Journal of Network Intelligence ©2025 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **10**, Number **4**, November **2025**

Super-resolution of UAV Images Based on Improved Generative Adversarial Networks

Tien-Wen Sung, Zheng-Jiang Xiao*, Qingjun Fang

Fujian Provincial Key Laboratory of Big Data Mining and Applications
Fujian University of Technology, Fuzhou 350118, China
tienwen.sung@gmail.com, 1594525797@qq.com, 89100941@qq.com

You-Te Lu

Department of Information and Communication
Southern Taiwan University of Science and Technology, Tainan 710301, Taiwan
yowder@stust.edu.tw

Thi-Minh-Phuong Ha

Information Technology Center
Hanoi Law University, Hanoi 10000, Vietnam
phuonghtm@hlu.edu.vn

*Corresponding author: Zheng-Jiang Xiao

Received April 23, 2024, revised December 9, 2024, accepted August 28, 2025.

Abstract. *The super-resolution algorithm combined with the generative adversarial network produces images that more accurately preserve the details and textures of real high-resolution images. so as to help UAVs improve the quality of their own images at a lower cost, so as to meet the applications in agriculture, geographic surveying and mapping and other fields. Therefore, this study proposes an improved super-resolution algorithm for generating adversarial networks, aimed at improving the quality of aerial images captured by UAVs. Firstly, an improved one-dimensional convolutional attention mechanism is designed and integrated into the generator model, by enhancing the model's focus on contextual information and reducing the parameters of the generator in the generative adversarial network, the process of image reconstruction is improved. Then, the discriminator is transformed into a U-shaped network with skip connections, allowing it to simultaneously learn global and local features, thereby enhancing its ability to learn from real images. By comparing image quality and image perception quality as objective data, experimental results demonstrate that the method proposed in this study achieves high precision in super-resolution reconstruction and rich texture details.*

Keywords: UAV images; super-resolution images; generative adversarial networks; attention mechanisms

1. **Introduction.** As an unmanned aerial vehicle, UAVs can fly in the air and perform a variety of tasks, such as taking photographs and videos, and performing ground monitoring, which makes them highly useful in the fields of agriculture [1,2], ocean monitoring [3], environmental monitoring [4], geographic information systems [5], and other fields.

With the advancement of UAV technology, UAVs have been greatly improved in terms of operation, control, and safety. Compared with traditional aircraft, UAVs have the advantages of smaller size and lower cost, so they can provide users with more affordable options and become a new type of image acquisition tool that can be widely used. However, UAV images can encounter many problems, including low image quality, low image resolution, blurry images, high image noise, etc. These problems may make it difficult for UAV images to be effectively identified and processed, which will affect the application of UAV images in agriculture, marine monitoring, environmental monitoring, geographic information systems and other fields. For example, in traffic highway inspections, the transmission of drone signals is easily interfered with, resulting in low-resolution and noisy images transmitted by drones [6], which can mask the extent of damage to cracks and gutters in the path, making it more difficult for monitors to observe and repair [7,8].

One method that is frequently used to efficiently increase the resolution of images is image super-resolution reconstruction [9–11]. The amount of pixels per square inch in an image, or pixel density, is generally referred to as image resolution. The information from one or more low-resolution photographs can be combined with image super-resolution technology to create high-resolution images. In this approach, image super-resolution technology can be applied to improve the quality and clarity of an image, regardless of the original image's low resolution. greater pixels mean greater detail, clarity, and information may be seen in high-resolution photographs.

The most direct way to improve the UAV's capture of high-resolution images is to upgrade the camera equipment mounted on the UAV, and enhance the image quality by loading better chips and sensors. However, this method will increase the cost of UAVs highly, and the iteration is slow, the improvement is limited, and it is not suitable for widespread use, resulting in the inability to meet people's needs for high-resolution UAV images. Therefore, for the images taken by UAVs, the use of image super-resolution technology can make the images clearer and more detailed, thereby improving the image quality, and reducing the cost of UAV shooting, and the iteration speed of the algorithm is faster than that of hardware equipment.

In summary, research into super-resolution algorithms for UAV images is critical for improving image quality, boosting visual effects, extracting information, and lowering UAV costs.

1.1. **Related work.**

As one of the main sources of information obtained by human beings and one of the main communication media on the Internet, images are becoming more and more closely integrated with the widespread application of artificial intelligence technology [12–15]. Therefore, image super-resolution algorithms have evolved into two main categories: classic interpolation-based approaches and deep learning-based methods.

Classic interpolation-based super-resolution algorithms are usually implemented by inserting new pixels between existing pixels, thereby increasing the resolution of the image [16]. For instance, the bicubic [17] interpolation method takes into account the 16 nearest pixels in the target image. It then weights each point according to its distance from the target and gray value. Finally, the weighted superposition method takes into account the target pixel's gray value. Such processing can estimate the value of the target pixel more accurately, thus realizing high-quality interpolation of the image. Nevertheless, this kind of super-resolution algorithms can only be applied to smooth images, the processing effect on complex images is not ideal, and can only reconstruct known information, can not add new information. Therefore, this type of method is commonly used in the enlargement and reduction of images.

The mapping relationship between low-resolution and high-resolution images is learned by the deep learning-based super-resolution technique. To accurately capture these relationships, a substantial amount of data is required for training. As deep learning technology progresses, notably convolutional neural networks (CNN), more super-resolution algorithms utilize deep learning techniques. Dong et al. first proposed SRCNN [18], a deep learning model for super-resolution tasks, which is mainly composed of three types of convolutional layers, which are used for image extraction, nonlinear mapping, and reconstruction. Initially, the low-quality picture undergoes enhancement via the conventional interpolation method to enlarge it to the dimensions of the high-quality picture. Afterward, the characteristics undergo additional extraction and alteration through a convolutional layer, allowing the acquired traits to better fit the reconstruction of high-quality pictures. Ultimately, the high-resolution pictures are generated via the reconstruction layer. Later, the research team proposed FSRCNN [19] on the basis of SRCNN, which can not only use low-resolution images as input, but also increase the network depth, and change the final reconstruction layer to the structure of upsampling and deconvolution layer to obtain high-resolution images, which has faster processing speed and better results than SRCNN. However, as the quantity of network layers rises, the issues of model deterioration and sluggish convergence become increasingly severe in deep learning.

Additionally, there exists a super-resolution algorithm that integrates deep learning with a Generative Adversarial Network (GAN) [20]. Kim et al. [21] were the first to incorporate ResNet [22] into the super-resolution algorithm, addressing the slow convergence issue of conventional convolutional networks and enhancing the quality of the resulting super-resolution images. Following Kim and colleagues' work, Ledig and team integrated the generative adversarial network with image super-resolution, presenting the SRGAN [23] network model. This model introduced the perceptual loss function, enhancing the visual fidelity perceived by the human eye in the reconstruction outcomes, thus rendering the generated image more akin to reality in both detail and style. In 2018, Wang et al. [24] replaced the residual network in SRGAN with a new residual network RRDB combined with DenseNet [25] on the basis of SRGAN, and used a more aggressive adversarial loss function of generator and discriminator, which well removed the artifacts in the SRGAN generated images and improved the presentation of texture details in the generated images. Since 2020, to tackle the issue of GAN' inability to extract global feature information, numerous researchers have integrated attention mechanisms [26, 27] into GAN, such as the GDCA model in which Nguyen et al. [28] fused channel attention mechanisms with SRGAN, which effectively improved the texture details and visual effects of SRGAN-generated images.

1.2. **Motivation and contribution.** While traditional super-resolution algorithms offer the advantages of simplicity and speed, they heavily rely on neighborhood information, often resulting in relatively blurred edges in the reconstructed image. The deep learning-driven super-resolution algorithm leverages the robust feature extraction prowess of convolutional neural networks to proficiently revive the intricacies of drone imagery, even during high-level magnification reconstructions. This ensures that the image edges retain their clarity, thereby substantially elevating the super-resolution image quality. However, two main challenges arise when applying deep learning methods. Firstly, with a small number of network layers, the model may not extract sufficient features, resulting in unclear texture details. Moreover, augmenting the network layer count may result in performance deterioration.

To effectively enhance the quality of UAV aerial images, an enhanced model called ESRGAN-UA was proposed based on the super-resolution model ESRGAN. The study compared experimental results and introduced key innovations and contributions as follows:

(1) The GAN model struggles with feature expression, and combining it with the traditional attention mechanism often leads to increased model complexity. To tackle this

concern, we introduce a streamlined attention mechanism crafted to bolster model efficacy sans escalating complexity.
(2) The discriminator in the GAN model lacks effective guidance for the generator and struggles to incorporate both global features and local details. To address this issue, we redesigned the conventional VGG [29] classification discriminator into a U-shaped discriminator with skip connections. This transformation enables the discriminator to provide more detailed pixel feedback to the generator, better assessing the realism of each pixel, and facilitating the learning of both global features and local details.
(3) Addressing the challenge of training the GAN mode, we incorporate Spectral Normalization Regularization [30,31] into the discriminator, adding a Lipsitz stability constraint to enhance its performance. This constraint can make the model have strong stability when dealing with input perturbations, which makes the training process of the network more reliable, and can effectively suppress artifacts.
(4) The ESRGAN model produces richer and more detailed image texture details compared to the SRGAN model. However, it demonstrates inferior performance in image quality evaluation metrics like PSNR [32] and SSIM [33]. To provide a more objective and comprehensive assessment of model performance, LPIPS [34] was introduced as an additional metric for image perception quality evaluation. This enables assessment of the model's performance from two viewpoints: image quality and perceptual image quality.

2. The proposed ESRGAN-UA model.

2.1. ESRGAN model.

ESRGAN is a super-resolution model that utilizes GAN. The concept of GAN is influenced by zero-sum games in game theory, where one player's gains are offset by another player's losses, resulting in a net gain or loss of zero. Consequently, the entire GAN model comprises two neural networks: a generator and a discriminator. The generator aims to produce data resembling real data, while the discriminator endeavors to distinguish between real and generated data.

The generator's network structure diagram, depicted in Figure 1, takes a low-resolution image (LR) as input. LR's feature map is extracted through convolutional and PreLu activation layers within the model. Subsequently, the feature map passes through 23 RRDB sequentially, and the RRDB structure diagram is shown in Figure 2 and Figure 3, which is composed of a convolutional layer and a LeakyRelu activation layer. The RRDB module extracts features from the previous feature map deeply to obtain a new feature map. It then proceeds with upsampling layers to increase image size and, finally, converts the feature map dimension to 3 through a convolutional layer to output the super-resolution image (SR).

The network model structure diagram of the discriminator, depicted in Figure 4, is crafted to distinguish between authentic high-resolution images and super-resolution images generated by the generator. Initially, the high-resolution image slated for classification undergoes processing into a feature map via a succession of eight convolutional blocks, each comprising a convolutional layer followed by a LeakyReLU activation layer. Subsequently, the resulting feature map is segmented into non-overlapping regions using average pooling. These segmented regions are then interconnected through a fully connected layer. Ultimately, the quality score of the input high-resolution image is produced via a Sigmoid activation layer. Typically, the score of an authentic high-resolution image tends to surpass that of a generated super-resolution image. This prompts the generator to craft super-resolution images closely mirroring authentic high-resolution counterparts.

Compared to the loss function of traditional GANs, ESRGAN employs a more aggressive adversarial function. This function introduces a relative disparity between the real

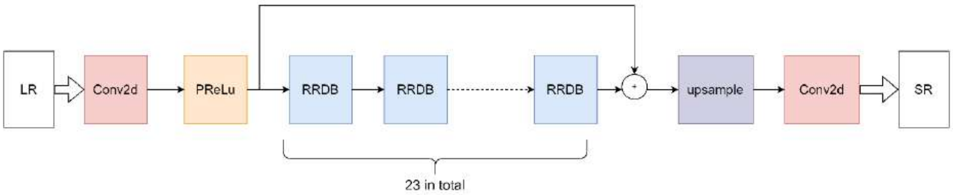

FIGURE 1. The ESRGAN generator model

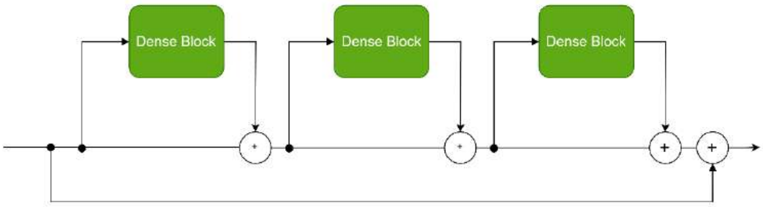

FIGURE 2. The RRDB structure

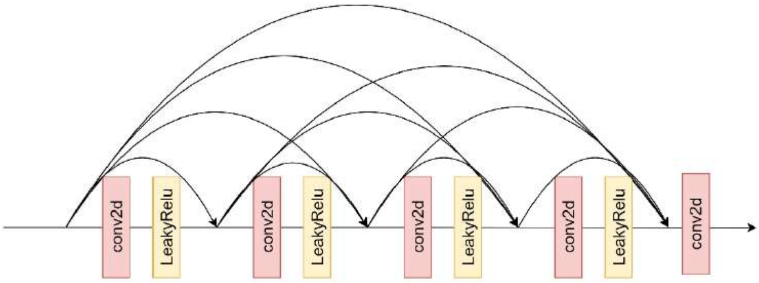

FIGURE 3. The Dense Block structure

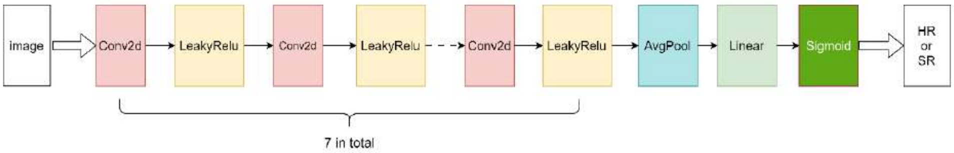

FIGURE 4. The ESRGAN discriminator model

image and the generated image. In simple terms, it enables the discriminator to quantify how much more realistic the real image is compared to the generated image. This is calculated using Formulas (1) and (2):

$$\begin{aligned} L_D =& -E_{x_1\sim P_r, x_2\sim P_f}\left[\log\left(\delta\left(D\left(x_1\right)-D\left(x_2\right)\right)\right)\right] \\ & - E_{x_1\sim P_r, x_2\sim P_f}\left[\log\left(1-\delta\left(D\left(x_2\right)-D\left(x_1\right)\right)\right)\right] \end{aligned} \tag{1}$$

$$\begin{aligned} L_G^{'} =& -E_{x_1\sim P_r, x_2\sim P_f}\left[\log\left(1-\delta\left(D\left(x_1\right)-D\left(x_2\right)\right)\right)\right] \\ & - E_{x_1\sim P_r, x_2\sim P_f}\left[\log\left(\delta\left(D\left(x_2\right)-D\left(x_1\right)\right)\right)\right] \end{aligned} \tag{2}$$

In the above two formulas, P_r denotes the real image, P_f denotes the generated image, $\delta(\cdot)$ signifies the Sigmoid function, $D(\cdot)$ denotes the mass fraction returned by the discriminator, E denotes the mathematical expectation, L_D denotes the discriminator's loss function, and $L_G^{'}$ denotes the generator's content loss function. The described formulas quantify the gap between the real image and the generated image, evaluate the disparity between the generated image and the real image, and iteratively refine the generated image to approach the real image.

ESRGAN uses the same perceived loss [23, 35] as SRGAN, but differs in that it uses the features that precede the activation layer for calculation. This discrepancy arises because the features post-activation tend to be highly sparse, particularly in deep networks. Such sparse activation offers weak supervision, leading to suboptimal performance. Additionally, utilizing activated features can induce inconsistencies in brightness between the reconstructed and original images. The formulas for computing perceptual loss and generator loss are as follows:

$$L_p = \frac{1}{hwc} \sum_{i,j,k} \left(\varphi^l_{i,j,k} \left(I^{HR} \right) - \varphi^l_{i,j,k} \left(I^{SR} \right) \right)^2 \tag{3}$$

$$L_G = L_p + \ \alpha L^{'}_G + \ \beta L_1 \tag{4}$$

In the above two equations, L_p denotes the perceptual loss function, L_G denotes the generator loss function, I^{HR} denotes a high-resolution image, I^{SR} denotes a super-resolution image generated by the generator, $\varphi^l_{i,j,k}$ denotes the feature map of the l-th layer of the VGG16, where h, w, and c stand for the height, width, and number of channels of the feature map of the l-th layer, respectively. Additionally, α and β are both constants.

The ESRGAN model, characterized by deep network layers and high complexity, boasts a considerable number of parameters. However, despite its sophistication, the super-resolution images it generates often lack sufficient feature expression ability. For instance, pixel features at the edges may sometimes overshadow those of the main subject, leading to poor performance in common image quality evaluations. To tackle these challenges, our study introduces an image super-resolution reconstruction model termed ESRGAN-UA, an enhancement over the original ESRGAN.

2.2. **Improved attention mechanisms.** Attention mechanisms in deep learning take various forms, with common types including channel-domain attention, spatial attention, and mixed-domain attention. Among these, spatial attention predominantly concentrates on modulating weights across various spatial positions in the feature map, accentuating the significance of each position for the given task. By learning spatial attention weights, the network can prioritize spatial locations that are more influential for the current task, enhancing its perceptual ability. However, this may lead to a reduction in the richness of extracted features by the network.

Conversely, the channel attention mechanism fine-tunes weights among distinct channels in the feature map to underscore the significance of each channel for the task. By acquiring knowledge of channel attention weights, the network can prioritize channels that exert greater influence on the task, thereby augmenting its representational capacity. However, this may result in reduced sensitivity to spatial information in the image.

Lastly, the hybrid attention mechanism combines channel attention and spatial attention simultaneously, adjusting weights across both channels and spatial positions in the feature map. Through acquiring knowledge of cross-channel and spatial attention weights, the network can adeptly capture crucial channels and spatial positions within the feature map, thereby bolstering performance and enhancing generalization capability. However, implementing the mixed-domain attention mechanism is more complex compared to single attention mechanisms, potentially increasing computational costs and model parameters.

In summary, this study introduces a modified attention residual module. This module integrates a modified channel attention mechanism with a residual connection. By integrating these components, pivotal information within the image is accentuated to extract and fortify essential features, thereby amplifying the network's acuity towards details. Consequently, the refined features are fused with the original features through residual connections, facilitating the network in proficiently learning and reconstructing the image's intricate structure. The configuration of the comprehensive attention residual module is delineated in Figure 5.

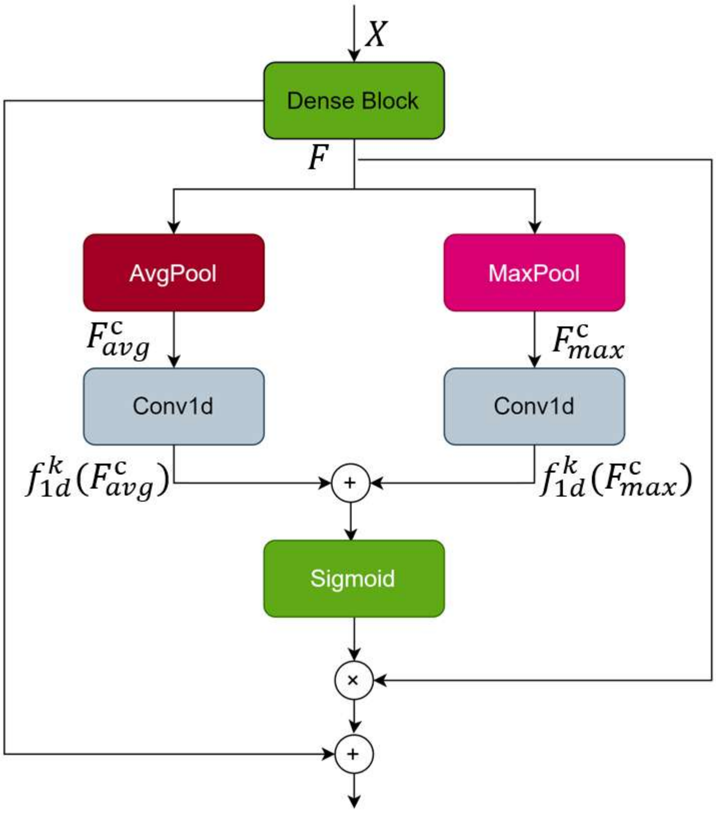

Figure 5. Dense Block with an attention mechanism

Initially, global average pooling and global maximum pooling were employed to aggregate spatial information from the feature map F extracted from the Dense Block, resulting in channel descriptors denoted as F^c_{avg} and F^c_{max}. Subsequently, a one-dimensional convolution with a kernel length of k was applied to integrate information within k neighborhoods of the channel. The information from the two pooling layers was aggregated using the sigmoid function, identifying the important regions in the feature map. Finally, the weight of this region is multiplied by the feature map before pooling, allowing the network to selectively enhance important features. This process facilitates more efficient extraction and utilization of information within the image, thereby ultimately enhancing the network's performance and generalization capability across diverse tasks.

To circumvent the issue of gradient vanishing in deep networks, the feature map acquired via the attention mechanism is merged with the original feature map through residual connection. The amalgamated feature map is subsequently incorporated into the backbone network, constituting a residual network structure. This design preserves the critical features extracted by the attention mechanism, while directly passing them to subsequent layers through residual connections. This approach aids in better gradient flow within the deep network, thereby enhancing the network's training stability and performance. The calculation process is as follows:

$$M_A(F) = F \cdot \ \sigma\left(f_{1D}^{k}\left(F_{avg}^{c}\right) + f_{1D}^{k}\left(F_{\max}^{c}\right)\right) \tag{5}$$

$$X = X + \ M_A(F) \tag{6}$$

In the above two equations, X denotes the feature map before adding the residual block, σ represents the Sigmoid function, F represents the feature map passed down from the Dense Block, and f_{1D}^{k} denotes a one-dimensional convolution operation with a kernel size of k. The parameter matrix obtained by our attention mechanism after applying a one-dimensional convolution with a kernel size of k is expressed as follows:

$$W^u = \begin{bmatrix} w_1^1 & \cdots & w_1^k & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w_2^2 & \cdots & w_2^{k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w_C^{C-k+1} & \cdots & w_C^C \end{bmatrix} \tag{7}$$

In the above equation, w_i^j represents the channel weight of the j-th neighboring channel for the i-th channel. Through this parameter matrix, it can be guaranteed that each grouping has a certain correlation, so as to reduce the loss of information, and speed up the calculation, for a certain channel of the feature map, only need to consider its common influence with k adjacent channels in the calculation. As depicted in the following formula:

$$f_i = \sigma(\sum_{j=1}^{k} w_i^j y_i^j), \ y_i^j \in F_i^k \tag{8}$$

In the above equation, F_i^k denotes the channel information of the k-th neighboring region of the i-th channel in the feature map F, and f_i denotes the he focal information of the i-th channel in the feature map F. Facilitating the interaction of local cross-channel information requires capturing an appropriate range. Thus, we adopted an adaptive formula to compute the convolution kernel size. By analogy with the proportional relationship between the quantity of channels and the convolution length in grouped convolutions, we determined that the coverage range for cross-channel information interaction is contingent upon the channel quantity c. Expressed as follows is this non-linear mapping function:

$$c = \ 2^{n \cdot k - m} \tag{9}$$

Therefore, the convolution kernel is calculated as follows equation (10):

$$k = \ \tau\left(\frac{\log_2 c}{n} + \frac{m}{n}\right) \tag{10}$$

In the above two equations, $\tau\left(x\right)$ denotes the nearest odd integer to x, n and m are constants, usually n take 2 and m take 1.

Pooling is indispensable for better information aggregation from the feature map and reduction of parameter count. However, traditionally, only average pooling has been commonly used. We believe that incorporating both maximum pooling and average pooling methods can provide different representations of features, and the differences between the features aggregated by these two methods can help generate a more detailed attention channel map. Therefore, we consider increasing the use of maximum pooling to be beneficial. To achieve this, we perform both global average pooling and global maximum pooling on the transmitted feature map. Subsequently, we employ one-dimensional convolution to individually fuse the channel features acquired from these two pooling methods. Finally, we adjust the weight of this region by the feature map pre-pooling, enabling the model to concentrate more on the crucial areas of the image.

Although the CBAM [36] attention mechanism allows the network to focus on contextual information, it compresses channels, leading to the network discarding some feature map information during computation, ultimately reducing the feature expression capability of the attention mechanism. Therefore, in this study, we directly perform one-dimensional convolution on the channel feature map after global pooling, allowing local cross-channel information interaction with a convolution kernel size of k. This approach avoids feature map compression and preserves feature map information. Given that cross-channel information interaction encompasses solely k parameter information, the need for weight sharing among all channels results in a reduced parameter count compared to the usage of CBAM.

2.3. **U-Shaped Discriminator.** The primary role of the discriminator is to comprehend the data distribution, acting as a component of the adversarial loss function within the generative network. This facilitates the provision of learning signals for the generator, aiding it in generating outputs that are increasingly realistic. However, the discriminator is often used as a classification network, focusing solely on learning the most discriminative differences between real and synthetic images. As a result, it typically either focuses on global structures or pays more attention to local details. Requesting the discriminator to simultaneously learn both global semantics and details may result in the loss of its expressive power.

Therefore, we have transformed the network structure of the discriminator from the traditional VGG classification model to a U-shaped network model with skip connections. This allows the discriminator to provide more detailed pixel feedback to the generator while maintaining the overall structure of the image. The U-shaped discriminator network structure is shown in Figure 6.

This network is an encoder-decoder network, where from left to right in Figure 6 are the encoder, decoder, and classifier. The encoder expands the feature channel count to eight times that of the input channels using three convolutional layers, while gradually reducing the resolution of the feature maps through three downsampling layers. The decoder performs upsampling, mapping high-dimensional feature maps to low-dimensional ones through three convolutional layers, and then obtaining high-resolution feature maps through upsampling layers. The dashed lines with arrows denote skip connections, which facilitate the transmission of information between the encoder and decoder for images of the same resolution, ensuring that the number of channels in the feature maps remains consistent with the original input. Finally, the classifier identifies the maximum value of the classification features as the prediction result.

To merge more low-level features and maintain the integrity of details within the feature maps, the U-Net network used for semantic segmentation adopts non-padding convolution to connect the cropped upsampled feature maps with the downsampled feature maps. However, this cropping operation leads to a mismatch in the sizes of the input and output images, resulting in the loss of a considerable amount of information in UAV aerial images. For example, in experiments, we found that the texture details of roads in UAV aerial images are ignored, thereby affecting the super-resolution effect of the road areas.

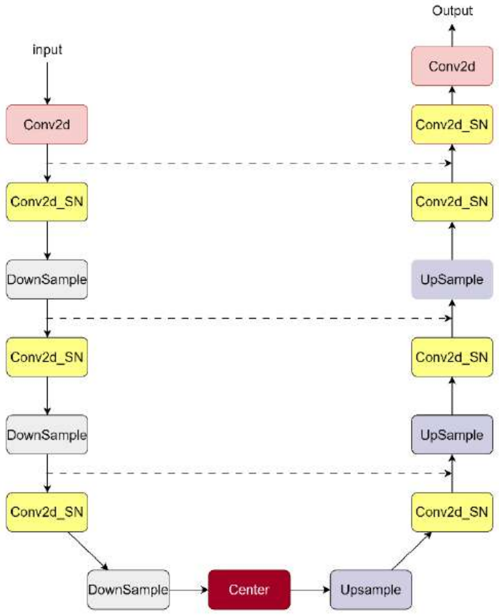

Figure 6. U-shaped Discriminator

Therefore, we abandoned the cropping strategy to achieve consistency in input and output sizes. This improvement allows the skip connection structure to better integrate detailed features, reduce information loss, and decrease the network's size and parameters. As a result, the computing speed of the network is not only improved, but also the image segmentation effect is maintained.

2.4. **Spectral Normalization Regularization.** Although modifying the discriminator model of ESRGAN enables it to learn both global and local feature information simultaneously, and reduces the network's size, the addition of skip connections also increases the complexity of the model, leading to training instability. Therefore, we have incorporated Spectral Normalization Regularization into the discriminator and added a Lipshitz stability constraint [37] to the discriminator, which is calculated as follows:

$$\frac{||D\left(x_1\right) - D(x_2)||_2}{||x_1 - x_2||_2} \leq K, \forall x_1, x_2 \tag{11}$$

In the above equation, $||\cdot||$ is denoted as 2 norms, K is a constant.

By adding Lipschitz stability constraints, neural networks exhibit strong stability when handling input perturbations. Even with minor changes in input data, the network's output remains relatively stable. This stability enhances the reliability of network training and facilitates reaching a good convergence state more easily. For instance, when a slight modification is made to a few pixels of an image A to obtain another image B, if the discriminator assigns vastly different scores to these two images, it indicates instability in the discriminator, which is detrimental to model training.

3. Experimental results and analyses.

3.1. **Experimental conditions and dataset used.** Our research was trained on three publicly available datasets: DIV2K, OutdoorSceneTraining, and UavidTrain, with the first two datasets being high-resolution high-frequency information datasets and the last being high-resolution drone aerial images. During the training phase, we randomly cropped regions of 96×96 pixels from the training set images to serve as high-resolution images (HR). Subsequently, we obtained corresponding low-resolution images (LR) by compressing the HR images using four-fold JPEG compression. This not only enriches the dataset and avoids overfitting, but also helps the discriminator learn the detailed texture of the high-resolution image to guide the generator to produce a more detailed image. In this study, training was conducted using the NVIDIA GEFORCE RTX 3060 8G version of the graphics card, utilizing a batch size of 16 samples per iteration. We employed the Adam optimizer and executed a total of 200 epochs of training. The learning rate was initially established at 1×10^{-4} for the initial 100 epochs, followed by a reduction to 1×10^{-5} for the subsequent 100 epochs.

3.2. **Evaluation metrics.** To evaluate the effectiveness of the enhanced algorithm, this study utilized two assessment criteria: image quality and image perceptual quality. The evaluation of image quality is based on PSNR and SSIM. PSNR serves as a prevalent metric for signal reconstruction quality, extensively employed in fields like image compression. It quantifies the correlation between the maximum signal and background noise of both the reference and reconstructed images, with higher values indicating reduced image distortion. There are three evaluation methods for PSNR evaluation of color pictures. The first involves calculating the PSNR for each of the three RBG channels and then taking the average. The second method computes the average variance of the three RGB channels divided by three. The third converts the image to YCbCr format and solely calculates the PSNR of the Y component, representing luminance. We adopted the third method to evaluate the PSNR of color images. The formula is as follows:

$$PSNR = 10 \cdot \ log\frac{MAX^2}{MSE} \tag{12}$$

$$MSE = \frac{1}{W * H}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left[I_{HR}(i,j) - I_{SR}(i,j)\right]^2 \tag{13}$$

In the provided formulas, MAX denotes the maximum pixel value within the image, while W and H denote the width and height of the image, respectively. I_{HR} signifies the pixel value of the real image, I_{SR} signifies the pixel value of the generated image.

Considering properties such as brightness, contrast, and image structure, SSIM evaluates the structural similarity between two images. SSIM measures image similarity by computing the mean, variance, and covariance. Higher SSIM values indicate greater resemblance between the images. The SSIM formula is as follows:

$$SSIM(x,y) = \frac{(2u_x u_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{14}$$

The equation above includes u_x denoting the mean of x, u_y representing the mean of y, σ_x^2 signifying the variance of x, σ_y^2 indicating the variance of y, and σ_{xy} representing the covariance of x and y. Additionally, two constants, denoted as c_1 and c_2, are involved in the computation.

The other metric is based on the LPIPS image perception quality assessment standard, which extracts features using deep learning models and calculates and compares the feature differences to get the perceptual similarity of the two images. Greater image similarity is indicated by a lower LPIPS value. This evaluation approach is more in line with human perception compared to the first standard. The formula for calculating LPIPS is as follows:

$$d\left(I_{HR}, I_{SR}\right) = \sum_{l} \frac{1}{H_l W_l} \sum \left|\left|w_l \odot \left(\widehat{y}^l - \widehat{y}_0^l\right)\right|\right|_2^2 \tag{15}$$

In the above equation, H_l and W_l denote the height and width of the feature map at layer l of the LPIPS network, respectively, $\widehat{y}^l$ denotes the feature map of I_{HR} at layer l, $\widehat{y}_0^l$ denotes the feature map of I_{SR} at layer l, $\odot$ for scaling, and w_l denotes the layer l vector of the LPIPS.

In the evaluation process, we perform tests by comparing the reconstructed images with the corresponding original high-resolution counterparts. For each image in the test set, we calculate its PSNR, SSIM, and LPIPS values. Finally, we average these three evaluations metrics to obtain the final test results.

3.3. **Comparative experiments.** We designated the ESRGAN model that solely integrates the attention mechanism as ESRGAN-A. Furthermore, the model incorporating the U-shaped discriminator network on top of ESRGAN-A was named ESRGAN-UA. To validate the optimization outcomes of the proposed model, three enhanced versions of SRGAN [28,38,39] and two improved versions of ESRGAN [40,41] were replicated. Ablation experiments were conducted, comparing them with ESRGAN+CBAM. Ten models in total were assessed on three test sets: BSDS100, UavidTest, and Urban100, with increasing image complexity. The PSNR, SSIM, and LPIPS values for each test set, with a scaling factor of 4, were acquired, as depicted in the tables presented in Table 1 and Table 2:

Table 1. Image quality assessment of each model on three test sets (PSNR/SSIM)

Model	BSDS100	UavidTest	Urban100
SRGAN [23]	24.9413/0.6486	**24.7253**/0.6176	23.3626/0.7021
GDCA [28]	25.0611/0.6670	24.4771/**0.6179**	23.3484/0.7016
Zhang et al. [38]	24.3937/0.6630	23.9540/0.6034	22.7358/0.6857
Xie et al. [39]	20.9553/0.5494	20.3521/0.5232	19.8173/0.5274
ESRGAN [24]	24.1513/0.6324	23.2754/0.5734	22.3430/0.6789
ESRGAN+CBAM	22.5015/0.5395	22.1067/0.4844	21.0179/0.6055
ESRGAN-A	24.5974/0.6404	23.4251/0.5863	22.5332/0.6744
SOUPGAN [40]	24.8641/0.6481	23.7256/0.5878	23.2172/0.6925
SOCAGAN [41]	24.6747/0.6265	23.6972/0.5728	23.1036/0.7052
ESRGAN-UA	**25.6289/0.6804**	24.5476/0.6150	**23.5330/0.7270**

Upon comparing the two evaluations in Table 1 and Table 2, it is evident that the integration of the CBAM attention mechanism into the ESRGAN model has led to a reduction in the values of the three assessments. However, the ESRGAN-A with our improved attention mechanism is improved compared to ESRGAN, and only the SSIM evaluation on the Urban100 dataset is slightly inferior. The image detail of the Urban100 data is the most complex and abundant among the three test sets, which means that our ESRGAN-A outperforms ESRGAN in super-resolution for low- to medium-to-medium texture complex images.

On the other hand, as depicted in Table 1, the proposed ESRGAN-UA model exhibits only a slight decrement in PSNR and SSIM evaluations within the UavidTest dataset. However, it remains among the top performers, showcasing optimal performance in other datasets. Although the texture complexity of the UavidTest dataset is not as high as that of the Urban100 dataset, the information contained in a single image is the highest

Table 2. Image quality assessment of each model on three test sets (LPIPS)

Model	BSDS100	UavidTest	Urban100
SRGAN [23]	0.0830	0.0982	0.1172
GDCA [28]	0.0840	0.0975	0.1268
Zhang et al. [38]	0.0838	0.1012	0.1184
Xie et al. [39]	0.1390	0.1628	0.2381
ESRGAN [24]	0.0916	0.1016	0.1275
ESRGAN+CBAM	0.1158	0.1225	0.1540
ESRGAN-A	0.0802	0.1011	0.1206
SOUPGAN [40]	0.0869	0.0993	0.1250
SOCAGAN [41]	0.0888	0.1061	0.1382
ESRGAN-UA	**0.0752**	**0.0846**	**0.1062**

because it is a high-altitude aerial image taken by a UAV. This shows that our improved discriminator model can improve the problem of insufficient detail texture richness in the ESRGAN-A model, and can further extract the overall information within the image. Reviewing Table 2, it becomes evident that the proposed ESRGAN-UA algorithm outperforms others in the evaluation of image perceptual quality. This outcome underscores the superior super-resolution efficacy of the ESRGAN-UA model, particularly in enhancing image detail texture.

To validate the effectiveness of the attention mechanism utilizing one-dimensional convolution as mentioned in this study, we utilized the parameters() function provided by PyTorch to ascertain the parameter count within both the generator and discriminator networks. The proposed model in this study is an enhanced version of ESRGAN, exhibiting a substantial variation in depth compared to the SRGAN class. Consequently, the parameter count between these two model types will differ. Therefore, we compared our model with the SOCAGAN model within the same ESRGAN system, and the parameter counts are illustrated in Table 3.

Table 3. Model parameter

Model	Generator	Discriminator
ESRGAN	16910406	23565505
SOCAGAN	17269206	23565505
ESRGAN+CBAM	16921653	23565505
ESRGAN-UA	**16910275**	**4376832**

As depicted in Table 3, the parameters of the generator in our proposed ESRGAN-UA model demonstrate a decrease of 131 following the integration of the attention mechanism module. Conversely, the generator model parameters of SOCAGAN with the CA attention mechanism experience a significant increase. This divergence arises from our enhanced attention mechanism's localized interaction with cross-channel information within a region using one-dimensional convolution with a kernel size k. Moreover, all channels share weight information, thereby mitigating parameter proliferation. Since SOCAGAN maintains the same discriminator model as ESRGAN, we utilized an identical discriminator model for reproducibility, ensuring consistency in discriminator model parameters. We can see that compared with ESRGAN, the parameters of the proposed discriminator model are greatly reduced, only 18.57% of the parameters of the original discriminator model. This is because our discriminant network adopts a fully convolutional layer, which reduces the fully connected layer compared with the VGG classification model, so the discriminator parameters are also greatly reduced.

In summary, compared to the performance decline observed in ESRGAN with the addition of the CBAM attention mechanism, our attention mechanism not only effectively enhances the performance of ESRGAN but also reduces the generator's parameters, consequently improving its training speed. We further augmented the model performance of ESRGAN-UA by modifying the discriminator model, leading to significant reductions in model parameters and further enhancing training speed.

3.4. **Analysis of Results.** To showcase the effectiveness of the proposed algorithm, we randomly selected an aerial image from the test set. We applied six models for comparison: Bicubic, SRGAN, GDCA, Zhang's model, ESRGAN, and SOUPGAN, each with a magnification factor of 4x. The reconstructed images are shown in Figure 7.

From Figure 7, it's evident that when processing high-frequency information is more frequent, ESRGAN displays numerous artifacts in regions with high-frequency information, whereas images generated by the Bicubic algorithm exhibit the lowest clarity. This also illustrates the limitations of traditional super-resolution algorithms, which perform poorly as the magnification factor increases. In order to see a more detailed comparison, we took two parts of the aerial image, the grass trail and the sidewalk, and compared them experimentally, resulting in Figures 8 and 9.

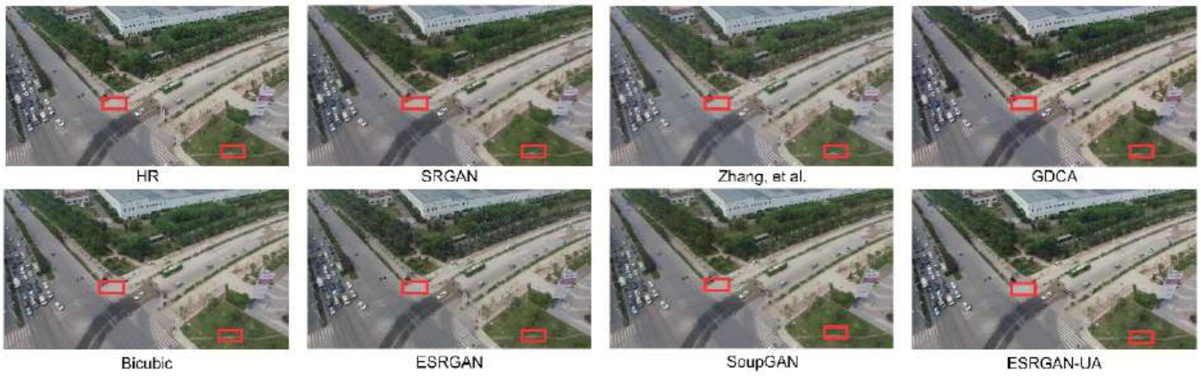

Figure 7. Comparison of high altitude aerial photos

In Figure 8, we can observe that the texture of the image generated by the Bicubic algorithm becomes difficult to discern after enlargement, with only the general shape visible. Comparing the line shapes of the pathways, our proposed ESRGAN-UA model produces the best results, with the highest similarity, smoother lines, clearer edges, and fewer artifacts. Additionally, comparing the texture of the grass beside the pathway, our generated grass details are richer, without severe overfitting like ESRGAN, and without clumping together like other models. The sense of hierarchy between the pathway and the grass is also more pronounced. This indicates that our proposed network has made significant improvements in enriching detail textures and effectively addressing the issue of edge blurriness.

As we can see from Figure 9, the Bicucub algorithm still handles the details in a blurred manner. While other models produce a certain number of artifacts, ours produces the fewest artifacts. SRGAN and Zhang et al.'s model has fuzzy zebra crossings, and there are many artifacts between the sharpened lines of the lines, while the zebra crossings of the ESRGAN model eliminate the blurring of edge lines, but the number of artifacts is still quite large. SOUPGAN has a lot of jaggedness on the lines and the lines are not

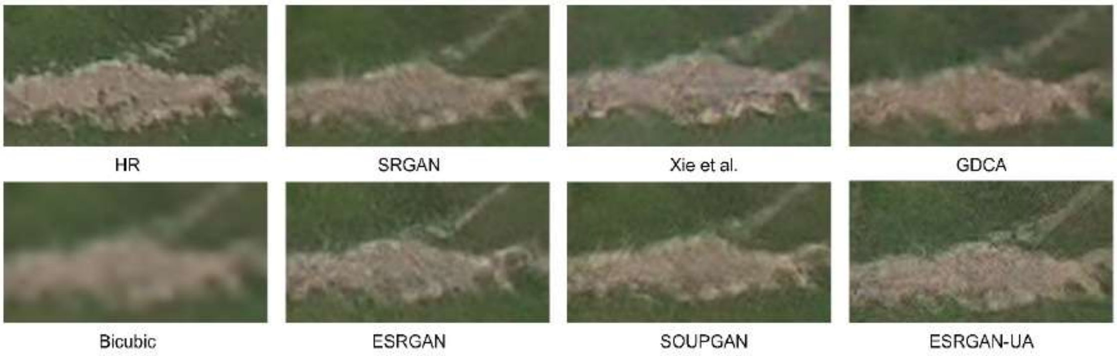

FIGURE 8. Comparison of high altitude aerial photos

soft enough. The closest to the HR diagram is GDCA and our model, but the edge lines of the zebra crossing in GDCA are blurred, while the lines generated by our model are smoother and softer, and the lines and the ground are more layered, which is weaker than the abruptness of ESRGAN.

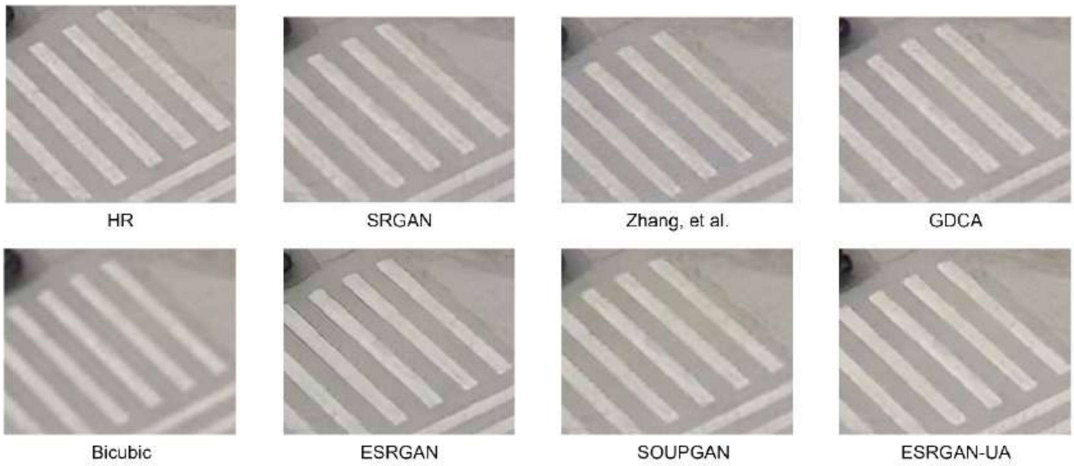

FIGURE 9. Comparison of high altitude aerial photos

In summary, the improved discriminator model proposed by us can not only provide richer details and textures for super-resolution graphs, deal with the problem of GAN edge blurring, but also further suppress the appearance of artifacts and reduce the difficulty of network training.

4. **Conclusion.** This study proposed a super-resolution model, ESRGAN-UA, which integrates an improved attention mechanism, thereby enhancing the extractor's ability to extract key information from images without increasing model complexity. Subsequently, we restructured the discriminator to act as both a classifier and a segmenter, allowing for better extraction of global features and local details to guide the generator in producing textures closer to those of real high-resolution images. Finally, by introducing Spectral Normalization Regularization, model training became more stable, suppressing the occurrence of artifacts. The experimental results affirm that our proposed method effectively enhances the texture details of super-resolution images while mitigating artifacts. However, there still exists the issue of insufficient handling of high-frequency information leading to overfitting, thus future research will focus on better extraction and separation of high-frequency information.

Acknowledgment. This work was partially supported by the Fujian University of Technology under Grant GY-Z20016 and GY-Z19005.

REFERENCES

[1] S. Ban, W. Liu, M. Tian, Q. Wang, T. Yuan, Q. Chang, and L. Li, "Rice leaf chlorophyll content estimation using uav-based spectral images in different regions," *Agronomy*, vol. 12, no. 11, p. 2832, 2022.

[2] A. Vázquez-Ramírez, D. Mújica-Vargas, A. Luna-Álvarez, M. Matuz-Cruz, and J. d. J. Rubio, "Real-time detection of bud degeneration in oil palms using an unmanned aerial vehicle," *Eng*, vol. 4, no. 2, pp. 1581–1596, 2023.

[3] J. Jessin, C. Heinzlef, N. Long, and D. Serre, "A systematic review of uavs for island coastal environment and risk monitoring: Towards a resilience assessment," *Drones*, vol. 7, no. 3, p. 206, 2023.

[4] K. Liu, L. Xia, and J. Xu, "An information management system of land resources based on uav remote sensing," *International Journal of Information and Communication Technology*, vol. 23, no. 2, pp. 107–125, 2023.

[5] Y. Takata, H. Yamada, N. Kanuma, Y. Ise, and T. Kanda, "Digital soil mapping using drone images and machine learning at the sloping vegetable fields in cool highland in the northern kanto region, japan," *Soil Science and Plant Nutrition*, vol. 69, no. 4, pp. 221–230, 2023.

[6] T. Wu, X. Guo, Y. Chen, S. Kumari, and C. Chen, "Amassing the security: An enhanced authentication protocol for drone communications over 5g networks," *Drones*, vol. 6, no. 1, p. 10, 2021.

[7] R. Zhao, Y. Huang, H. Luo, X. Huang, and Y. Zheng, "A framework for using uavs to detect pavement damage based on optimal path planning and image splicing," *Sustainability*, vol. 15, no. 3, p. 2182, 2023.

[8] B. Fan, Y. Li, R. Zhang, and Q. Fu, "Review on the technological development and application of uav systems," *Chinese Journal of Electronics*, vol. 29, no. 2, pp. 199–207, 2020.

[9] D. Sara, A. K. Mandava, A. Kumar, S. Duela, and A. Jude, "Hyperspectral and multispectral image fusion techniques for high resolution applications: A review," *Earth Science Informatics*, vol. 14, no. 4, pp. 1685–1705, 2021.

[10] Y. K. Ooi and H. Ibrahim, "Deep learning algorithms for single image super-resolution: a systematic review," *Electronics*, vol. 10, no. 7, p. 867, 2021.

[11] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016.

[12] T.-W. Sung, P.-W. Tsai, T. Gaber, and C.-Y. Lee, "Artificial intelligence of things (aiot) technologies and applications," *Wireless Communications and Mobile Computing*, vol. 2021, p. 9781271, 2021.

[13] Y. Xia, N. Ravikumar, J. P. Greenwood, S. Neubauer, S. E. Petersen, and A. F. Frangi, "Super-resolution of cardiac mr cine imaging using conditional gans and unsupervised transfer learning," *Medical Image Analysis*, vol. 71, p. 102037, 2021.

[14] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2154–2164.

[15] T.-W. Sung, B. Zhao, and X. Zhang, "An adaptive dimension differential evolution algorithm based on ranking scheme for global optimization," *PeerJ Computer Science*, vol. 8, p. e1007, 2022.

[16] N. Z. F. N. Azam, H. Yazid, S. A. Rahim *et al.*, "Super resolution with interpolation-based method: A review," *IJRAR-International Journal of Research and Analytical Reviews (IJRAR)*, vol. 9, no. 2, pp. 168–174, 2022.

[17] F. Aràndiga, "A nonlinear algorithm for monotone piecewise bicubic interpolation," *Applied Mathematics and Computation*, vol. 272, pp. 100–113, 2016.

[18] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13.* Springer, 2014, pp. 184–199.

[19] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14.* Springer, 2016, pp. 391–407.

[20] I. 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.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[24] 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*, September 2018.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[26] D. Zheng, W. Dong, H. Hu, X. Chen, and Y. Wang, "Less is more: Focus attention for efficient detr," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6674–6683.

[27] Y. Rao, D. Wu, M. Han, T. Wang, Y. Yang, T. Lei, C. Zhou, H. Bai, and L. Xing, "At-gan: A generative adversarial network with attention and transition for infrared and visible image fusion," *Information Fusion*, vol. 92, pp. 336–349, 2023.

[28] T. Nguyen, H. Hoang, and C. D. Yoo, "Gdca: Gan-based single image super resolution with dual discriminators and channel attention," *arXiv preprint arXiv:2111.05014*, 2021.

[29] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[31] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A spectral convolutional neural network model based on adaptive fick's law for hyperspectral image classification," *Computers, Materials & Continua*, vol. 79, no. 1, 2024.

[32] D. R. I. M. Setiadi, "Psnr vs ssim: imperceptibility quality assessment for image steganography," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8423–8444, 2021.

[33] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through fsim, ssim, mse and psnr—a comparative study," *Journal of Computer and Communications*, vol. 7, no. 3, pp. 8–18, 2019.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[35] R. Abbas and N. Gu, "Improving deep learning-based image super-resolution with residual learning and perceptual loss using srgan model," *Soft Computing*, vol. 27, no. 21, pp. 16 041–16 057, 2023.

[36] W. Wang, X. Tan, P. Zhang, and X. Wang, "A cbam based multiscale transformer fusion approach for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817–6825, 2022.

[37] B. Zhang, D. Jiang, D. He, and L. Wang, "Rethinking lipschitz neural networks and certified robustness: A boolean function perspective," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 398–19 413, 2022.

[38] T. Zhang and Y. Zhou, "Super-resolution algorithm combining attention mechanism with srgan network," *Journal of Chinese Computer Systems*, vol. 42, no. 12, pp. 2587–2591, 2021.

[39] H. Xie, T. Zhang, W. Song, S. Wang, H. Zhu, R. Zhang, W. Zhang, Y. Yu, and Y. Zhao, "Super-resolution of pneumocystis carinii pneumonia ct via self-attention gan," *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106467, 2021.

[40] K. Zhang, H. Hu, K. Philbrick, G. M. Conte, J. D. Sobek, P. Rouzrokh, and B. J. Erickson, "Soup-gan: Super-resolution mri using generative adversarial networks," *Tomography*, vol. 8, no. 2, pp. 905–919, 2022.

[41] J. Zhao, Y. Ma, F. Chen, E. Shang, W. Yao, S. Zhang, and J. Yang, "Sa-gan: A second order attention generator adversarial network with region aware strategy for real satellite images super resolution reconstruction," *Remote Sensing*, vol. 15, no. 5, p. 1391, 2023.