# Image Super-Resolution Reconstruction Method Based on Improved Generative Adversarial Network

Feng Xiong
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 1023465977@qq.com

Yu Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: ly1351737248@163.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: yujun@xatu.edu.cn

Chaoyi Dong
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 545861148@qq.com

Zhiyi Hu
Engineering Design Institute
Army Research Loboratory
Beijing, 100042, China
E-mail: 763757335 @qq.com

Hongpei Zhang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 2464094384@qq.com

***Abstract*—To address the challenges of low reconstruction accuracy and insufficient model generalization in image super-resolution (ISR) under complex degradation scenarios, this paper proposes an improved method that integrates generative adversarial networks (GAN) and vision transformers (ViT). First, in the generator module of Real-ESRGAN, some residual-in-residual dense blocks (RRDB) are replaced with ViT modules, leveraging the self-attention mechanism to enhance global feature modeling. This enables the model to better capture global information while preserving local details in complex scenes. Experimental results demonstrate that the improved model achieves PSNR gains of 0.59dB/0.45dB and SSIM improvements of 0.018/0.056 in ×2/×4 upscaling tasks on the Urban100 dataset, while also exhibiting excellent performance on benchmark datasets such as Set14. This method significantly enhances image reconstruction quality under complex degradation conditions, providing an effective technical solution for practical applications such as security surveillance, remote sensing monitoring, and target reconnaissance.**

***Keywords-Image Super-Resolution; Generative Adversarial Network; Vision Transformer; Self-Attention Mechanism***

## I. INTRODUCTION

Image super-resolution technology holds significant application value across multiple domains [1], particularly in critical fields such as medical imaging, video surveillance, and remote sensing imaging. With the increasing reliance on low-resolution images in video surveillance, the demand for high-definition images has become more urgent, especially in security applications where clear facial and license plate details are crucial. In the field of remote sensing imaging [2], acquiring high-resolution images is essential for geographic information systems, agricultural monitoring, and environmental protection. However, the high cost of satellite imaging and environmental interference make super-resolution technology a powerful alternative solution [3] [4]. Similarly, medical imaging [5] faces comparable challenges, as high-quality medical images play a decisive role in disease diagnosis and treatment. Super-resolution technology can enhance image quality without additional equipment costs.

Therefore, developing efficient and low-cost algorithms to improve low-resolution image quality has become a pressing challenge in the field of image processing.

Current super-resolution techniques can be primarily categorized into traditional methods and modern deep learning-based approaches. Traditional methods such as interpolation-based and reconstruction-based approaches can enhance image clarity to some extent, but they generally suffer from insufficient high-frequency detail recovery and poor adaptability to complex scenarios. With the rise of deep learning, convolutional neural network (CNN)-based SR methods have gradually become mainstream. However, these approaches typically rely on fixed degradation models, making them less effective in handling complex degradation scenarios such as motion blur, rotational distortion, and JPEG compression artifacts [6]. To further improve the adaptability of super-resolution technology, Generative Adversarial Networks (GANs) [7] have been introduced. Methods like SRGAN [8] and ESRGAN [9] significantly enhance texture details and visual quality through adversarial training. Nevertheless, challenges such as fixed degradation assumptions and training instability persist. Recently, Transformer-based architectures [10], such as Vision Transformer (ViT)[11], have demonstrated powerful global feature modeling capabilities. However, they still face challenges in computational efficiency and practical deployment.

To address these challenges, this paper introduces the ViT-Base module into the Real-ESRGAN [12] generator, forming a hybrid "shallow CNN - intermediate ViT - deep CNN" architecture. By leveraging self-attention mechanisms, the proposed model enhances global contextual dependency modeling and local feature extraction, thereby improving the balance between high-frequency detail recovery and global feature representation. This leads to superior visual perceptual quality and textural detail preservation. Through comparative experiments and visualization analysis, the proposed method demonstrates remarkable superiority in complex degradation scenarios. Experimental results validate its consistent performance improvement across multiple benchmark datasets, as well as robust generalization capability under diverse degradation conditions.

## II. Improved Real-ESRGAN-Based Image Super-Resolution Method

### *A. Residual-in-Residual Dense Block (RRDB)*

This study adopts Real-ESRGAN as the baseline architecture, with technical improvements primarily focused on network structure optimization. Compared to traditional generative adversarial architectures, the model introduces Residual-in-Residual Dense Blocks (RRDB) as a structural innovation. These blocks integrate multi-level residual connections and dense feature reuse mechanisms, significantly enhancing the network's feature extraction capability and gradient propagation efficiency. Specifically, the RRDB employs a cascaded residual learning framework, which effectively extracts multi-scale image features while optimizing information flow efficiency. Moreover, the network eliminates Batch Normalization (BN) layers, further improving its overall performance.

The residual block demonstrates the effect of removing the Batch Normalization (BN) layer. As a technique proposed in 2015 to accelerate the convergence of neural networks, BN can effectively address the vanishing gradient problem, speed up training, and prevent overfitting. However, in generative adversarial networks for super-resolution reconstruction, the BN layer performs mean and variance normalization on all data. With the increase in network depth, it may introduce artifacts into the generated images, especially when generating high-resolution images. To avoid the instability caused by the BN layer, Real-ESRGAN chooses to remove it, thereby improving the quality of image reconstruction, reducing computational complexity, and effectively enhancing computational efficiency when measured by Peak Signal-to-Noise Ratio (PSNR).

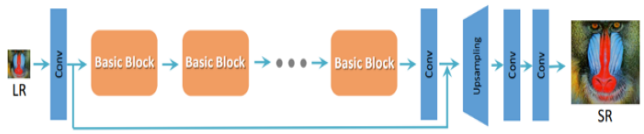


Figure 1. Replace the original Resblock with the RRDB structure

As shown in Figure 1, Real-ESRGAN employs the Residual-in-Residual Dense Block (RRDB) as the core building unit (the detailed structure of the module is shown in Figure 2), replacing the basic Residual Block (Resblock) in traditional generative networks. With the increase in the number of network layers, the effective receptive field of the model is gradually expanded through the cumulative effect of successive convolutional operations. This characteristic enables it to integrate a broader range of contextual correlation information, thereby enhancing the geometric consistency of super-resolution reconstruction.

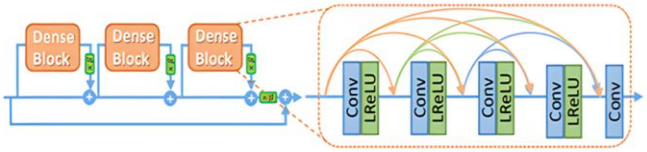


Figure 2. RRDB structure

In the adversarial training framework, this study adopts an improved relativistic discriminator architecture, whose objective function is defined as follows:

The mathematical expression of the discriminator loss is shown in equation (1):

$$L_D^{Ra} = -IE_{xr}\left[\log\left(D_{Ra}\left(x_r, x_f\right)\right)\right] - IE_{xf}\left[\log\left(1 - D_{Ra}\left(x_f, x_r\right)\right)\right] \quad (1)$$

The mathematical expression of the generator's adversarial loss is shown in equation (2):

$$L_D^{Ra} = -IE_{xr}\left[\log\left(D_{Ra}\left(x_r, x_f\right)\right)\right] - IE_{xf}\left[\log\left(1 - D_{Ra}\left(x_f, x_r\right)\right)\right] \quad (2)$$

Here, $x_f = G(xi)$ represents the generated high-resolution image, $x_r$ represents the real high-resolution image, and $xi$ is the low-resolution input image. This optimization approach has significant advantages in terms of edge details and image quality of the generated images.

In addition, regarding the perceptual loss calculation strategy, Real-ESRGAN innovatively sets the feature extraction point before the activation layer of the VGG network, with the comparison effect shown in Figure 3. Compared to SRGAN, which calculates the loss after the activation layer, this improvement effectively avoids the feature sparsification problem caused by the ReLU function.

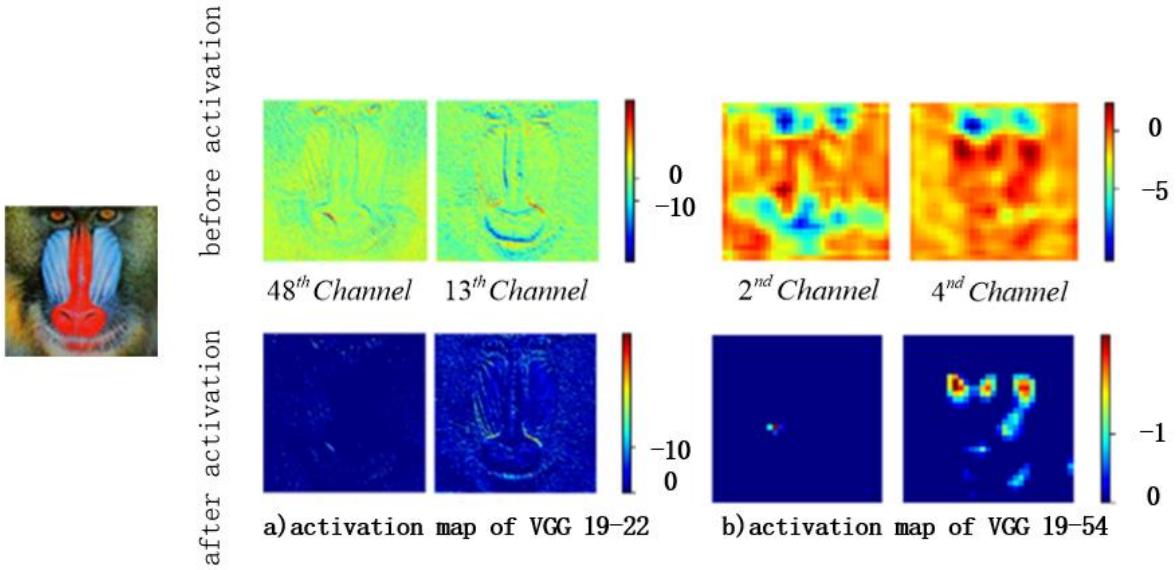


Figure 3. Representative feature maps before and after activation

### *B. The self-attention mechanism of Vision Transformer*

Despite the excellent performance of Real-ESRGAN based on Convolutional Neural Networks (CNN) in local feature extraction and detail recovery, its Residual-in-Residual Dense Block (RRDB) module is limited by the local receptive field characteristic of convolutions. Although stacking convolutional layers can gradually expand the receptive field, CNNs still struggle to effectively model global context information in complex scenes, such as high-frequency texture-dense urban architecture or long-range dependent remote sensing images.

As shown in Figure 4, the Vision Transformer (ViT) overcomes the locality limitations of CNNs through its self-attention mechanism, enabling the direct modeling of long-range dependencies between arbitrary pixels. Specifically, ViT divides the input image into non-overlapping patches and utilizes Multi-Head Self-Attention (MSA) to dynamically compute the correlation weights between different regions, thereby capturing global context. Moreover, the positional encoding design of ViT preserves the spatial information of the image, avoiding the issue of spatial information loss caused by unordered inputs in traditional Transformers.

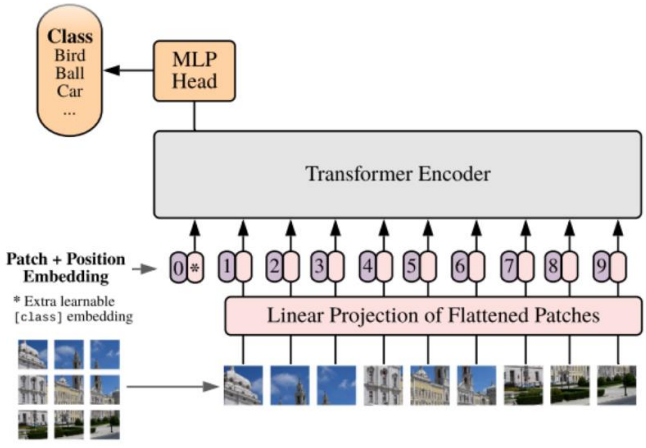


Figure 4. The architecture diagram of ViT

As shown in Figure 5, this paper intuitively compares the essential differences in the feature modeling mechanisms of ViT and CNN. The red box indicates the region of interest for the current operation, and the darker the heatmap color, the higher the attention weight. CNNs employ convolutional operations with local receptive fields, gradually expanding the receptive field through hierarchical stacking. In contrast, ViT directly establishes global dependencies through its self-attention mechanism. Experiments have shown that ViT can dynamically capture the correlations between distant pixels (e.g., the high-frequency structure formed by the four corner points of a building's window frame). Compared to the local processing approach of CNNs, ViT demonstrates a significant advantage in detail recovery.

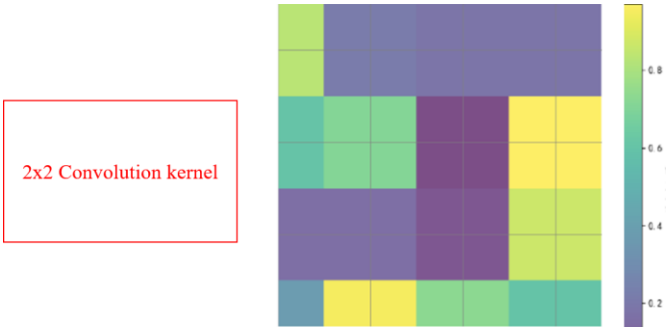


Figure 5. Comparison of CNN's local receptive field (left) and ViT's global attention weight distribution (right).

### *C. The Overall Structure of the Improved Network*

The generator network of Real-ESRGAN consists of 23 Residual-in-Residual Dense Blocks (RRDBs), which are primarily used to gradually extract low-resolution image features and enhance image details. The RRDB module captures local features and a certain range of global information through multi-level residual and dense connections. However, when dealing with high-frequency information, such as small targets or long-range dependencies, it may fail to adequately model complex global features, resulting in insufficient enhancement of local features.

To address this issue, this paper proposes to enhance the global feature modeling capability by introducing the Vision Transformer (ViT) module. ViT has strong global modeling capabilities, especially in capturing long-range pixel dependencies and overall image structure. By incorporating the ViT module, the shortcomings of the RRDB module in modeling global features can be effectively compensated for, particularly in improving the global consistency of generated

images and the detail restoration in complex scenes.

To ensure a balance between local and global feature modeling while controlling computational overhead, this paper replaces the convolutional and RRDB modules in layers 9 to 11 of the Real-ESRGAN generator with the ViT-Base module. This modification further enhances the network's ability to model global features. The improvement not only boosts the overall performance of the network but also shows significant advantages, especially in handling small-object super-resolution tasks. The improved network architecture is shown in Figure 6.

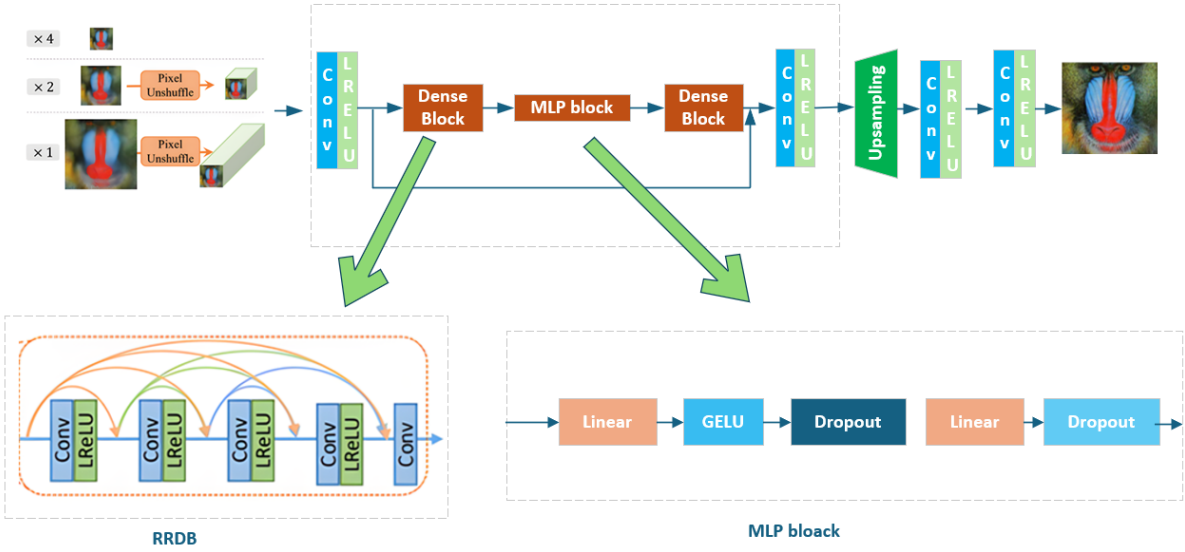


Figure 6. The improved generator network structure

In the process of introducing the ViT module, this paper selected the ViT-Base model as the foundational module to replace the RRDB module. The multi-head self-attention mechanism of ViT can effectively capture global features and improve the global consistency of images through long-range dependency modeling. However, ViT has a large number of parameters, and due to its global attention mechanism, it may introduce significant computational overhead when used within the GAN generator framework. To avoid these potential issues, this method only replaces layers 9 to 11 in the generator, retaining the convolutional and RRDB modules in the shallow and deep layers to ensure a balance between local detail recovery and global feature modeling capabilities.

As shown in Table 1, the table presents the specific parameter settings of the ViT-Base module. These parameters are based on the basic structure of ViT and have been adjusted to meet the requirements of the generator network.

TABLE I. PARAMETER CONFIGURATION OF THE VIT-BASE MODULE

| ***Module Component*** | ***Key Parameters*** | ***Main Function*** |
|---|---|---|
| Patch Embedding | Input Size：64×64×64<br>Patch Size：8×8<br>Output Dimension：768 | Divide the feature map into 64 patches to reduce computational complexity. |
| Multi-Head Attention | Number of Heads：12<br>Dimension per Head：64 | Establish global correlations to enhance complex feature modeling. |

TABLE I(Continued Table). PARAMETER CONFIGURATION OF THE VIT-BASE MODULE

| *Module Component* | *Key Parameters* | *Main Function* |
|---|---|---|
| Feed-Forward Network | MLP Structure：768→3072→768<br>Activation Function：GeLU | Perform nonlinear transformations of features to stabilize training in conjunction with LayerNorm. |
| Residual Connection | Application Position：After each MSA/FFN sublayer | Prevent gradient vanishing and accelerate convergence. |
| Stacked Structure | Number of Encoder Layers：12<br>Total Parameters：86M | Construct a deep feature transformer to improve global context modeling capabilities. |

## III. Experimental Design and Result Analysis

### *A. Experimental Environment and Configuratio*

The experimental environment of this paper is based on the Ubuntu 18.04 operating system, with an RTX 2080Ti GPU and 16 GB of memory. The experiments were essentially conducted using the officially recommended parameter settings and were implemented using the Python 3 and PyTorch frameworks. The training set consists of the DIV2K dataset (800 images) and the Flickr2K dataset (2650 images), totaling 3450 high-resolution images covering a wide range of scenes, including natural landscapes, urban areas, and architecture. The test sets include the Set5, Set14, BSD100, and Urban100 benchmark datasets to evaluate the model's performance across different scenarios.

### *B. Experimental design*

In this experiment, we used the DIV2K dataset (800 images) and the Flickr2K dataset (2650 images) as the training datasets, totaling 3450 high-resolution images, covering a variety of natural and man-made scenes. To simulate the local degradation features in real-world scenarios, the images were first randomly cropped into 256×256 pixel blocks, and then various data augmentation operations were performed to enhance the model's robustness to complex degradation conditions. The augmentation operations included: applying horizontal or vertical flipping with a 50% probability and randomly rotating the images within the range of −90° to 90° to improve the model's adaptability to directional changes; for color perturbation, the brightness was uniformly sampled and adjusted by ±15%, and the contrast and saturation were uniformly sampled and adjusted by ±10% to simulate different lighting and imaging environments; for noise injection, Gaussian noise with a standard deviation ranging from 0 to 0.03 was added, and JPEG compression noise with a quality factor ranging from 20 to 95 was introduced to simulate the degradation of images during acquisition, transmission, or compression processes.

The training process consists of two stages: pre-training and formal training, with the overall procedure designed to ensure the stability of training and the progressive improvement of model performance. During the pre-training stage, the generator is initialized and optimized using only the L1 loss function for 20k iterations, with a fixed learning rate of $2\times10^{-4}$. Performance validation is conducted every 17,280 iterations. This stage effectively mitigates the mode collapse issue caused by an overly powerful discriminator in the early stages of adversarial training, with the training loss stabilizing around 0.03 after 20k iterations.

Subsequently, the formal training stage commences, lasting for 500k iterations. The

generator and discriminator are jointly optimized using a combination of L1 loss, perceptual loss, and adversarial loss. The learning rate is dynamically adjusted to further enhance training stability. During the training process, the loss curve is recorded every 216 iterations, and the PSNR and SSIM metrics are output every 1,728 iterations to evaluate the model's performance in terms of reconstruction quality. As shown in Figure 7, the training loss drops below 0.025 after the 300kth iteration, indicating that the model has essentially converged, with stable and reliable training results.

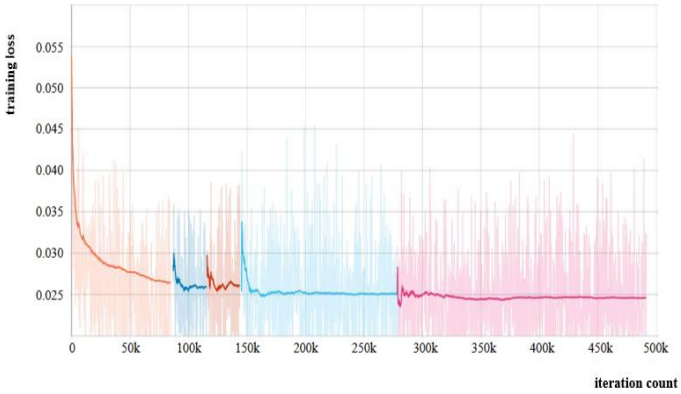


Figure 7. The improved generator network structure

### *C. Comparative Experiments and Analysis*

To verify the performance of the improved model, comparative experiments were conducted with several mainstream super-resolution methods, including VDSR, EDSR, RCAN, and SwinIR. Quantitative metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) were used for evaluation on the Set5, Set14, BSD100, and Urban100 datasets, supplemented by visual comparisons. The results are shown in Table 2.

TABLE II. COMPARISON OF PSNR (DB) / SSIM FOR DIFFERENT METHODS ON BENCHMARK DATASETS UNDER ×2 AND ×4 SUPER-RESOLUTION

| ***Multiple*** | ***Model*** | ***Set5*** ***PSNR (dB) /SSIM*** | ***Set14*** ***PSNR (dB) /SSIM*** | ***BSD100*** ***PSNR (dB) /SSIM*** | ***Urban100*** ***PSNR (dB) /SSIM*** |
|---|---|---|---|---|---|
| ×2 | SRCNN | 36.66/0.9542 | 32.42/0.9063 | 31.36/0.8918 | 29.50/0.8946 |
| | EDSR | 38.11/0.9603 | 33.92/0.9180 | 32.46/0.9015 | 32.93/0.9355 |
| | RCAN | 38.31/0.9614 | 34.15/0.9209 | 32.63/0.9027 | 33.34/0.9410 |
| | HAN | 38.34/0.9618 | 34.18/0.9214 | 32.68/0.9032 | 33.41/0.9422 |
| | NLSA | 38.35 / 0.9620 | 34.20 / 0.9217 | 32.69 / 0.9036 | 33.48 / 0.9430 |
| | SwinIR | 38.40 / 0.9624 | 34.25 / 0.9221 | 32.71 / 0.9040 | 33.57 / 0.9438 |
| | Our | 38.51 / 0.963 | 34.40 / 0.9235 | 32.85 / 0.9054 | 33.95 / 0.9465 |
| ×4 | SRCNN | 30.49 / 0.8630 | 27.50 / 0.7510 | 26.91 / 0.7105 | 24.54 / 0.7260 |
| | EDSR | 32.65 / 0.9000 | 28.95 / 0.7918 | 27.80 / 0.7449 | 26.98 / 0.8085 |
| | RCAN | 32.78 / 0.9004 | 29.06 / 0.7932 | 27.89 / 0.7468 | 27.14 / 0.8125 |
| | HAN | 32.80 / 0.9008 | 29.08 / 0.7939 | 27.91 / 0.7473 | 27.24 / 0.8140 |
| | NLSA | 32.82 / 0.9010 | 29.10 / 0.7942 | 27.93 / 0.7480 | 27.31 / 0.8153 |
| | SwinIR | 32.88 / 0.9014 | 29.14 / 0.7948 | 27.96 / 0.7488 | 27.45 / 0.8165 |
| | Our | 33.00 / 0.9025 | 29.28 / 0.7962 | 28.10 / 0.7505 | 27.73 / 0.8200 |

As shown in Table 2, the proposed method in this paper demonstrates excellent performance in the image super-resolution task, especially under multi-scale magnification and complex scenarios (such as Urban100). Both quantitative and qualitative experiments indicate that our method outperforms existing approaches in terms of PSNR and SSIM, including traditional networks like SRCNN, EDSR, and RCAN, as well as advanced models like SwinIR.

Specifically, at the ×2 magnification scale, our method achieves a PSNR of 33.95 dB on the Urban100 dataset (an improvement of 0.38 dB over SwinIR) and an SSIM of 0.9465 (an increase of 0.027). Our method also performs outstandingly on benchmark datasets such as Set5 and Set14, demonstrating its generalizability. For the more challenging ×4 task, our method achieves a PSNR of 33.00 dB on Set5 (an improvement of 0.12 dB) and a PSNR of 27.73 dB with an SSIM of 0.8200

on Urban100, showing significant performance advantages.

*D. Visualization Analysis*

As shown in Figure 8, the proposed method in this paper demonstrates significant advantages in texture recovery in complex scenes, particularly in terms of clarity and naturalness in high-frequency areas such as building edges and texture patterns. Comparative experiments indicate that the reconstruction results of SRCNN are blurry and contain artifacts; although EDSR shows some improvement, the details are still not clear; RCAN and HAN enhance clarity but struggle to accurately recover complex textures; NLSA suffers from structural blurring and color distortion; and SwinIR exhibits texture reconstruction errors in images such as Barbara and img062.

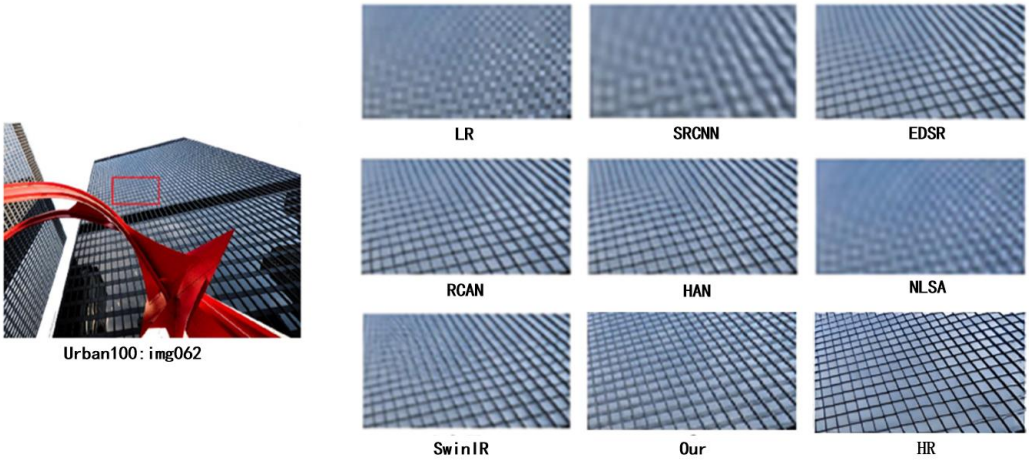


Figure 8. Comparison of 4x upsampling experimental results for img062.png in the Urban100 dataset.

In contrast, the proposed method in this paper demonstrates more accurate and clear texture edge recovery in the tests of Barbara and img062. In terms of visual perception, the SR results of Barbara are very close to the HR image, especially in the detail restoration of book textures. In the img062 image, the shape and structure of the window grid are closest to the HR image, fully reflecting the superiority of the proposed method in handling high-frequency information in complex scenes.

## IV. CONCLUSIONS

This paper addresses the issues of low reconstruction accuracy and insufficient model generalization ability in image super-resolution reconstruction under complex degradation scenarios by proposing an improved method that combines Generative Adversarial Networks (GANs) and Vision Transformers (ViT). By replacing some residual modules in the Real-ESRGAN generator with ViT modules, the method leverages the self-attention mechanism to enhance global feature modeling while retaining the advantages of local detail extraction, significantly improving the reconstruction performance in complex scenes. Experimental results demonstrate that the proposed method achieves a PSNR improvement of 0.59 dB/0.45 dB (for ×2/×4 magnification) and an SSIM improvement of 0.018/0.056 on the Urban100 dataset, verifying the effectiveness of the proposed approach. This research not only proves the superiority of the hybrid architecture in image super-resolution tasks but also provides a new technical idea for balancing global modeling and local detail recovery through a hierarchical feature fusion strategy. In practical applications, the proposed method shows great potential in fields such as security surveillance and remote sensing imagery, effectively enhancing the quality and usability of low-resolution images. Future research can further explore model lightweight design, cross-modal super-resolution reconstruction, and adaptive degradation modeling to improve the model's practicality and generalization ability and

to promote the application and development of super-resolution technology in more real-world scenarios. It can be concluded that combining the strengths of CNNs and Transformers is an effective way to enhance the performance of image super-resolution.

## REFERENCES


[1] Zhang Fang, Zhao Dongxu, Xiao Zhitao, et al. Research Progress on Single Image Super-Resolution Reconstruction Technology. Acta Automatica Sinica, 2022, 48(11): 2634-2654.

[2] Liu Chenglu. Research on Image Super-Resolution Reconstruction Algorithm Based on Improved Generative Adversarial Network. Yanshan University, 2022.

[3] Wang Yuanhang. Research on Super-Resolution Reconstruction of Remote Sensing Images for Small Target Detection Based on Deep Learning. Harbin Institute of Technology, 2022.

[4] Xu Zhengpu. Research on Super-Resolution Reconstruction Algorithm for Single Space Target Image Based on Deep Learning. Xidian University, 2021.

[5] Lin Yi. Research on Medical Image Super-Resolution Algorithm Based on Attention Mechanism. Anhui University, 2023.

[6] WANG X, XIE L, DONG C, et al. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 1905-1914.

[7] Liang Junjie, Wei Jianjing, Jiang Zhengfeng. A Survey on Generative Adversarial Networks (GAN). Journal of Computer Science and Exploration, 2020, 14(01): 1-17. Zhang Yi, Shen Kexin, Liu Kaili, et al. Research on Surveillance Image Enhancement Based on SRGAN Model.Computer Programming Skills and Maintenance, 2024, (07): 144-146 + 173.

[8] Ni Jie, Liu Qingyuan, Zhou Li. Research on Historical Image Super-Resolution Reconstruction Using Improved Real-ESRGAN Model. Journal of Information and Management, 2025, 10(1): 65-77

[9] Niu Kai, Jia Ronghao, Wei Guohui, et al. Pneumonia Auxiliary Diagnosis Based on Hybrid CNN and Transformer Model. Computer Systems Application, 2025, 34(02): 216-222.

[10] Liu Wenting, Lu Xinming. Research Progress on Transformer in Computer Vision. Computer Engineering and Applications, 2022, 58(06): 1-16.

[11] Zhang Yuxin, Zeng Xue, Wang Yongzeng, et al. Application of Improved Real-ESRGAN Model in Low-Resolution Image Enhancement of Airport Baggage. Logistics Technology and Applications, 2024, 29(12): 180-184.

[12] Zhang Chao. Research on Real-World Image Super-Resolution Reconstruction Based on Deep Learning. Southeast University, 2023.

[13] Ma Huanhuan. Blind Image Quality Assessment Based on Joint Attention Mechanism and Multi-Level Features. Nanjing University of Posts and Telecommunications, 2022.

[14] Zhang Chao. Research on Real-World Image Super-Resolution Reconstruction Based on Deep Learning. Southeast University, 2023.