multi-scale fusion, and Retinex theory. For instance, the adaptive guided filtering method combined with global-local mapping[21] enhances local details and optimizes global contrast, effectively improving image clarity. However, in complex scenes, this method may lead to over-enhancement and is highly sensitive to filtering parameters, limiting its generalization ability.

To further enhance contrast stability, a method combining multi-scale guided filtering and contrast-limited adaptive histogram equalization (CLAHE)[22] decomposes images into different detail levels and applies CLAHE to the base layer to enhance overall brightness and contrast. While this method performs well in low-contrast image enhancement, it may cause excessive contrast in local regions, affecting visual consistency. Moreover, Song et al.[23] proposed an adaptive histogram equalization method combining guided filtering and Gaussian filtering. This method first decomposes the original infrared image into a base layer and a detail layer using guided and Gaussian filtering, then applies adaptive histogram equalization to the base layer to improve overall brightness and contrast. For the detail layer, guided filtering is used for regularization, and a detail gain function is constructed to enhance detail information. Finally, the base and detail layers are fused to obtain the enhanced

infrared image. Although this method achieves effective detail recovery, it may still result in detail smoothing and is highly dependent on parameter selection, limiting its adaptability in complex scenarios.

Additionally, the multi-scale Retinex and sequential guided filtering method[24] incorporates the Retinex model for multi-scale processing and employs sequential guided filtering to enhance details while suppressing noise, thereby improving the visual quality of infrared images. However, the non-linear processing of Retinex may lead to detail smoothing, and its strong dependence on parameter selection limits its adaptability. Although filtering-based enhancement methods have made significant progress in detail recovery and contrast optimization, they still face challenges such as noise sensitivity, parameter dependence, and localized over-enhancement.

In recent years, deep learning technology has achieved remarkable advancements in infrared image enhancement. Unlike traditional methods, deep learning approaches learn image enhancement mappings in an end-to-end manner, enabling adaptive adjustment strategies that achieve more precise detail preservation and noise suppression. For instance, Cheng[25] proposed a lightweight adversarial generative network that enhances images through multi-layer feature fusion and multi-scale loss optimization while significantly reducing computational complexity. Tang[26] developed an edge gradient enhancement method that substantially improves gradient features in infrared and visible image fusion. Xie[27] incorporated Retinex-Net to develop a low-light infrared image enhancement method, effectively improving overall image quality. However, deep learning-based methods still face challenges such as data scarcity and high computational costs.

Generative adversarial network

Since their introduction in 2014, generative adversarial networks (GANs) have achieved remarkable progress in various image processing and computer vision domains[28]. In recent years, their applications have expanded significantly, demonstrating enhanced capabilities in medical imaging, image fusion, image restoration, image generation, image colorization, and super-resolution[29].

In medical imaging, GANs have been employed to generate synthetic medical images, addressing challenges such as patient privacy protection and data scarcity. For instance, they can produce realistic MRI and PET scans, improving the training of diagnostic models[30]. In the field of image fusion, researchers have developed a lightweight GAN based algorithm that integrates the convolutional block attention module (CBAM) and Depthwise Separable Convolution (DSConv), enabling efficient fusion of visible and infrared images for real-time applications on embedded devices[15]. In satellite image processing, GANs have been applied to super-resolution tasks, learning from high-resolution data to reconstruct fine details in low-resolution images, thereby enhancing clarity and information richness[31]. In the realm of super-resolution, the Photo-Realistic Super-Resolution GAN (SRGAN) has been widely used to upscale low-resolution images while preserving intricate details, significantly improving image quality[32]. For image restoration and inpainting, models such as the Contextual Attention GAN have demonstrated strong capabilities in reconstructing damaged images with high fidelity[33].

GANs have also played a key role in image-to-image translation. Conditional GANs (cGANs) learn transformation mappings between different domains, enabling effective style transfer[34]. In text-to-image synthesis, GAN based models have been used to generate semantically coherent images from textual descriptions. Reed et al.[35] proposed an early model for this task, while Zhang et al.[36] introduced a conditional GAN capable of generating clear images under adverse weather conditions, such as rain and snow. Another approach by Zhang et al.[37] utilizes a stacked GAN model to produce high-resolution images based on textual descriptions. Unsupervised image enhancement has also benefited from GANs, as demonstrated by Ni et al.[38], who developed a model that learns image-to-image mappings without requiring supervised training data.

GANs have further advanced image colorization. Shafiq and Lee[39] proposed a method that combines a transformer-based architecture with a GAN framework, effectively capturing global image context and enhancing visual quality. In medical image synthesis, Ju et al.[40] introduced the Hybrid Augmented Generative Adversarial Network (HAGAN), which preserves structural textures and tissue details. This model incorporates an Attention Mixed Generator, a Hierarchical Discriminator, and a Reverse Skip Connection between the Discriminator and Generator.

Overall, these developments underscore the transformative impact of GANs on contemporary image processing and computer vision, demonstrating their versatility and efficacy across a wide spectrum of applications. To elucidate the architectural evolutions that have driven these advances, we now review both generator and discriminator designs.

Generator

The generator serves as the principal component of a generative adversarial network (GAN), with its architectural design decisively impacting the quality and diversity of generated images. Early implementations relied on multilayer perceptrons (MLPs), which were insufficient for synthesizing high-resolution content. This limitation motivated the transition to deep convolutional neural networks (DCNNs), which exhibited marked improvements in image fidelity. In recent advancements, several novel generator frameworks have emerged, StyleGAN2[41] introduces path length regularization and an expanded network capacity to ensure stable convergence and enhanced sample diversity at high resolutions; U-GAT-IT[42] integrates adaptive layer-instance normalization with attention mechanisms, enabling unsupervised image-to-image translation models to emphasize semantically critical regions and refine textural details; Real-ESRGAN[43] employs purely synthetic training data augmented by higher-order degradation modeling and embeds a U-Net–based discriminator alongside spectral normalization within the generator, thereby achieving superior performance in real-world super-resolution scenarios; TransGAN[44] provides the first demonstration that a wholly transformer-based generator can accomplish high-fidelity, high-resolution image synthesis, establishing a new paradigm for generator design. These cutting-edge architectures collectively offer valuable guidance and theoretical underpinnings for the development and optimization of the W-GAN generator in the present study.

Discriminator

The discriminator serves as a fundamental component of GANs, exerting a decisive influence on adversarial training stability and the quality of generated images. Liu et al.[45] developed a Polarization Image Quality Discriminator that performs weighted evaluations across different polarization channels to guide the fusion network's training, thereby significantly enhancing contrast and preserving edge details in the fused output. Cong et al.[46] introduced a Dual-Discriminator architecture that imposes adversarial constraints on style and content separately, effectively improving the realism and aesthetic quality of the enhanced images. Song et al.[47] proposed a Triple-Discriminator framework, wherein local patches from infrared images, visible-light images, and their difference images are independently classified; this design accentuates differential information to extract fine-grained details and salient object contours, markedly boosting clarity and texture retention in the fusion results. Zhang et al.[48] devised Dual Markovian Discriminators that independently assess patches from infrared and visible modalities and jointly train with the generator to estimate and optimize both distributions; this method obviates the need for manual rule design, adapts end-to-end to cross-modal features, and achieves lower parameter counts and computational complexity compared to fuzzy logic–based approaches, rendering it more suitable for real-time infrared enhancement. In contrast, fuzzy-logic–based discriminators address uncertainty by embedding membership functions and rule-based inference within the network. However, these rule bases are typically handcrafted from expert knowledge and are not amenable to joint end-to-end optimization with network weights, thereby impairing the model's ability to generalize to complex imaging scenarios[49]. As a result, such discriminators frequently exhibit excessive smoothing, loss of fine structural details, and substantial computational overhead, which limits their applicability to real-time infrared image enhancement.

Attention mechanisms in image enhancement

In recent years, with the rapid advancement of deep neural networks in image processing, attention mechanisms have emerged as a pivotal technique for enhancing image enhancement performance due to their ability to adaptively weight salient regions in feature maps while suppressing redundant information. Xu et al.[50] proposed an underwater image enhancement method based on cross-attention, which establishes mutual attention between different channels and spatial positions to adaptively focus on critical visual cues, significantly improving the quality of low-contrast and color-distorted underwater scenes. Zhou et al.[51] designed a Pixel-Weighted Channel Attention Module (PCAM) that captures inter-channel dependencies and adaptively recalibrates channel features based on the degree of image degradation, substantially enhancing detail preservation and visual fidelity in image restoration. Chen et al.[52] introduced a multi-attention framework grounded in non-local attention, embedding the non-local means concept into convolutional neural networks to adaptively capture long-range contextual information via weighted aggregation of global pixel pairs, thereby markedly improving detail recovery and visual quality in multi-exposure low-light images. Dong et al.[53] proposed a shared-weight attention mechanism in a CNN–graph attention network for hyperspectral image classification, wherein identical weight vectors are utilized across all graph attention layers to compute inter-node attention coefficients, significantly reducing parameter counts and enhancing computational efficiency. Concurrently, Wang et al.[54] incorporated Multi-Head Self-Attention (MHSA) within the LeWin Transformer module of Uformer, enabling more effective capture of long-range dependencies and multi-scale features in image denoising, deraining, and deblurring tasks, resulting in notable improvements in restoration quality. Collectively, these attention mechanisms, by dynamically emphasizing critical information and suppressing noise, have demonstrably elevated the representational capacity and generalization ability of enhancement models across diverse tasks, thereby providing robust support for image enhancement in underwater, infrared, polarization, hyperspectral, and multi-exposure low-light scenarios.

Methods

Basic principles of generative adversarial networks

The basic principle of generative adversarial networks[55] (GAN) is to train two models through games: generator and discriminator. The generator and discriminator train and optimize each other during the game. The generator receives random noise as input and generates samples through a series of transformation and mapping operations. The discriminator receives the samples generated by the generator and the real samples and outputs the probability value of judging the sample to be a real sample. The goal of the generator is to generate as realistic samples as possible to fool the discriminator so that it cannot accurately distinguish between generated samples and real samples, while the goal of the discriminator is to judge the authenticity of the samples as accurately as possible. The two compete with each other, learn from each other, and adjust their parameters to improve performance by constantly alternating optimization training, until the distribution of samples generated by the generator is close enough to the distribution of real samples, and the discriminator cannot effectively distinguish them. The basic structure diagram of GAN is shown in Fig. 1.

As shown in Fig. 1, random noise is input to the predefined generator model for training. During the training process, the generator G will generate a fake image $G(z)$ similar to the real image distribution. The generated $G(z)$ is input into the discriminator$D$ together with the corresponding real image x. The discriminator model will identify the two images and give a judgment result. The optimization goals of the generator model and the discriminator model confront each other, and their losses can be expressed as formula (1):

$$\min_{G}\max_{D} V(G,D) = E_{x\sim P_{data}(x)}[logD(x)] + E_{z\sim P_{noise}(z)}[log(1-D(G(z)))] \tag{1}$$

among them, E is the expected value of the distribution function, x is the real sample, z is random noise, $P_{data}(x)$ is the distribution of the real sample, $P_{noise}(z)$ is the distribution of the noise, and $G(z)$ is the sample generated

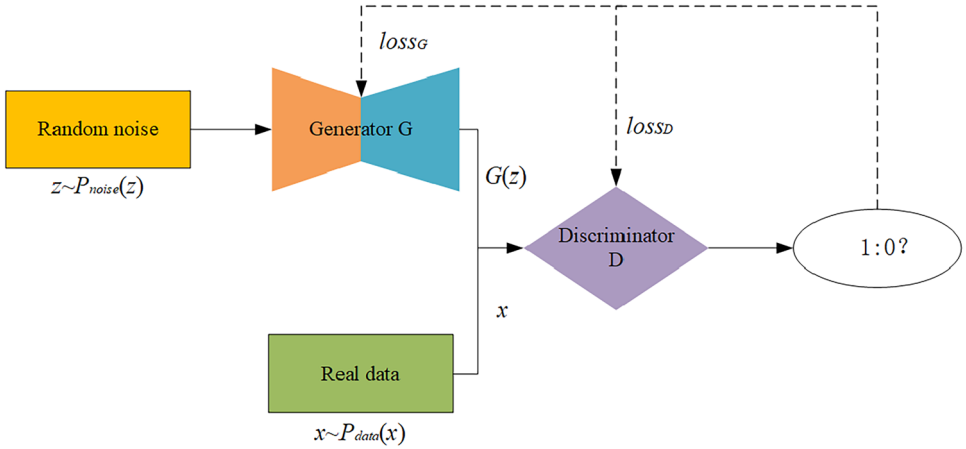


Fig. 1. Generative adversarial network structure.

by the generator network. $D(x)$ is the probability that the discriminator network determines that the real sample is true, and $D(G(z))$ is the probability that the discriminator network determines that the generated sample is true.

Image enhancement network based on 2D-GAN

To enhance infrared images, this chapter designs a dual-decoding generative adversarial network (2D-GAN), whose network structure is shown in Fig. 2. The generation network of 2D-GAN consists of a basic encoding-decoding network, internal skip connections, external skip connections, and cross-level attention modules. The discriminative network adopts a Markov discriminator containing four convolutional layers.

In Fig. 2a, the input of the 2D-GAN generation model is an infrared image of size 128×128. In the first encoding stage, convolution and batching with a kernel size of 4×4 and a step size of 2 are mainly used. Normalization and Leaky ReLU (Leaky Rectified Linear Unit) activation are used to extract features. The decoding stage uses transposed convolution with a kernel size of 4×4 and a stride of 2, along with batch normalization and Leaky ReLU activation. Internal skip connections are used to receive feature information from the encoding stage, thereby reducing information loss in this process.The operation of the second stage is similar to the first stage, but it takes the intermediate image output by the first stage as input. It then encodes this image and integrates the feature information of the first stage through external skip connections during encoding. During decoding, the second stage uses cross-level attention to weigh the coding features of the first stage. It does so by using both the corresponding coding feature information from the first stage and the feature information obtained by decoding from the previous layer. These features are then integrated with the coding features of the second stage to complete the decoding process.

In the generative model of the 2D-GAN network, more comprehensive information is captured through the encoding process, and this information is gradually restored during the decoding stage, which helps the network better understand the relationship between different areas in the image and can also effectively improve The network's ability to understand and enhance image details. Compared with one decoding, two decodings can better reconstruct the fine structure and edge details of the image, thereby improving the quality and clarity of the image. In general, the network structure used in this article can provide stronger feature expression capabilities and better detail retention capabilities, which helps the model achieve better performance and results in image processing tasks.

The discriminative model structure used by 2D-GAN is shown in Fig. 2b. The real and generated images of size 128×128 are converted into block outputs of size 8×8 through 5 convolutional layers. The first 4 convolutional layers use a kernel size of 4×4 and a stride of 2, with batch normalization and Leaky ReLU activation. The last layer uses a stride of 1. The discriminative model divides the generated image and the real image into multiple overlapping patches and performs independent evaluation and discrimination on each patch. By comparing the similarity of each patch, the discriminative model can discover inconsistencies and errors in local areas and punish them. Compared with the global discriminator, the Markov discriminator used in this article pays more attention to the local details and structure of the image and can evaluate the authenticity of the generated image in more detail, allowing the generative model to generate higher-quality images.

Internal skip connections

In the encoder–decoder network, the skip connection plays the role of transferring the feature information extracted by the low-level network to the high-level network[56]. It enables the network to retain the high-resolution information of the image and helps to enhance the image texture and some details.

The inner skip connection designed in this paper aims to solve the problem of loss of detailed information and the problem of gradient vanishing in the infrared image enhancement task. By establishing a direct connection between the encoder and the decoder, the inner skip connection realizes the reuse of feature information, which

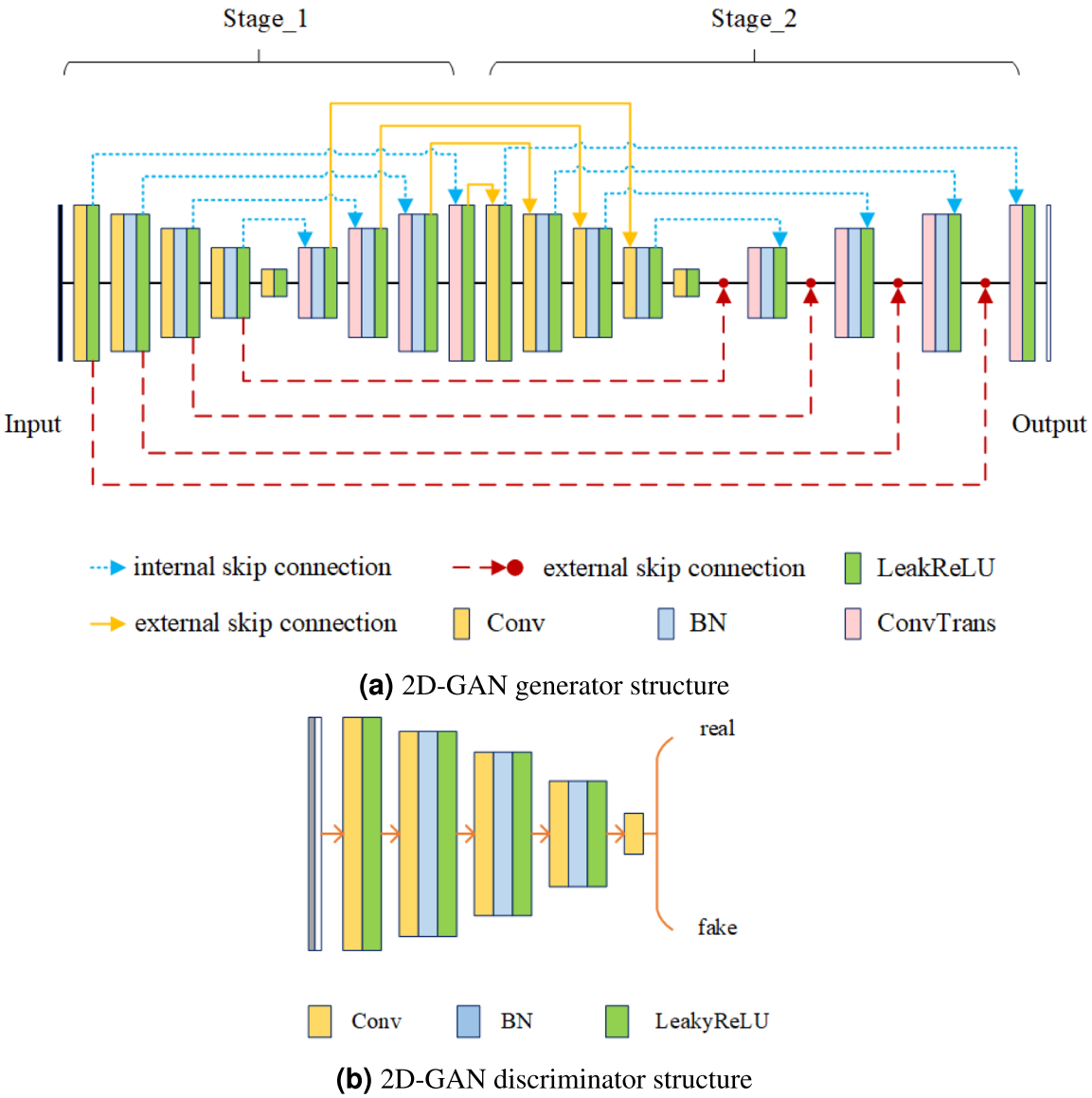


Fig. 2. 2D-GAN network structure.

allows the detailed features of the lower layers in the network to be transferred to the higher layers and fused with the semantic information of the higher layers, thus improving the overall performance, expressiveness, and generative effects of the network. For the network, skip connections can be regarded as an information shortcut, allowing gradients to be propagated directly back to the lower-level network, avoiding the problem of gradient disappearance during network training. This enables the 2D-GAN network to stack layers deeper to better capture the high-level semantics of images.

Specifically, in the network decoding stage, the inner skip connection passes the feature map f_e^i of the same resolution corresponding to the encoding stage to the decoding stage. The feature map f_d^{i-1} from the previous decoding stage is processed by transposed convolution, batch normalization, and activation function to obtain a feature map with doubled resolution. This feature map is then concatenated with f_e^i to obtain the feature map $f_d{}^i$ of this decoding stage.This process can be expressed by formula (2):

$$f_d^i = Concat(ReLU(BN(ConvT(f_d^{i-1}))), f_e^i) \tag{2}$$

among them, $ConvT(\cdot)$ is the transposed convolution operation with a kernel size of 4, a stride of 2, and a padding of 1, which doubles the size of the feature map. *BN* is the batch normalization operation with a momentum parameter of 0.8. The activation function is denoted as *ReLU*, and *Concat* represents the concatenation of feature maps along the channel dimension.

Internal skip connections play an important role in the main architecture of the network. They enable the network to have better context awareness, attend to both low-frequency structural information and high-

frequency detail information in the image, promote image generation and detail recovery, and help improve the quality and accuracy of image enhancement.

External skip connections

The external skip connections transfer characteristic information from the decoder in the previous encoding-decoding network to the encoder in the subsequent encoding-decoding network. This approach improves the encoder's understanding and representation of input data while reducing the risk of information loss during encoding. Additionally, it helps maintain a high-quality feature representation, enhancing the network's perceptual ability and image generation quality.

As shown in Fig. 3, the external skip connection performs channel dimensionality reduction on the features f_i^{d1} from the first decoding stage and passes them to the second encoding stage. These features are then compared with the corresponding features from the second encoding stage at the same depth. After integration and downsampling, the next-stage encoding feature f_{i+1}^{e2} is obtained.

This process can be expressed as formula (3):

$$f_{e_2}^{i+1} = conv_2(ReLU(BN(conv_1(concat(conv_1(f_{d_1}^i, f_{e_2}^i)))) \tag{3}$$

where $f_{d_1}^i$ is the feature at depth *i* in the first decoding stage. $conv_1$ is the convolution operation that preserves the spatial size of the input feature map while reducing the number of channels by half. $conv_2$ is the convolution operation that reduces the spatial size of the input feature map by half.

In the architecture of the generative network, the introduction of external skip connections helps to optimize the training and generation effects of the network. By transferring the feature information of the decoder in the first stage to the encoder in the later stage, the network can not only obtain the high-level semantic information from the first decoding stage, but also, when combined with the inner skip connection in the first encoding-decoding network, extract detailed feature information from one encoding stage in the third stage. In this way, the integration of information from multiple parties allows the network to utilize richer feature representations to generate more accurate and detailed image results. This connection method also helps alleviate the vanishing gradient problem of the network and accelerates model convergence.

Cross-level attention

In infrared image enhancement tasks, over-enhancement often occurs in the pursuit of higher quality infrared images. Over-enhancement refers to the over-processing of the image, resulting in sharpening, unnaturalness, color distortion, and other problems in the enhanced image. To avoid over-enhancement, this paper designs a cross-level attention module. By adding an attention module to the network, the model can automatically learn the weight of each position in the feature map, thereby processing different areas to varying degrees during the enhancement process and helping the model focus more on retaining the naturalness and coherence of the image.

Under the adjustment of the attention mechanism, the model can avoid excessive enhancement of these areas by giving lower weights to the existing brightness and details in the image, thus avoiding sharpening, serious exposure, and other artifacts. In addition, the cross-level attention module can achieve a weighted fusion of shallow network features and deep features, which improves the expressive ability of the model and the quality of the generated results, allowing the network to better capture the semantic information and detailed features of the image.

The structure of the cross-level attention module is shown in Fig. 4. Its two inputs are the first encoding stage feature $f_{e_1}^i$ and the second decoding stage feature $f_{d_2}^{i-1}$. The output is the weighted feature $\hat{f}_{e_1}^i$.

In Fig. 4, convolution and batch normalization operations with a kernel size of 1×1 are performed before the feature maps are added. *Sigmoid* represents a function that maps the data to a value between 0 and 1. After weighing the features, it is still necessary to combine the feature information from the second encoding stage to ensure that the detailed features of the image are effectively enhanced. It is worth noting that in the second

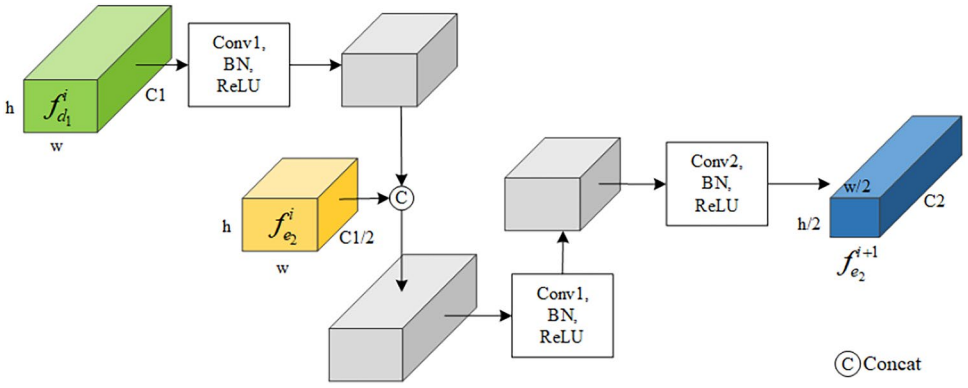


Fig. 3. External skip connection structure.

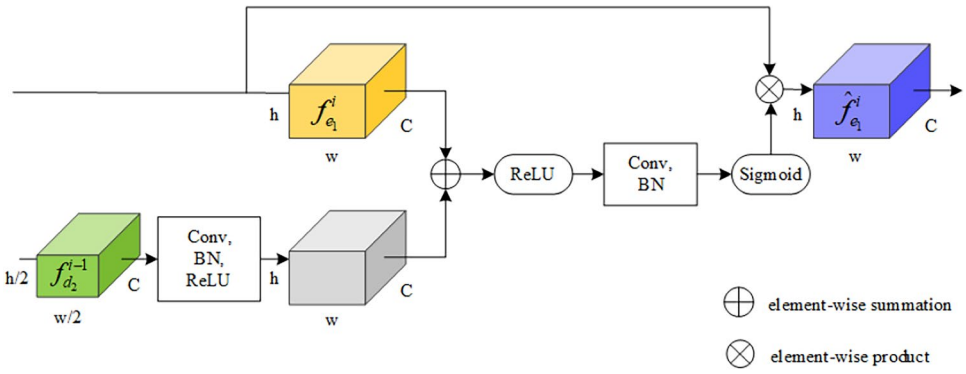


Fig. 4. Cross-level attention mechanism structure.

decoding stage, the target of cross-level attention is the feature from the first encoding stage. This is equivalent to the network decoding the features from the first encoding stage twice, and the second decoding also incorporates information from the first decoding stage. The advantage of this is that it can prevent feature loss in multiple rounds of encoding and decoding and help the network obtain higher-level semantic information.

Loss function

The proposed 2D-GAN network utilizes multiple loss functions to optimize the image generation process and achieve high-quality image enhancement. Specifically, the loss functions include Wasserstein GAN with Gradient Penalty (WGAN-GP), Mean Squared Error (MSE) loss, Visual Geometry Group (VGG) loss, and EfficientNet loss. By comprehensively considering the contributions of these loss functions, the quality of the generated images can be effectively improved.

Adversarial losses include generation loss and discriminative loss, which employ adversarial training to encourage the generator to generate more realistic and visually coherent images. Unlike Least Squares GAN (LSGAN)[57], our method adopts Wasserstein GAN with Gradient Penalty (WGAN-GP)[58], which improves training stability by enforcing the Lipschitz constraint. The generator G aims to generate realistic enhanced images, while the discriminator D is optimized to distinguish real images from generated ones. The adversarial loss is defined as follows:

$$\begin{cases} L_D = \mathbb{E}_{x\sim P_{\text{fake}}}[D(G(x))] - \mathbb{E}_{y\sim P_{\text{data}}}[D(y)] + \lambda\mathbb{E}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \\ L_G = -\mathbb{E}_{x\sim P_{\text{fake}}}[D(G(x))] \end{cases} \tag{4}$$

where y is the reference image, x is the input image to be enhanced, and $G(x)$ is the generated image from the generator. The third term in L_D is the gradient penalty, which enforces the 1-Lipschitz constraint on the discriminator, improving training stability and reducing the risk of mode collapse.

The MSE loss function is based on Gaussian prior, and its function is to improve the network's enhancement effect on infrared images. It compares the generator-enhanced image and the reference image pixel-by-pixel and calculates the square of the difference between them. Specifically, for each pixel, the mean square error loss function calculates the difference between the enhanced image and the reference image, squares the difference, and then sums or averages all pixels. Its definition formula is shown in formula (5):

$$L_{MSE} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}\left[\left\|y_{i,j} - G(x)_{i.j}\right\|_2^2\right] \tag{5}$$

where W and H are the width and height of the image, respectively. The mean square error (MSE) loss function encourages the model to generate an enhanced image that closely matches the reference image at each pixel position. If the enhanced image is very similar to the reference image at the pixel level, the MSE loss approaches zero, whereas a large discrepancy results in a higher loss value. The MSE loss function is sensitive to outliers, but its negative impact can be mitigated through the combination of multiple loss weights.

Perceptual Loss aims to measure the perceptual difference between the generated image and the target image. It measures the similarity between generated and target images by comparing their differences in high-level feature representations. It is able to focus more on the perceived quality of an image rather than just pixel-level differences. By introducing perceptual losses, generative models can better capture the semantic and textural features of target images, thereby generating more realistic and high-quality images.

VGG loss is a perceptual loss function based on the VGG network. The VGG network is a deep convolutional neural network that consists of multiple convolutional layers and pooling layers, and can extract semantic information and texture features of the image. The pre-trained VGG network has learned rich image feature representations by being trained on the large-scale image dataset ImageNet. The core idea of VGG loss is to use the pre-trained VGG network to extract feature representations of images and compare the differences in these feature representations between the generated image and the target image. Its expression is shown in formula (6):

$$L_{VGG} = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} \left\| \phi^{i,j,k}(y) - \phi^{i,j,k}(G(x)) \right\|_2 \tag{6}$$

where *C*, *W*, and *H* are the number of channels, width, and height of the feature map respectively, and $\phi(x)$ represents the feature representation of the image in the Block5_conv2 layer in the VGG network.

EfficientNet Loss follows a similar principle but extracts feature representations from a pre-trained EfficientNet-B0 model, focusing on global structure and efficient feature encoding. Unlike pixel-wise losses, it leverages hierarchical features to preserve both fine details and overall coherence. The formulation is given in Eq. (7):

$$L_{Eff} = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} \left\| \phi_{Eff}^{i,j,k}(y) - \phi_{Eff}^{i,j,k}(G(x)) \right\|_2 \tag{7}$$

where $\phi_{Eff}(x)$ denotes the features extracted from the final feature layers of EfficientNet-B0.

The total loss used by the 2D-GAN network is a weighted combination of the above losses and is defined as follows in formula (8):

$$Loss = L_G + \lambda_1 L_{MSE} + \lambda_2 L_{VGG} + \lambda_3 L_{Eff} \tag{8}$$

where λ_1, λ_2, and λ_3 are the weighting coefficients for MSE loss, VGG loss, and EfficientNet loss, respectively.

The adversarial loss and MSE loss ensure that the generated image closely resembles the target image at the pixel level, while the combination of VGG and EfficientNet allows the generator to capture both local fine details and global structural consistency. This integrated approach enhances the realism and perceptual fidelity of the generated images.

Experimental results and analysis

Data settings

Experimental data

1. ImageNet database

Low-contrast grayscale images resemble infrared images in their characteristics. The images lack clarity, and the boundaries between the target and background are blurred. Traditional methods are generally trained on preprocessed visible-light images. In our experiment, 5004 images were randomly selected from the benchmark dataset ImageNet[59] for training, and 500 images were used for validation. First, these images were converted into grayscale to obtain high-quality grayscale images. Then, a random contrast function was applied within a predefined contrast factor range to reduce contrast. Gaussian noise and blur were further added to decrease image brightness, thereby generating the corresponding low-quality grayscale image. As shown in Fig. 5, the first row contains high-quality reference images, while the second row presents the corresponding low-quality input images.

2. Sober–Drunk database

The Sober–Drunk Database infrared image dataset was created by Koukiou from the University of Patras in Greece. It collects infrared images of the face, eyes, sides, and hands of people before and after drinking. Relevant

Fig. 5. High-quality reference image and low-quality input image examples.

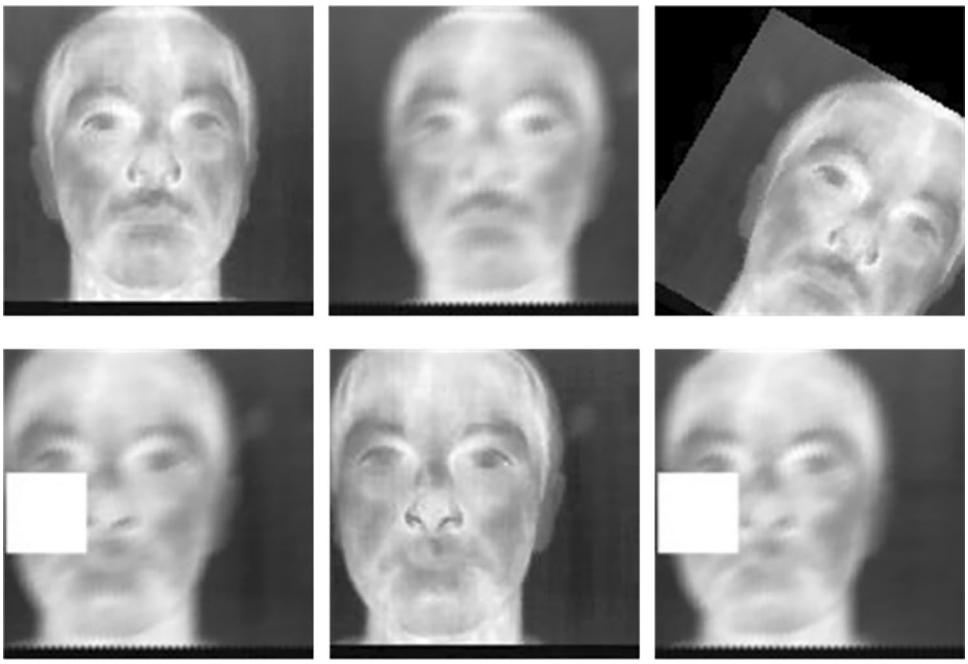


Fig. 6. Infrared facial image data augmentation.

Accessories	Parameter
Operating system	Windows 11
Processor	AMD Ryzen 7 5800X
Graphics card	NVIDIA GeForce RTX 3080
Machine with RAM	32GB
Experiment platform	PyCharm2021

Table 1. Experimental environment configuration.

data were collected through FLIR's A10 thermal infrared camera, which operates at a wavelength of 7.5 μm to 13.0 μm[60]. Each time, a sequence of 50 thermal infrared image frames with a resolution of 128×160 was collected, and the sampling period between frames was 100 ms.

In this article, due to the relatively short sampling period between frames, the similarity between some images is high, so the sample images need to be screened and quantitatively enhanced to prevent model overfitting. Therefore, some images were selected from the original dataset as basic sample data. Each sample contains infrared images from four perspectives of the face, eyes, hands, and sides. The basic sample images were horizontally flipped, rotated, cropped, partially covered, and blurred, combining multiple augmentation methods to enhance the data. An example of data enhancement on an ordinary facial infrared image is shown in Fig. 6. The final number of sample images obtained was 1120, including 640 drinking samples and 480 non-drinking samples. The experiments in this chapter used facial data for model training, and the training set, validation set, and test set were divided in a ratio of 7:2:1.

Environment and parameter settings

During the experiment, the system used was Windows 11; the graphics card model was NVIDIA GeForce RTX 3080; the processor was a 3.8GHz AMD Ryzen 7 5800X 8-Core Processor; the programming language was Python; and the deep learning framework used was PyTorch. The parameters of this experiment were designed as follows: the batch size was set to 8, the number of epochs to 101, the initial learning rate to 0.0003, and the Adam optimizer was used. The environment configuration used in the experiment was shown in Table 1.

To assess the computational complexity of the proposed model, we evaluated the number of floating-point operations (FLOPs), parameter count, and inference time. The FLOPs and parameter count were computed using the thop library, while the inference time was obtained by running the model 100 times on a single image and calculating the average execution time. The results are summarized in Table 2.

Evaluation index

Image quality evaluation methods can be divided into two categories: subjective evaluation and objective evaluation. Among them, the subjective evaluation method is based on statistical significance. It uses a sufficient

Model	FLOPs (G)	Parameters (M)	Inference time (ms)
Generator	61.32	12.18	9.417
Discriminator	3.46	1.46	1.113

Table 2. Computational complexity and inference time.

number of observers to make a qualitative assessment of the image quality. Objective evaluation methods are mainly divided into two types: full-reference and no-reference methods[61].

Full-reference image quality evaluation assesses an image by analyzing the difference between it and an ideal reference image[62]. Commonly used full-reference image quality evaluation metrics include Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

The peak signal-to-noise ratio is based on image pixels and evaluates image quality by calculating the difference between corresponding pixels of the evaluated image and the reference image. Its calculation is given in formula (9):

$$PSNR = 10 \times lg \frac{{MAX_I}^2}{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|R(i,j)-F(i,j)\right|^2} \tag{9}$$

among them, MAX_I is the maximum pixel value of the image, typically set to 255. *R* is the reference image, *F* is the image to be evaluated, and *M* and *N* are the width and height of the image, respectively. A higher *PSNR* value generally indicates better image quality. However, since it is based on the global statistics of pixel values, it overlooks local visual factors. In some cases, an image may have a high *PSNR* value, yet its perceptual quality as judged by the human eye may be poor.

Structural similarity is based on structural information. The similarity between the structures of two images is computed based on the correlation between image pixels, and the image quality is measured in terms of brightness, contrast, and structure. It is defined in formula (10):

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y+C_1)(2\sigma_{XY}+C_2}{(\mu_X^2+\mu_Y^2+C_1)(\sigma_X^2+\sigma_Y^2+C_2)} \tag{10}$$

where μ_X and μ_Y denote the average gray values of image *X* and image *Y*, σ_{XY} denotes the covariance between them, σ_X^2 and σ_Y^2 are their variances, and C_1 and C_2 are constants used to stabilize the denominator when it is close to zero. *SSIM* takes values in the range [0, 1], where a value closer to 1 indicates greater similarity between the two images. *SSIM* evaluates image quality from multiple aspects and is one of the most important objective metrics.

The reference-free method, also known as the no-reference evaluation method, breaks away from the dependence on a reference image and evaluates image quality based on the statistical characteristics of the image. Commonly used no-reference image quality evaluation metrics include Average Gradient (AG), Information Entropy (IE), and Enhancement Measure Evaluation (EME).

The average gradient can measure the detail and texture variations in the image, reflecting its ability to represent fine structures. The specific formula is given by formula (11):

$$AG = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{(\Delta xF(i,j))^2+(\Delta yF(i,j))^2} \tag{11}$$

where $\Delta_x F(i,j)$ and $\Delta_y F(i,j)$ respectively represent the first-order differences in the *x* and *y* directions of pixel (*i*, *j*) in the evaluated image *F*. A higher *AG* value indicates greater image brightness and richer detail.

Information entropy quantifies the amount of information in an image from the perspective of information theory. It is defined in formula (12):

$$IE = -\sum_{i=0}^{255} p(i)log_2 p(i) \tag{12}$$

where *p*(*i*) denotes the probability of gray level *i* occurring in the image. A higher *IE* value generally indicates better image quality.

Comparative experiment

To validate the effectiveness of the proposed method, we compare our model with several publicly available infrared image enhancement algorithms. Specifically, we consider traditional enhancement techniques such as histogram equalization (HE) and contrast-limited adaptive histogram equalization (CLAHE), which are widely used for improving image contrast. Additionally, we include guided filtering-based image enhancement (GF)[63], Single-Scale Retinex (SSR)[64], and Multi-Scale Retinex (MSR)[65], which leverage image decomposition and illumination correction to enhance details while preserving structural consistency. Beyond these conventional methods, we also compare our method with deep learning-based approaches, including the Brightness-based Convolutional Neural Network for Thermal Image Enhancement (TIECNN)[66], which employs convolutional

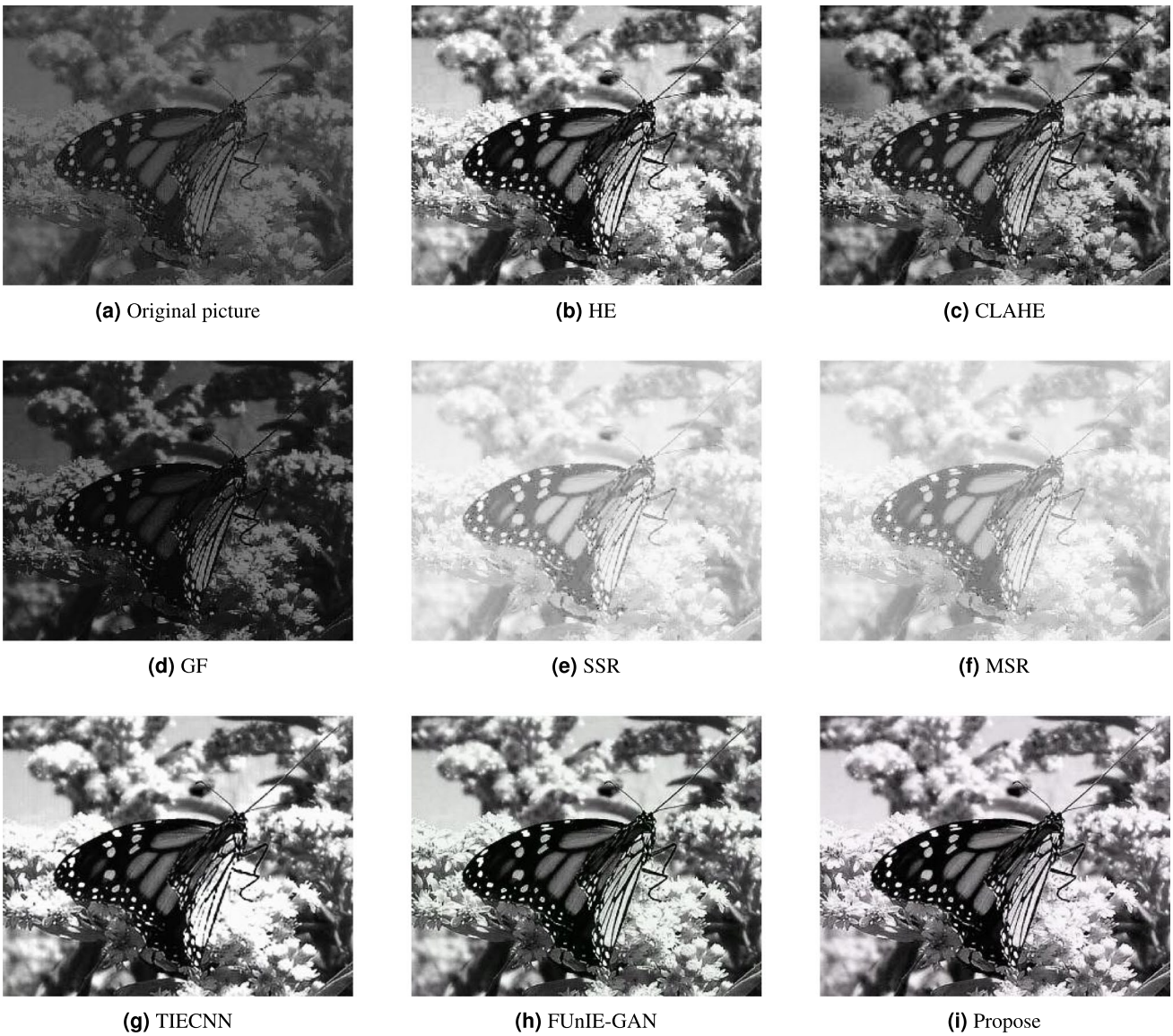
neural networks to improve the perceptual quality of thermal images, and the Fully-Convolutional Underwater Image Enhancement GAN (FUnIE-GAN)[67], a generative adversarial network originally designed for underwater image enhancement but also applicable to infrared image enhancement tasks. This comprehensive comparison enables a rigorous evaluation of our method's performance in contrast to both traditional and deep learning-based approaches.

Each algorithm was tested comparatively on Dataset 1. The evaluation metrics used include Average Gradient (AG), Information Entropy (IE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). The experimental results are shown in Figs. 7, 8, 9, and Table 3.

Histogram Equalization (HE) significantly enhances image contrast, as indicated by its high AG values. However, this often leads to structural distortions and increased noise. As shown in Fig. 8b, HE introduces blocky artifacts, particularly in high-brightness regions, causing unnatural intensity distributions. This effect is further reflected in its low PSNR (9.1228) and SSIM (0.3746) scores in Example 3, indicating a significant loss of structural coherence. Overexposed regions in Figs. 7b and 8b highlight its instability under varying illumination.

CLAHE partially mitigates these issues by locally adjusting contrast, leading to improvements in PSNR and SSIM compared to HE. However, as seen in Figs. 7c and 8c, CLAHE struggles with uneven brightness adaptation across different regions. This is particularly evident in Fig. 9c, where darker regions remain under-enhanced, limiting visibility improvements in low-light conditions.

Guided Filtering (GF), Single-Scale Retinex (SSR), and Multi-Scale Retinex (MSR) provide moderate enhancements but show limited improvement in image clarity. Their lower AG values suggest weaker contrast enhancement, while their PSNR values, all below 16, indicate a lack of fine detail preservation. As seen in Figs. 7d–f and 8d–f, these methods enhance brightness but introduce excessive smoothing, leading to a loss of

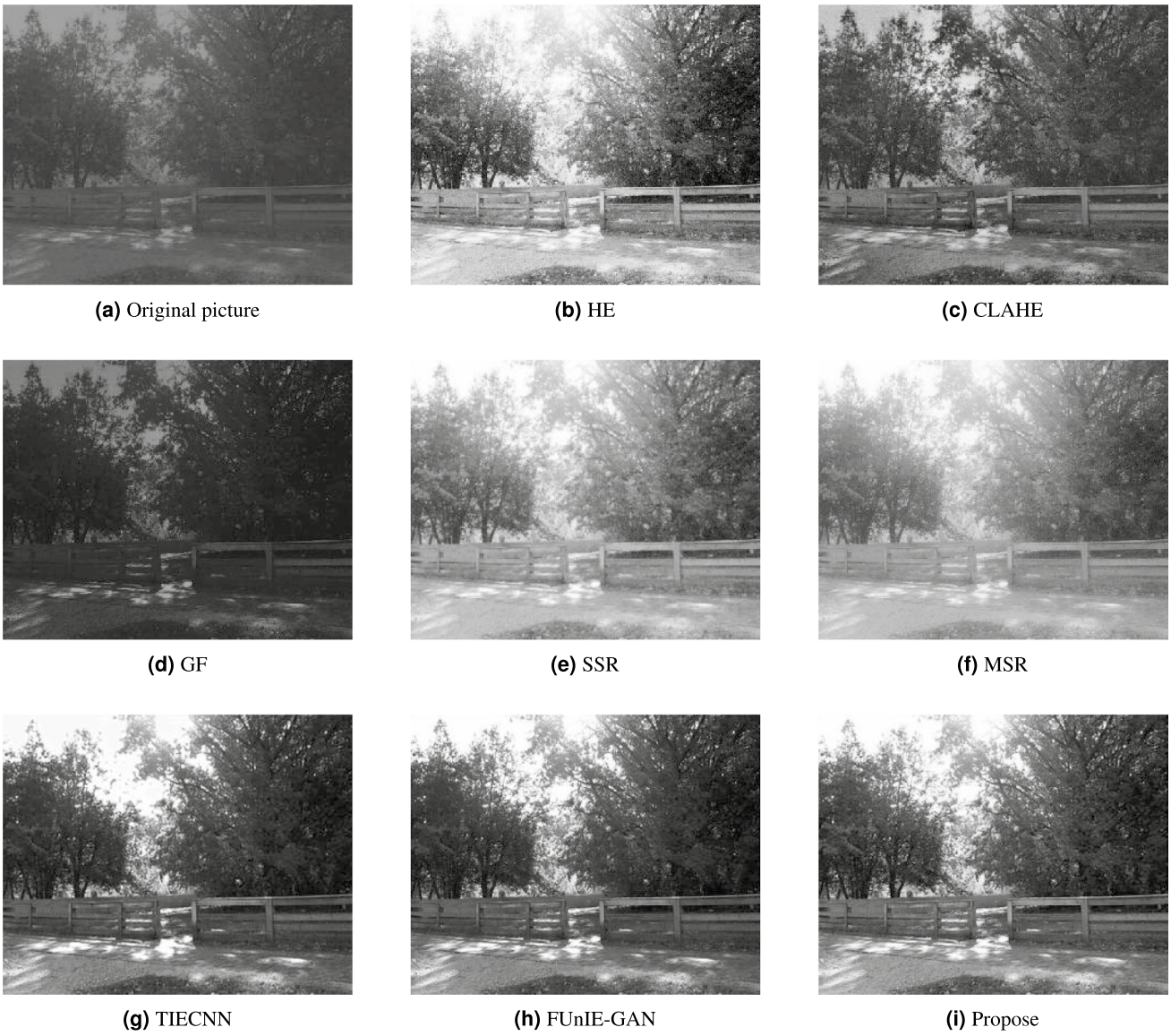


(a) Original picture **(b)** HE **(c)** CLAHE

(d) GF **(e)** SSR **(f)** MSR

(g) TIECNN **(h)** FUnIE-GAN **(i)** Propose

Fig. 7. Example 1 experimental results.



(a) Original picture **(b)** HE **(c)** CLAHE

(d) GF **(e)** SSR **(f)** MSR

(g) TIECNN **(h)** FUnIE-GAN **(i)** Propose

Fig. 8. Example 2 experimental results.

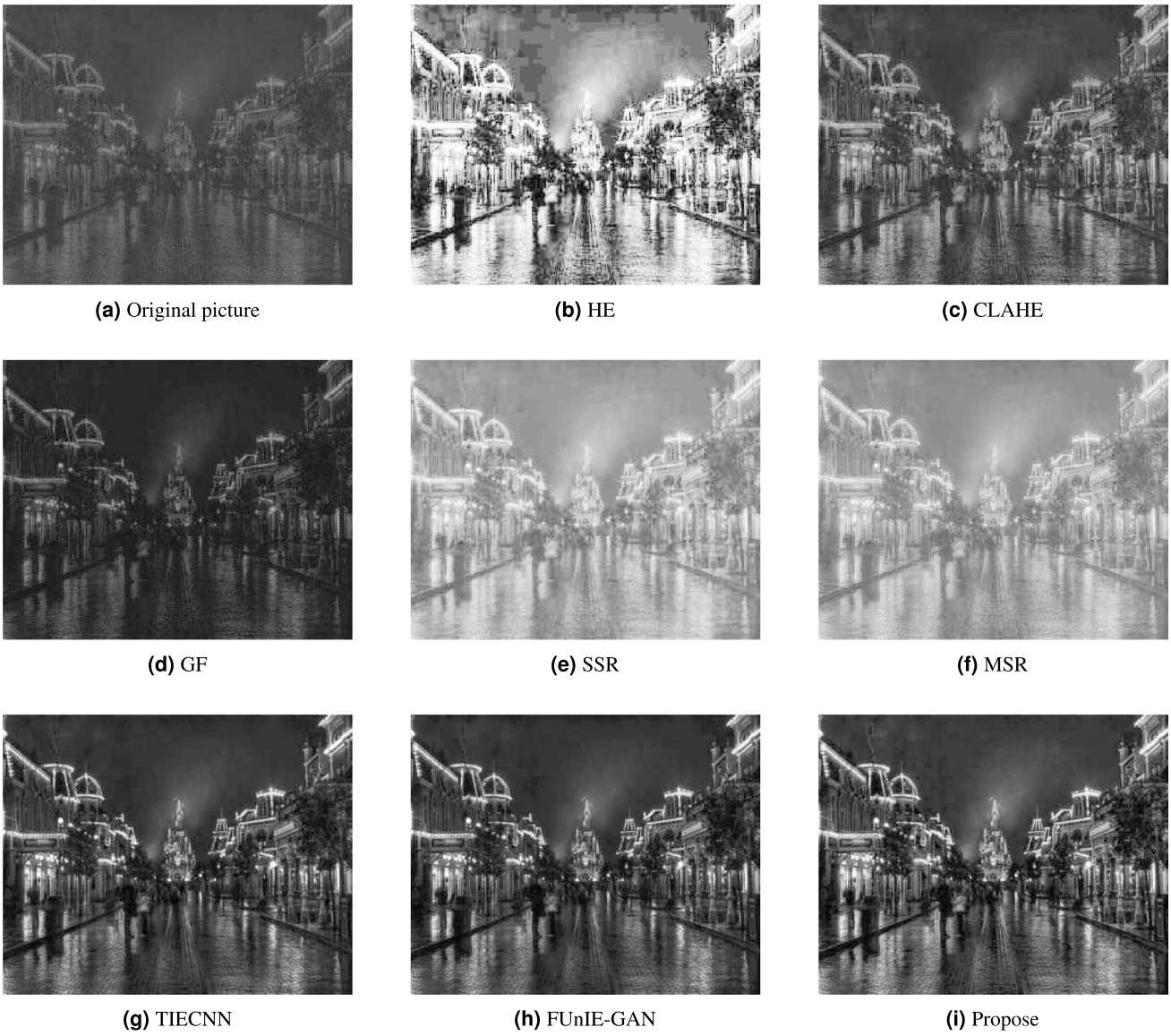
high-frequency textures. This effect is particularly noticeable in Fig. 9f, where fine structures appear blurred. Their SSIM values, remaining below 0.5, further confirm their limited ability to retain the original image structure.

Although traditional methods enhance image contrast, they often introduce noise and fail to preserve fine structural details. To address these limitations, deep learning-based methods have been explored. TIECNN achieves higher PSNR and SSIM scores than conventional methods, yet it tends to overexpose bright areas, leading to loss of detail, as seen in Figs. 7 and 8. Similarly, FUnIE-GAN achieves higher SSIM and PSNR scores compared to traditional methods, indicating better structural preservation and noise suppression. However, as seen in Figs. 8h and 9h, the method struggles to retain details in dark regions, leading to texture flattening and reduced perceptual quality.

In contrast, the proposed 2D-GAN based method consistently outperforms all other approaches in terms of PSNR, SSIM, AG, and IE, ensuring both contrast enhancement and fine detail preservation. The self-attention mechanism employed in the model enables better retention of local and global structural information, ensuring high-quality enhancement under varying illumination conditions. As shown in Table 3, the 2D-GAN method consistently outperforms TIECNN and FUnIE-GAN across all four evaluation indicators. These findings underscore the effectiveness and superiority of the proposed approach in infrared image enhancement, further validated by both quantitative analysis and qualitative visual assessments presented in Figs. 7, 8, and 9.

Ablation experiment

To further verify the effectiveness of the network structure and each proposed component, ablation experiments were conducted on both the overall network architecture and its main components. The main structures and



(a) Original picture **(b)** HE **(c)** CLAHE

(d) GF **(e)** SSR **(f)** MSR

(g) TIECNN **(h)** FUnIE-GAN **(i)** Propose

Fig. 9. Example 3 experimental results.

	Method	OG	HE	CLAHE	GF	SSR	MSR	TIECNN	FUnIE	Ours
Img1	AG	6.0248	**21.5455**	20.3509	10.9852	7.8593	7.9757	19.6100	21.7028	21.3267
	IE	6.2390	**7.9608**	7.7017	6.6686	6.6707	7.7514	7.8571	7.9113	7.9210
	PSNR	7.8692	23.3966	14.2653	8.2680	8.8879	9.1173	21.3133	25.0579	**25.5274**
	SSIM	0.3474	0.8459	0.7579	0.3994	0.4578	0.4520	0.8584	0.9166	**0.9258**
Img2	AG	5.1035	**21.4965**	18.1191	7.0424	12.0773	10.4535	14.2562	16.0262	17.1275
	IE	5.8133	**7.8804**	7.1668	5.9153	7.7584	7.7491	7.4697	7.5638	7.7395
	PSNR	10.6677	14.9974	16.1804	10.4561	15.0742	15.1718	21.5889	23.6008	**26.9968**
	SSIM	0.4025	0.6304	0.7429	0.3764	0.6604	0.6363	0.7852	0.8452	**0.8953**
Img3	AG	4.0684	**27.9084**	13.6946	6.4984	10.3271	9.9790	12.4966	13.4100	15.8442
	IE	5.0308	**7.1811**	6.8459	5.2812	6.9053	6.9398	6.6147	6.7015	7.0372
	PSNR	15.3018	9.1228	23.4671	15.9015	8.8818	9.1681	23.3151	24.0755	**27.2207**
	SSIM	0.4728	0.3746	0.7807	0.4856	0.4685	0.4754	0.8177	0.8379	**0.8839**

Table 3. Comparison of results of various infrared image enhancement methods. Best values for each metric are in bold.

components tested include a dual-decoding network structure, inner skip connections, external skip connections, and a cross-level attention module. For these four components, five different network configurations (C1–C5) were evaluated.Among them, C1 represents a network with a single encoding-decoding structure. C2 adds inner skip connections based on C1. C3 adopts a U-shaped dual-decoding structure as the main body of the network and incorporates both inner and external skip connections. C4 also employs a U-shaped dual-decoding structure as the main body but removes the external connections and introduces the cross-level attention module. Finally, C5 uses both inner and external skip connections along with the attention mechanism on top of the U-shaped dual decoding structure, forming the complete network design of this paper.

The dataset and evaluation metrics used in this experiment are the same as those in the previous section and will not be detailed here. Figs. 10 and 11 show examples of test-set images under different network configurations. The specific results of the ablation experiments are presented in Table 4.

As can be seen from Fig. 10 and 11, compared with the original image, although the C1 network effectively enhanced the brightness and contrast of the image, the image became blurred, lost a lot of details, and the enhanced image introduced grid noise. Overall The quality is poor; the human visual effects of the C2 and C3 networks are better. The difference is that the tones are different but the details are better preserved; the C4 network enhances the image contrast and further enhances the brightness of the bright parts of the image. At the same time It weakens the brightness of the dark parts, and also obviously introduces grid noise, and loses a lot of dark details in the original image; the visual enhancement effect of the C5 network is also better, and the brightness and contrast are better enhanced. And the details remain intact.

In Table 4, C2 adds inner joins to C1. Compared with the experimental results of the two, except for the IE indicator, C2 is not much different from C1, and the results of other indicators are far better than C1. Through

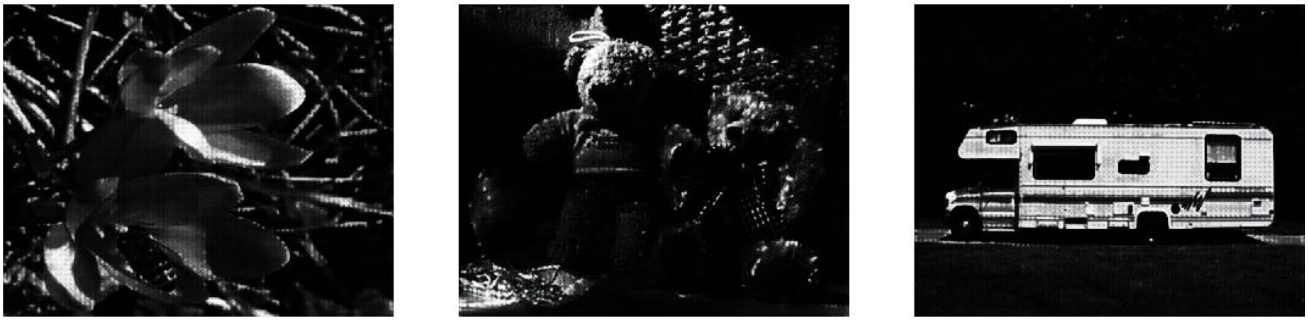
(a) Original picture

(b) C1

(c) C2

Fig. 10. Examples of experimental results for different network configurations.

(a) C3



(b) C4

(c) C5

Fig. 11. Examples of experimental results for different network configurations.

Main structure of the network	c1	c2	c3	c4	c5
Inner skip connections	×	√	√	√	√
external skip connections	×	×	√	×	√
Cross-level attention	×	×	×	√	√
AG	6.4158	11.0332	11.8112	6.7855	**12.5988**
IE	6.9469	6.8513	6.9448	4.1879	**7.1194**
PSNR	21.5558	25.3429	**28.6659**	13.7420	27.7329
SSIM	0.5269	0.8353	0.8788	0.1951	**0.8811**

Table 4. Ablation study on the effectiveness of key components. Best values for each metric are in bold.

comparison, it can effectively Prove the effectiveness and importance of internal skip connections; compared with C5, C3 does not add cross-level attention. From the experimental results, although C5 is equivalent to C3 results in average PSNR and SSIM indicators, C5 has better performance in AG and IE. The performance on all indicators is significantly better than C3, with AG and IE values increasing by 6.67% and 2.51% respectively, indicating the effectiveness of cross-level attention in the network; C5 adds external skip connections based on the structure of C4. Compared with C4, the results have extremely high improvements in the four indicators

Attention mechanism	Pixel attention	Multi-head attention	Cross-level attention
AG	11.9877	12.0637	**12.5988**
IE	6.9873	7.0633	**7.1194**
PSNR	**28.7329**	27.6342	28.6535
SSIM	0.8648	0.8634	**0.8811**

Table 5. Comparison of different attention mechanisms in the ablation study. Best values for each metric are in bold.

Loss function	L1	L2	L3	L4	L5
WGAN-GP	×	√	√	√	√
MSE	×	×	√	√	√
VGG	×	×	×	√	√
EfficientNet	×	×	×	×	√
AG	8.4531	10.7532	11.2134	11.7658	**12.5988**
IE	5.1027	6.5321	6.8312	6.9725	**7.1194**
PSNR	22.4358	25.7312	27.0123	27.6859	**28.6535**
SSIM	0.7921	0.8451	0.8623	0.8745	**0.8811**

Table 6. Ablation study on the effectiveness of loss functions. Best values for each metric are in bold.

of AG, IE, PSNR, and SSIM. From a numerical point of view, the four indicators have increased by 85.67%, 70.00%, 101.8%, and 351.6% respectively, which fully verifies that Understand the effectiveness and importance of external skip connections. Combined with the experimental results of the three network structures of C2, C3, and C4, compared with the ordinary encoding and decoding network, the dual decoding network structure can only show its advantages when external skip connections are added, which fully proves the advantages of external skip connections. importance. C5 is the complete network proposed in this article, and its comprehensive experimental results are also the best.

Combining the ablation experiment results from Figs. 10 and 11 and Table 4, we can see—based on both human visual inspection and subjective-objective indicators—that each component (inner skip connections, external skip connections, and the cross-level attention module) is effective and crucial. These findings fully confirm the superiority of the network proposed in this article.

To better investigate the effectiveness of the cross-level attention mechanism, we compare it with pixel attention and multi-head attention. By analyzing the data presented in Table 5, we evaluate the impact of different attention mechanisms on model performance and verify the advantages of cross-level attention.

Experimental results demonstrate that the cross-level attention mechanism outperforms other attention mechanisms in multiple key metrics. In terms of AG, cross-level attention achieves the highest score, surpassing both pixel attention and multi-head attention. This indicates that cross-level attention can more effectively enhance image edge details and texture features, thereby improving overall image sharpness.Regarding IE, cross-level attention achieves 7.1194, outperforming both pixel attention and multi-head attention, suggesting that it retains more information during the image enhancement process and thus increases the richness of image content. Furthermore, in SSIM, cross-level attention achieves the highest score, significantly higher than pixel attention and multi-head attention, indicating its ability to better preserve structural consistency and produce enhanced images that more closely resemble the original.For PSNR, cross-level attention achieves 28.6535, which is close to pixel attention but significantly better than multi-head attention, demonstrating its effectiveness in maintaining image quality and reducing noise.

Overall, the cross-level attention mechanism exhibits superior performance across multiple evaluation metrics, effectively enhancing image clarity, richness, and structural integrity. These findings further validate the feasibility and advantages of the proposed mechanism in image processing tasks

In addition to network structure, the choice of loss functions plays a crucial role in the enhancement process. To further investigate the impact of different loss components, we designed five loss configurations for comparative analysis. L1 employs only a simple adversarial loss to evaluate its role in generating realistic images; L2 uses only MSE to strengthen pixel-level supervision and improve enhancement quality; L3 employs only VGG loss to analyze the role of perceptual loss in maintaining semantic consistency; L4 utilizes only EfficientNet loss to explore its contribution to global structural fidelity; and L5 adopts the complete loss combination (WGAN-GP, MSE, VGG, and Efficient Net) to provide comprehensive optimization constraints. The experimental results are presented in Table 6.

The ablation study evaluates the impact of different loss components on the enhancement process by comparing five configurations (L1–L5). The results highlight the contributions of adversarial, pixel-level, and perceptual losses in optimizing image quality.

Using only an adversarial loss results in the weakest performance, with SSIM dropping to 0.7921 and AG at 8.4531, indicating that while adversarial training helps generate realistic images, it lacks sufficient constraints

Fig. 12. Infrared facial image enhancement results.

	Original image	After enhancement
AG	4.8826	6.4653
IE	6.9263	7.1802
PSNR	18.3806	21.4032
SSIM	0.9189	0.9641

Table 7. Comparison of the values of various indicators before and after Sober–Drunk data set enhancement.

for accurate reconstruction. Adding MSE significantly improves image quality, increasing SSIM to 0.8451 and PSNR to 25.7312, by enforcing pixel-wise supervision, reducing artifacts, and enhancing detail preservation. However, relying solely on MSE does not fully capture high-level structural information, limiting perceptual improvements.

Introducing VGG loss further refines structural consistency and texture sharpness, leading to a noticeable increase in SSIM to 0.8623, demonstrating the effectiveness of perceptual supervision. The addition of EfficientNet loss further enhances global coherence, complementing VGG loss by improving feature representation and maintaining natural contrast. The full loss configuration, which integrates all components, achieves the best overall performance, with AG reaching 12.5988 and SSIM improving to 0.8811, confirming that a combination of adversarial, pixel-wise, and perceptual losses leads to the most balanced and high-quality image enhancement.

The findings suggest that MSE is crucial for pixel accuracy, perceptual losses contribute to structural consistency, and EfficientNet further refines global coherence. The study validates that combining all loss functions results in optimal infrared image enhancement, striking a balance between realism, detail preservation, and structural fidelity.

Image enhancement results

The Sober–Drunk data set is enhanced using the 2D-GAN based infrared image enhancement method proposed in this article. Examples of images before and after enhancement are shown in Fig. 12.

By observing Fig. 12, it can be seen that the contrast of the image enhanced by the 2D-GAN network has been significantly improved, the overall image is clearer, and the details are well preserved. The average index results obtained from the experiment are shown in Table 7.

Combining the information in Fig. 12 and Table 7, it can be seen that all indicators of the enhanced image have improved, indicating that the infrared image enhancement method based on 2D-GAN can effectively enhance the quality of infrared images.

Discussion

Although our method generally exhibits strong enhancement performance, several challenges persist during training and evaluation. Chief among these is the model's tendency to over-enhance low-contrast images, resulting in unnatural textures and structural distortions. This issue largely stems from the model's high sensitivity to local intensity variations, occasionally leading to excessive sharpening and the loss of originally subtle details. When the intensity distribution is overly complex, the model also struggles to maintain balanced enhancement across different regions, ultimately producing inconsistent outputs.

To address these issues, we introduce a cross layer attention mechanism that dynamically allocates feature weights across different layers of the network, thereby preserving local details while maintaining global structural consistency. Additionally, by combining multiple loss functions, we enhance training stability and effectively mitigate mode collapse in adversarial training, further improving overall generation quality.

Despite these optimizations, certain limitations remain. The multi-scale processing and adversarial training introduce computational complexity, posing challenges for real-time applications—particularly in resource-constrained environments. The model also remains sensitive to input variations and occasionally produces

inconsistent enhancement under extreme intensity changes or highly dynamic backgrounds. Moreover, its generalization to real-world infrared images with various noise patterns and sensor characteristics is still an open question, requiring additional validation under diverse imaging conditions.

Future research will focus on improving computational efficiency, refining attention mechanisms, and strengthening generalization to ensure broader applicability. As infrared imaging gains increasing importance in security monitoring, industrial inspection, and medical diagnosis, these advancements will further enhance the practical utility and effectiveness of the proposed approach in real-world scenarios.

Conclusion

This paper presents a 2D-GAN based infrared image enhancement method is proposed to address the problems of low contrast, blurred details, and high noise in infrared images.In the proposed method, a dual decoder structure is introduced to enhance the network's ability to extract and represent key features, thereby improving the quality of enhanced images. Internal and external skip connections are designed to minimize information loss, preserve details, and maintain structural coherence during enhancement. A cross-layer attention module is developed to dynamically capture long-range dependencies and multi-scale context, improving texture clarity and structural coherence in generated images. A joint loss function is formulated to optimize enhancement quality, incorporating perceptual loss, deep feature similarity, and pixel-level supervision that work synergistically to refine local textures while preserving global structural coherence. The performance of the proposed method is evaluated on the ImageNet and Sober–Drunk datasets. Experimental results demonstrate that the 2D-GAN method achieves superior performance in infrared image enhancement, effectively improving visual quality and structural fidelity. In the future, we will further extend the method to larger datasets and explore its applicability in practical scenarios such as security monitoring, industrial inspection, and medical diagnosis, aiming to enhance its generalization capability and robustness.

Data availability

The datasets generated and analyzed during the current study are available from the following sources: the ImageNet database hosted by Princeton University and Stanford University at https://image-net.org/download, and the Sober–Drunk Database provided by Koukiou from the University of Patras, Greece, at https://github.com/YuYang88888/Sober–Drunk-Database. The code used in this study is available at https://github.com/Adriannajl/2D-GAN.

Received: 2 September 2024; Accepted: 9 June 2025
Published online: 01 July 2025

References

1. Sun, B. et al. Response mechanism of dynamic tensile mechanics in granite under microwave irradiation: Insights from experiments and simulations. *Rock Mech. Rock Eng.* https://doi.org/10.1007/s00603-024-03958-8 (2024).
2. Renn, N., Onyango, J. & McCormick, W. Digital infrared thermal imaging and manual lameness scoring as a means for lameness detection in cattle. *Vet. Clin. Sci.* **2**, 16–23 (2014).
3. Uemura, D. K., Shah, S. B., Regmi, P., Grimes, J. & Wang-Li, L. Low-cost calibration method for the infrared camera. *Appl. Eng. Agric.* **39**, 529–534. https://doi.org/10.13031/aea.15546 (2023).
4. Li, R. et al. Estimation of nitrogen content in wheat using indices derived from rgb and thermal infrared imaging. *Field Crop Res.* **289**, 108735. https://doi.org/10.1016/j.fcr.2022.108735 (2022).
5. Siami, M., Barszcz, T., Wodecki, J. & Zimroz, R. Design of an infrared image processing pipeline for robotic inspection of conveyor systems in opencast mining sites. *Energies* **15**, 6771. https://doi.org/10.3390/en15186771 (2022).
6. Yang, K., Xiang, W., Chen, Z., Zhang, J. & Liu, Y. A review on infrared and visible image fusion algorithms based on neural networks. *J. Vis. Commun. Image Represent.* https://doi.org/10.1016/j.jvcir.2024.104179 (2024).
7. Wang, D., Lai, R. & Guan, J. Target attention deep neural network for infrared image enhancement. *Infrared Phys. Technol.* **115**, 103690. https://doi.org/10.1016/j.infrared.2021.103690 (2021).
8. Chen, Y. et al. Implicit multi-spectral transformer: An lightweight and effective visible to infrared image translation model. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. (IEEE, 2024). https://doi.org/10.1109/IJCNN60899.2024.10650029.
9. Navrátil, L., Le Saux, V., Leclercq, S., Carrere, N. & Marco, Y. Infrared image processing to guide the identification of damage and dissipative mechanisms in 3d layer-to-layer woven composites. *Appl. Compos. Mater.* **29**, 1449–1477. https://doi.org/10.1007/s10443-022-10023-6 (2022).
10. Katircioğlu, F., Cingiz, Z., Çay, Y., Gürel, A. E. & Kolip, A. Performance assessment of a refrigeration system charged with different refrigerants using infrared image processing techniques. *Arab. J. Sci. Eng.* **46**, 12009–12028. https://doi.org/10.1007/s13369-021-05794-2 (2021).
11. Shi, L., Wen, Y.-B., Zhao, G.-S. & Yu, T. Recognition of blast furnace gas flow center distribution based on infrared image processing. *J. Iron. Steel Res. Int.* **23**, 203–209. https://doi.org/10.1016/s1006-706x(16)30035-8 (2016).
12. Binbin, Y. An improved infrared image processing method based on adaptive threshold denoising. *EURASIP J. Image Video Process.* **2019**, 5. https://doi.org/10.1186/s13640-018-0401-8 (2019).
13. Ma, Q., Zhao, M., Zheng, Y., Sun, L. & Ni, F. Infrared lmage detail enhancement based on adaptive conditional histogram equalization. *Infrared Technol.* **46**, 52–60 (2024).
14. Lu, G. et al. Gan-ha: A generative adversarial network with a novel heterogeneous dual-discriminator network and a new attention-based fusion strategy for infrared and visible image fusion. *Infrared Phys. Technol.* **142**, 105548. https://doi.org/10.1016/j.infrared.2024.105548 (2024).
15. Wu, Z. et al. A lightweight gan-based image fusion algorithm for visible and infrared images. In *2024 4th International Conference on Computer Science and Blockchain (CCSB)*, 466–470 (IEEE, 2024). https://doi.org/10.1109/CCSB63463.2024.10735676.
16. Cheng, J. et al. Lightweight infrared lmage enhancement network based on adversarial generation. *J. Signal Process.* **40**, 484–491. https://doi.org/10.16798/j.issn.1003-0530.2024.03.007 (2024).
17. Wang, H., Tang, Z., Jie, F., Liu, Q. & Zhang, S. An infrared lmage enhancement mapping method based on multi-scale multi-histogram fusion. *Electron. Opt. Control* **31**, 75–80 (2024).

18. Rivera-Aguilar, B. A., Cuevas, E., Pérez, M., Camarena, O. & Rodríguez, A. A new histogram equalization technique for contrast enhancement of grayscale images using the differential evolution algorithm. *Neural Comput. Appl.* https://doi.org/10.1007/s00521-024-09739-2 (2024).
19. Lu, P. & Li, J. Normal histogram matching and non-sharpening masking for infrared image enhancement. *Radio Commun. Technol.* **51**(2), 400-406 (2025).
20. Zhang, H., Ma, X. & Tian, Y. An image fusion method based on curvelet transform and guided filter enhancement. *Math. Probl. Eng.* **2020**, 9821715. https://doi.org/10.1155/2020/9821715 (2020).
21. Zhang, H., Chen, Z., Cao, J. & Li, C. Infrared image enhancement based on adaptive guided filter and global–local mapping. In *Photonics*, Vol. 11, 717 (MDPI, 2024).
22. Li, H. et al. Thermal infrared-image-enhancement algorithm based on multi-scale guided filtering. *Fire* **7**, 192. https://doi.org/10.3390/photonics11080717 (2024).
23. Song, H., Wang, Z., Cao, W., Zhang, Y. & Leng, X. Infrared image enhancement based on guided filtering and adaptive algorithm and its fpga implementation. *Microw. Opt. Technol. Lett.* **67**, e70105. https://doi.org/10.1002/mop.70105 (2025).
24. Yuan, Z. et al. Infrared image enhancement based on multiple scale retinex and sequential guided image filter. In *Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning*, 196–201 (2024). https://doi.org/10.1145/3654823.3654859.
25. Wu, G. et al. Infrared Image enhancement algorithm based on improved generative adversarial network. *Semicond. Optoelectron.* **44**, 782–787. https://doi.org/10.16818/j.issn1001-5868.2023060201 (2023).
26. Tang, H., Liu, G., Qian, Y., Wang, J. & Xiong, J. Egefusion: Towards edge gradient enhancement in infrared and visible image fusion with multi-scale transform. *IEEE Trans. Comput. Imaging*. https://doi.org/10.1109/tci.2024.3369398 (2024).
27. Xie, J. et al. Infrared image enhancement method for power equipment based on improved retinex-net. *Infrared Technol.* 1–13. http://kns.cnki.net/kcms/detail/53.1053.TN.20231218.0944.002.html (2024).
28. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, 2672–2680 (2014).
29. Wang, K. et al. Generative adversarial networks: Introduction and outlook. *IEEE/CAA J. Autom. Sin.* **4**, 588–598. https://doi.org/10.1109/JAS.2017.7510583 (2017).
30. Shin, H.-C. et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, 1–11 (Springer, 2018).
31. Liu, T.-J. & Chen, Y.-Z. Satellite image super-resolution by 2d rrdb and edge-enhanced generative adversarial network. *Appl. Sci.* **12**, 12311. https://doi.org/10.3390/app122312311 (2022).
32. Shamsolmoali, P., Zareapoor, M., Wang, R., Jain, D. K. & Yang, J. G-ganisr: Gradual generative adversarial network for image super resolution. *Neurocomputing* **366**, 140–153. https://doi.org/10.1016/j.neucom.2019.07.094 (2019).
33. Chen, Y. et al. Research on image inpainting algorithm of improved gan based on two-discriminations networks. *Appl. Intell.* **51**, 3460–3474. https://doi.org/10.1007/s10489-020-01971-2 (2021).
34. Xiong, F., Wang, Q. & Gao, Q. Consistent embedded gan for image-to-image translation. *IEEE Access* **7**, 126651–126661. https://doi.org/10.1109/access.2019.2939654 (2019).
35. Tan, H., Liu, X., Yin, B. & Li, X. Dr-gan: Distribution regularization for text-to-image generation. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 10309–10323. https://doi.org/10.1109/tnnls.2022.3165573 (2022).
36. Zhang, H., Sindagi, V. & Patel, V. M. Image de-raining using a conditional generative adversarial network. *IEEE Trans. Circuits Syst. Video Technol.* **30**, 3943–3956. https://doi.org/10.1109/tcsvt.2019.2920407 (2019).
37. Zhang, H. et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1947–1962. https://doi.org/10.1109/tpami.2018.2856256 (2018).
38. Ni, Z., Yang, W., Wang, S., Ma, L. & Kwong, S. Towards unsupervised deep image enhancement with generative adversarial network. *IEEE Trans. Image Process.* **29**, 9140–9151. https://doi.org/10.1109/tip.2020.3023615 (2020).
39. Shafiq, H. & Lee, B. Transforming color: A novel image colorization method. *Electronics* **13**, 2511. https://doi.org/10.3390/electronics13132511 (2024).
40. Ju, Z. et al. Hagan: Hybrid augmented generative adversarial network for medical image synthesis. arXiv preprint arXiv:2405.04902https://doi.org/10.1007/s11633-024-1528-ys (2024).
41. Karras, T. et al. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119 (2020).
42. Kim, J., Kim, M., Kang, H. & Lee, K. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830. https://doi.org/10.48550/arXiv.1907.10830 (2024).
43. Wang, X., Xie, L., Dong, C. & Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–1914 (2021).
44. Jiang, Y., Chang, S. & Wang, Z. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Adv. Neural. Inf. Process. Syst.* **34**, 14745–14758 (2021).
45. Liu, J., Duan, J., Hao, Y., Chen, G. & Zhang, H. Semantic-guided polarization image fusion method based on a dual-discriminator gan. *Opt. Express* **30**, 43601–43621 (2022).
46. Cong, R. et al. Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators. *IEEE Trans. Image Process.* **32**, 4472–4485 (2023).
47. Song, A., Duan, H., Pei, H. & Ding, L. Triple-discriminator generative adversarial network for infrared and visible image fusion. *Neurocomputing* **483**, 183–194 (2022).
48. Zhang, H., Yuan, J., Tian, X. & Ma, J. Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual Markovian discriminators. *IEEE Trans. Comput. Imaging* **7**, 1134–1147 (2021).
49. Mohammed, D. & Khudeye, R. S. Bridging techniques: A review of deep learning and fuzzy logic applications. *Artif. Intell. Robot. Dev. J.* **4**, 292–313 (2024).
50. Xu, S., Wang, J., He, N., Hu, X. & Sun, F. Underwater image enhancement method based on a cross attention mechanism. *Multimedia Syst.* **30**, 26 (2024).
51. Zhou, J., Sun, J., Zhang, W. & Lin, Z. Multi-view underwater image enhancement method via embedded fusion mechanism. *Eng. Appl. Artif. Intell.* **121**, 105946 (2023).
52. Chen, Q., Fan, J. & Chen, W. An improved image enhancement framework based on multiple attention mechanism. *Displays* **70**, 102091 (2021).
53. Dong, Y., Liu, Q., Du, B. & Zhang, L. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Trans. Image Process.* **31**, 1559–1572 (2022).
54. Wang, Z. et al. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693 (2022).
55. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144. https://doi.org/10.1145/3422622 (2020).
56. Wang, X., Ouyang, J., Li, D. & Zhang, G. Underwater object recognition based on deep encoding-decoding network. *J. Ocean Univ. China* **18**, 376–382. https://doi.org/10.1007/s11802-019-3858-x (2019).
57. Mao, X. et al. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802 (2017).

58. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 214–223 (PMLR, 2017).
59. Koukiou, G. & Anastassopoulos, V. Drunk person identification using thermal infrared images. *Int. J. Electron. Secur. Digit. Forensics* **4**, 229–243. https://doi.org/10.1504/ijesdf.2012.049747 (2012).
60. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90. https://doi.org/10.1145/3065386 (2017).
61. Fang, Y., Sui, X., Yan, J., Liu, X. & Huang, L. Progress in no-reference image quality assessment. *J. Image Graphics* **26**, 265–286 (2021).
62. Athar, S. & Wang, Z. Degraded reference image quality assessment. *IEEE Trans. Image Process.* **32**, 822–837. https://doi.org/10.1109/tip.2023.3234498 (2023).
63. He, K., Sun, J. & Tang, X. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1397–1409. https://doi.org/10.1109/TPAMI.2012.213 (2012).
64. Qi, J., Abera, D. E. & Cheng, J. Ps-gan: Pseudo supervised generative adversarial network with single scale retinex embedding for infrared and visible image fusion. *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.* **18**, 1766–1777 (2024).
65. Huang, Y., Lu, X., Quan, Y., Xu, Y. & Ji, H. Image shadow removal via multi-scale deep retinex decomposition. *Pattern Recogn.* **159**, 111126 (2025).
66. Lee, K., Lee, J., Lee, J., Hwang, S. & Lee, S. Brightness-based convolutional neural network for thermal image enhancement. *IEEE Access* **5**, 26867–26879. https://doi.org/10.1109/ACCESS.2017.2769687 (2017).
67. Islam, M. J., Xia, Y. & Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* **5**, 3227–3234. https://doi.org/10.1109/LRA.2020.2974710 (2020).

Acknowledgements

We thank the editors and anonymous reviewers for their valuable comments and suggestions, which have greatly improved this manuscript. This work was supported by the National Natural Science Foundation of China (Grant No. 52178357 and U23A2046), the 2020 Tianfu Technology Elite Project of Sichuan Province (Tianfu 10000 Talents Program; Grant No. 658), the Sichuan Youth Sci-Tech Innovation Team Project (Grant No. 2022JDTD0007), the Sichuan Science and Technology Program (Provincial Academy-University Cooperation Project) (Grant No. 2024YFHZ0022 and 2025YFHZ0015), the Natural Science Starting Project of Southwest Petroleum University (Grant No. 2022QHZ013 and 2022QHZ023).

Author contributions

Y. Y.:Writing–review&editing, Writing–original draft, Validation, Methodology, Investigation, Conceptualization. L. J.:Writing–original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. Q. H.:Writing–review & editing, Investigation. Q. C.:Writing–review&editing,Conceptualization. Q. Z.:Writing–review&editing, Conceptualization. X. S.:Writing–review & editing, Methodology.

Declarations

Competing interests

The author(s) declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025