Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 2, 2025

Implicit Function Super-Resolution Reconstruction Based on Group Propagation Vision Transformer

Jiacun Song[1, *]

[1]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
[*]Corresponding Author: 212209020057@home.hpu.edu.cn

Abstract: Reconstruction-based single-image super-resolution methods, while demonstrating excellent performance, often face challenges such as training instability, artifact generation, information loss, and insufficient control over global information. To address these challenges, we propose an implicit function super-resolution reconstruction algorithm based on Group Propagation Vision Transformer (GP-ViT). This method employs GP-ViT as an encoder to efficiently capture global contextual information through a group propagation mechanism, while reducing computational complexity and memory consumption, and significantly enhancing local feature extraction capabilities. In the decoding phase, the algorithm utilizes an implicit function continuous representation to decode image features, supporting super-resolution reconstruction up to 32 times, enabling the recovery of high-frequency details in a continuous manner and generating high-quality images. Experimental results show that compared to classical super-resolution models, our method has significant improvements in two key metrics, PSNR and SSIM, while effectively reducing artifacts and preserving more detailed information.

Keywords: Implicit function; Super-resolution; GP-ViT.

1. Introduction

In recent years, super-resolution technology based on deep learning has made significant progress. This technology focuses on enhancing the resolution of images, that is, restoring high-resolution images with high clarity and rich details from blurry and smaller-sized images, thereby increasing the display size of the images. As a long-standing research topic in the field of computer vision, super-resolution technology based on deep learning primarily relies on deep learning models to capture and learn the complex and intricate mapping relationships between low-resolution and high-resolution images. Through this process, the technology can achieve high-quality image super-resolution reconstruction, significantly improving the visual effects and detail representation of images, thereby optimizing and enriching the content and quality of image presentation.

In 2014, Dong et al. pioneered the introduction of SRCNN [1], marking the first application of deep learning technology in the field of image super-resolution. By constructing a three-convolutional-layer architecture, SRCNN effectively learned and established a mapping model between low-resolution and high-resolution images, achieving remarkable super-resolution results and laying the foundation for subsequent research in deep learning for super-resolution. With the rise of deep convolutional neural network models such as AlexNet, researchers began to integrate deeper network architectures into super-resolution tasks. Kim et al. subsequently proposed VDSR [2], significantly improving reconstruction performance by increasing the number of network layers, highlighting the potential of deep networks in the field of super-resolution. Shortly after, the introduction of the residual network (ResNet) concept opened new paths for super-resolution research. Inspired by this, ResNet-based super-resolution models, such as SRResNet, adopted residual structures to alleviate the vanishing gradient problem in deep network training. At the same time, other researchers explored super-resolution methods based on densely connected networks (DenseNet), such as SRDenseNet, which further enhanced image reconstruction quality by deeply mining and utilizing multi-level feature information. In the same year, attention mechanisms achieved breakthrough progress in the field of image super-resolution. Specifically, Zhang et al. proposed RCAN [3], a typical example that significantly improved the model's performance in restoring high-frequency details of images by integrating channel attention mechanisms. During this period, a large number of research papers emerged, continuously exploring and applying new variants and optimization strategies of various attention mechanisms, further advancing the development of image super-resolution technology. With the introduction of ViT, Transformers began to be applied in the field of computer vision, also bringing new breakthroughs to super-resolution technology. For example, Chen et al. proposed IPT [4], which adopted the original Transformer structure to perform low-level vision tasks such as image super-resolution through pre-trained models. The proposal of IPT marked the formal application of Transformers in the field of super-resolution.

GAN networks, due to their exceptional generative capabilities, have opened new avenues in the field of super-resolution. In 2016, Christian Ledig et al. proposed SRGAN [5], marking the first application of GAN technology in the field of super-resolution. The core of SRGAN lies in leveraging the adversarial training mechanism of GANs, enabling the generator to produce highly realistic and detail-rich high-resolution images. It learns the mapping relationship from low-resolution to high-resolution through a convolutional neural network, while the discriminator is responsible for distinguishing between generated images and real images, prompting the generator to continuously optimize. The introduction of SRGAN brought a completely new perspective to super-resolution technology. Subsequently, based on SRGAN, researchers have made numerous improvements and innovations, with ESRGAN [6] being particularly notable. In 2018, Xintao Wang et al. proposed ESRGAN, which optimized the network structure and loss

function while retaining the overall framework of SRGAN. ESRGAN replaced Residual Blocks with Dense Blocks and removed batch normalization layers (BN), thereby enhancing the quality and detail representation of the generated images. Additionally, ESRGAN introduced a combination of perceptual loss and adversarial loss to more effectively guide the training of the generator. Besides SRGAN and ESRGAN, other GAN-based super-resolution methods have also made significant progress. For example, BSRGAN [7] proposed an image degradation random permutation strategy based on ESRGAN, aiming to simulate various degradation scenarios of images in the real world, thereby effectively addressing the problem of blind image super-resolution. These improvements and innovations collectively drive the continuous development of super-resolution technology.

Previous super-resolution technologies mostly focused on non-blind super-resolution, which significantly differs from the low-quality images commonly encountered in real life. To better align with practical applications, blind super-resolution technology emerged, enabling super-resolution processing of images under unknown degradation kernels, thereby maximally simulating real-world low-quality images. In the early stages, blind super-resolution primarily relied on traditional image processing techniques and manually designed algorithms. However, with the rapid development of deep learning technologies, the learning methods for blind super-resolution gradually shifted towards data-driven and neural network-based approaches. Among these, SRMD [7], proposed by the Computer Vision Laboratory at ETH Zurich, stands out as the first blind image super-resolution method employing deep learning. SRMD combines the input image with degradation information (such as blur kernels, noise, etc.) and inputs them into the super-resolution model, allowing the model to adapt features according to specific degradation scenarios, thus covering multiple degradation types within a single model. Subsequently, explicit modeling and implicit modeling became the two mainstream directions in blind super-resolution learning methods. Explicit modeling methods typically adopt classical degradation models and attempt to estimate degradation parameters (such as blur kernels and noise levels) through various means. For example, the IKC [8] method, proposed by Jinjin Gu and Chao Dong's team, employs an iterative optimization approach to correct the estimation of blur kernels.

In our research paper, we have pioneered the proposal of an implicit neural representation framework based on Group Propagation Vision Transformer [9] (GPViT) encoding, serving as the core model for image processing tasks. This framework deeply integrates the advantages of Group Propagation Block (GP Block) encoding with the unique properties of implicit neural networks, bringing significant performance leaps to related tasks. Specifically, we ingeniously utilize GP Blocks to achieve efficient exchange of global information. Compared to traditional Transformer models, GPViT demonstrates superior performance in visual recognition tasks due to its exceptional global information processing capabilities and modular design. Furthermore, we introduce implicit neural networks as the basic architecture, greatly enriching the model's expressive power. Implicit neural networks map discrete image data to continuous functions, endowing the model with the ability to perform upsampling at arbitrary resolutions, i.e., generating high-quality images at different scales on demand. This continuous representation not only enhances the quality of image generation but also opens up new flexibility and possibilities for subsequent image processing workflows.

2. Related Work

The concept of Implicit Neural Representation (INR), as a cutting-edge data representation method, focuses on utilizing continuous functions to accurately capture and express the intrinsic features and states of complex data such as images and videos. Although the precise mathematical forms of these continuous functions are often difficult to determine directly, researchers, with their profound mathematical expertise and innovative thinking, ingeniously employ neural networks as a powerful tool to approximate these functions. In the field of image super-resolution, this concept has found preliminary yet highly promising applications, opening new avenues for enhancing image resolution. Specifically, researchers have innovatively applied INR functions that map two-dimensional coordinates to RGB values in image super-resolution tasks. By training neural networks to learn the mapping relationship from low-resolution to high-resolution images, significant improvements in image resolution are achieved. This method not only preserves the detailed information in the images but also makes the generated images more natural and realistic. In 2021, the introduction of the LIIF [10] method marked a significant step forward in super-resolution technology based on implicit neural representation. The LIIF method learns a neural implicit function that takes the feature mapping of low-resolution images and the coordinates of high-resolution images as joint inputs, and through the nonlinear transformation capabilities of neural networks, it precisely predicts the RGB values of high-resolution images. This method not only enhances the accuracy and efficiency of image super-resolution but also endows the model with the ability to query arbitrary continuous high-resolution coordinates during the testing phase. This means that users can generate images of any resolution as needed, greatly meeting the demands of various application scenarios.

In recent years, the Group Propagation Vision Transformer (GPViT) model has achieved remarkable success in the field of computer vision, attracting widespread attention from researchers. However, traditional Transformers face a significant challenge when processing high-resolution images: the computational complexity of their self-attention mechanism is proportional to the square of the image size, leading to substantial computational and memory demands, which severely limit the model's scalability and efficiency in practical applications. Although traditional Transformers have demonstrated exceptional capabilities in capturing global contextual information, enabling them to understand and utilize long-range dependencies within images, they may fall short in handling local feature extraction. The precise capture of local features is crucial for tasks such as image understanding, segmentation, and restoration, yet the performance of traditional Transformers in this aspect still requires improvement.

To address this challenge, the Group Propagation Vision Transformer (GPViT) ingeniously optimizes the Transformer model by introducing the innovative Group Propagation Block (GP Block). GPViT not only reduces computational complexity and enhances memory efficiency but also significantly improves the model's ability to extract local features. Specifically, the GP Block utilizes a fixed number of learnable group tokens to partition image features and

employs an efficient group propagation mechanism to exchange global information among the grouped features. This design enables GPViT to effectively capture global contextual information while maintaining high resolution, without compromising the fine-grained capture of local features.

3. Method

In our research paper, we innovatively integrated a framework based on implicit neural networks with the Group Propagation Vision Transformer model to construct an efficient and powerful image processing system. This combined strategy not only deeply reflects the respective advantages of the two models but also significantly enhances the performance and efficiency of image processing through their synergistic effects. Specifically, we selected the Group Propagation Vision Transformer as the core component for image encoding. Leveraging its efficient global information exchange capabilities, the Group Propagation Vision Transformer can deeply understand the complex structures and detailed features within images. This characteristic enables the model to more accurately identify and retain key information during image processing, thereby significantly improving the quality of image generation. On the other hand, we adopted a framework based on implicit neural networks as the fundamental support for the entire system. Implicit neural networks, with their unique representational capabilities and generalization performance, have demonstrated remarkable potential in handling high-magnification image information. This framework can efficiently process complex and variable image data, providing more flexible and powerful support for image encoding and generation. By incorporating implicit neural networks, our system can better adapt to images of different resolutions and contents, further broadening the application scenarios and scope. The overall framework is illustrated in Figure 1(a):

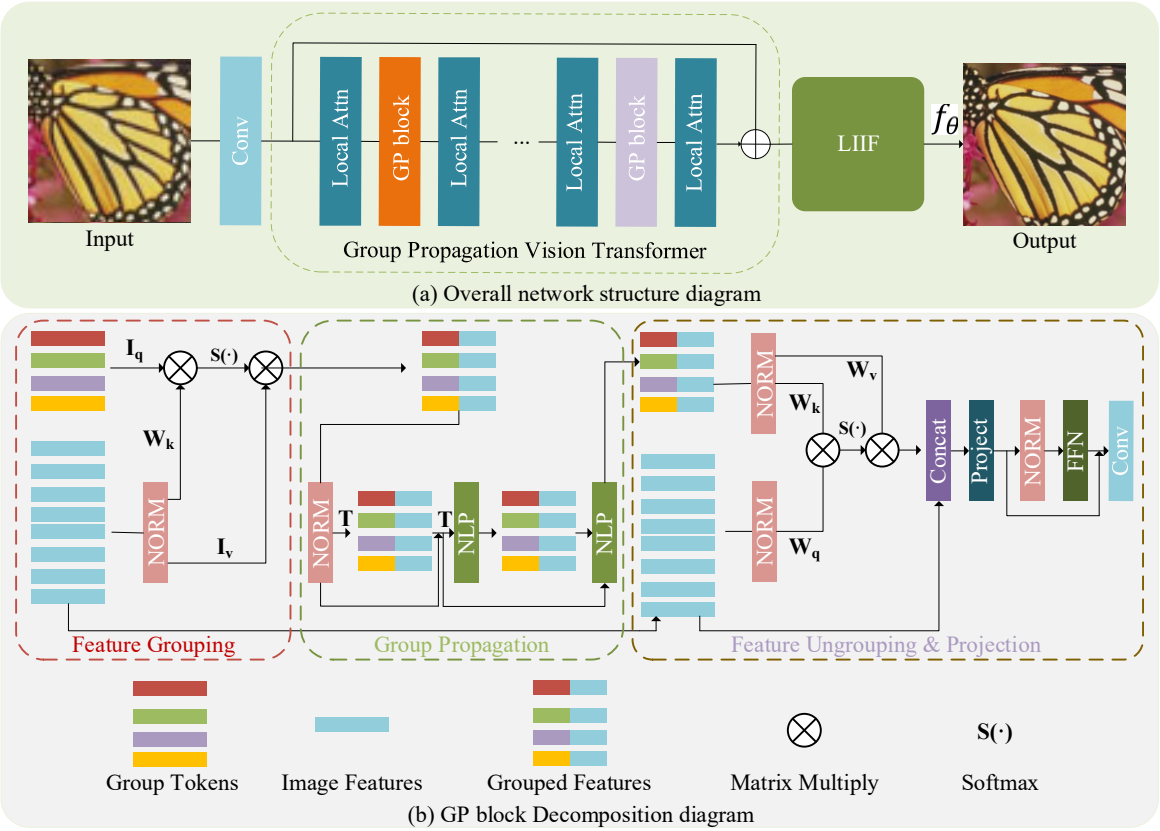

Figure 1. The overall network model

3.1. Implicit Neural Representations.

Implicit Neural Networks (INNs) are a special type of neural network architecture whose core lies in representing and solving problems through implicit functions. Unlike traditional explicit neural networks (such as fully connected networks, convolutional neural networks, etc.) that directly output prediction results, implicit neural networks are typically used to solve implicit equations of the form F(x) = 0, where F is a complex function represented by a neural network. The core innovation of LIIF (Local Implicit Image Function) lies in its unique local implicit function representation method. The central idea of this method is to map each local region of an image to a continuous, high-dimensional implicit space. Within this implicit space, the local features of the image are ingeniously encoded as parameters of the implicit function. By deeply learning these parameters, LIIF can accurately capture the local structures and texture details of the image, thereby achieving precise prediction of pixel values at any location in the image.

In the LIIF framework, each image $I^{(i)}$ is first converted into a 2D feature map $M^{(i)} \in \mathbb{R}^{H\times W\times D}$, which meticulously records the local feature information for each pixel in the image. Subsequently, LIIF utilizes a decoding function (typically a meticulously designed Multi-Layer Perceptron, MLP) f_θ (with θ as the parameters) to parse these feature maps and convert them into pixel values at arbitrary positions

in the image. This decoding process is deeply rooted in local implicit functions, and it is precisely due to this design that LIIF is able to make precise predictions for arbitrary positions in the image. The decoding function can be expressed as

$$s = f_\theta(z,\ x). \tag{1}$$

Where, the vector z represents the hidden feature encoding output from the encoder, which can also be understood as a latent feature representation. x represents the two-dimensional coordinate points in the image, used to locate specific positions within the image. s denotes the predicted signal value at the given position x, typically corresponding to the RGB color values of the image. In practical applications, we set the two dimensions of the two-dimensional coordinate x to vary within the ranges $[0,\ 2H]$ and $[0,\ 2W]$, where H and W represent the height and width of the image, respectively. Multiplying by 2 accounts for potential upsampling or super-resolution factors, allowing the coordinate system to cover higher-resolution image spaces. We refer to $M^{(i)}$ as the latent encoding map, which is a uniformly distributed feature representation in the continuous image domain. For each local region in the image $I^{(i)}$, $M^{(i)}$ provides a rich set of features. As shown in Figure 2, these latent encodings $M^{(i)}$ can be imagined as a series of feature points uniformly distributed in the continuous image space, with each feature point associated with a local region in the image. To predict the RGB value of the continuous image $I^{(i)}$ at any coordinate x_q from the latent encoding map $M^{(i)}$, we adopt the following definition: First, assign a corresponding two-dimensional coordinate to each feature point in the latent encoding $M^{(i)}$. Then, using a decoding function (typically a multi-layer perceptron (MLP) or another type of neural network), take the coordinate x_q and the nearest latent encoding (or locally interpolated features obtained through some method) as inputs, and output the predicted RGB value at that position. This process can be expressed by the following formula:

$$I^{(i)}(x_q) = f_\theta(z^*, x - v^*). \tag{2}$$

Where, z^* represents the latent code that is nearest to the coordinate x_q in the Euclidean sense (the nearest one to x_q in the figure is Z_{11}^*, and here z∗ corresponds to Z_{11}^*), and v∗ represents the coordinate of the latent code z^* in the image domain (in Figure 2, the coordinate of v^* is the coordinate of Z_{11}^*).

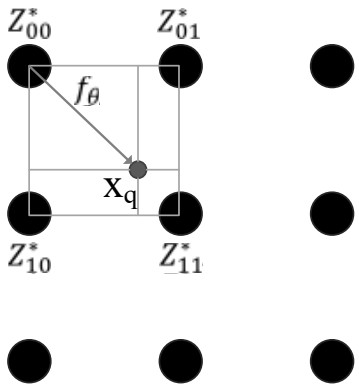

Figure 2. Two-dimensional feature map of LIIF

In addition, LIIF achieves continuous processing of image data by transforming discrete image pixel representations into continuous implicit function representations. This transformation endows LIIF with the ability to capture continuous variations in images, enabling it to perform precise interpolation and prediction at any position within the image. Overall, these characteristics of LIIF collectively form the basis for its powerful image processing and prediction capabilities.

3.2. Group Propagation Vision Transformer

In our model encoder, the core component is the Group Propagation Block (GP Block), whose structure is shown in Figure 1(b). The GP Block primarily consists of three key parts: Feature Grouping, Group Propagation, and Feature Ungrouping & Projection. In the first stage, image features are effectively grouped; then, in the second stage, global information is propagated among these grouped features; finally, in the last stage, the global information is fed back and integrated into the original image features.

3.2.1. Feature Grouping.

The input to the GP Block is an image feature matrix (represented by the light blue matrix block in Figure 1b) $X \in \mathbb{R}^{N\times C}$, where N denotes the total number of image features and C represents the dimension of each feature vector. In Feature Grouping,M learnable tokens are stored in the matrix $G \in \mathbb{R}^{M\times C}$, and then a multi-head attention mechanism is used for grouping, resulting in grouped features $Y \in \mathbb{R}^{M\times C}$. Feature Grouping can be expressed using Equation (4) as:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{p}}\right)V \tag{3}$$

$$Y = Concat_{\{h\}}\left(Attention\left(W_h^Q G_h, W_h^K X_h, W_h^V X_h\right)\right) \tag{4}$$

Where p is the channel index, h is the head index, and $W_h^{\{Q,K,V\}}$are the projection matrices for query, key, and value, respectively. Subsequently, we remove the feature projection layer, which is connected to the heads of the multi-head attention mechanism. The outputs of each head are then weighted and summed, with the weights determined based on the contribution of each head at each position.

3.2.2. Group Propagation.

After obtaining the grouped features $Y \in \mathbb{R}^{M\times C}$ in the previous step, global information propagation can be achieved in Group Propagation. Specifically, we adopt the structure of an MLPMixer, which consists of two consecutive multi-layer perceptrons (MLPs). The first MLP is used to mix information between groups, while the second MLP is used to mix information along the channel dimension.

After obtaining the grouped features $Y \in \mathbb{R}^{M\times C}$ in the previous step, global information propagation can be achieved through Group Propagation. The specific implementation is as follows: We adopt the MLPMixer structure, which is a lightweight and efficient neural network module capable of mixing information between groups and across channel dimensions while maintaining low computational complexity and parameter count. The MLP-Mixer consists of two consecutive multi-layer perceptrons

(MLPs) used for information fusion at different levels. The first MLP operates between groups, capturing global inter-group relationships by mixing information from different groups. The second MLP operates across channel dimensions within each group, enhancing feature expressiveness by mixing feature channels. This can be represented by the following equations:

$$\bar{Y} = Y + \left(MLP_1(LayerNorm(Y)^T)\right)^T \quad (5)$$

$$\tilde{Y} = \bar{Y} + \left(MLP_2(LayerNorm(\bar{Y})^T)\right)^T \quad (6)$$

Where Y is the input grouped feature matrix with dimensions$M \times C$, where M represents the number of groups and C represents the feature dimension of each group. LayerNorm(·) denotes the layer normalization operation, used to stabilize the training process. MLP_1 and MLP_2 are two multi-layer perceptrons used for information mixing between groups and across channel dimensions, respectively.$\bar{Y}$ is the output of the first MLP, serving as an intermediate result. $\tilde{Y}$ is the final output feature, containing global information both between and within groups. Through the MLP-Mixer, the grouped feature Y is updated to $\tilde{Y}$, which not only retains the information of the original features but also incorporates global contextual information.

3.2.3. Feature Ungrouping & Projection

The goal of Feature Ungrouping is to recombine the updated grouped features $\tilde{Y}$ after Group Propagation with the original image features X, thereby propagating global information back to the image features. This process is achieved through a Transformer Decoder Layer, where the grouped features $\tilde{Y}$ serve as the Key and Value, and the image features X serve as the Query. First, the Multi-Head Attention mechanism is used, with the image features X as the Query and the grouped features $\tilde{Y}$ as the Key and Value, to extract global information from the grouped features. The output U of the attention mechanism contains the global contextual information extracted from the grouped features. Subsequently, the attention output U is concatenated with the original image features X to form the fused features. A linear projection matrix is then used to map the concatenated features to the same dimension as the image features X, resulting in $\bar{Z}$. Next, a Feed-Forward Network (FFN) is employed to further transform the features $\bar{Z}$, enhancing their expressive power. Residual connections ($\bar{Z}$ + FFN($\bar{Z}$) are used to retain the original information, yielding $\tilde{Z}$. Finally, after a convolutional operation, the enhanced features Z are obtained as the output of the GP Block. In Z, both global contextual information and local detail information are preserved.

4. Experiment

4.1. Details of the experiment

In this paper, we conducted an in-depth analysis of the model's performance, particularly employing a scaling factor of X4 for image super-resolution. During the experiments, we chose the efficient and flexible Pytorch deep learning framework as the technical foundation, which not only ensured the convenience of algorithm implementation but also enhanced computational efficiency. To support this complex and large-scale experiment, we utilized a hardware environment equipped with four high-performance Tesla A100 GPUs, significantly accelerating the model training and validation processes. In terms of specific implementation details, we meticulously set the experimental parameters to maximize the model's learning effectiveness. The batch size was set to 96, a choice that ensures efficient memory utilization while promoting the stability and convergence speed of model training. Additionally, we adopted 48x48 pixel patches as the basic units of input data, a size chosen to balance computational resource constraints with the model's ability to capture image features. After careful tuning, we ultimately set the learning rate to 0.0001, a smaller rate that helps the model converge smoothly during the initial training phase, avoiding local optima. Furthermore, we set the total number of training iterations to 1000 epochs, ensuring that the model has sufficient time to learn from the data and achieve the desired generalization performance.

We choose the DIV2K dataset for model training, which contains 800 natural scenery images in the training set and 100 images in the validation set. Leveraging the rich image resources of DIV2K, the model can learn more features and improve its generalization ability. At the same time, using a validation set from the same source to evaluate the model ensures the accuracy and reliability of the evaluation results, comprehensively reflecting the model's performance.

4.2. Evaluate benchmarks and indicators

In this study, to comprehensively and thoroughly evaluate the performance of our model, we have carefully selected five representative benchmark datasets, namely Set5 [11], Urban100 [12], Manga109 [13], BSD100 [14], and DIV2K100 [15]. These datasets not only cover a variety of image types such as natural scenes, urban landscapes, and manga, but also exhibit a wide range of representativeness in terms of resolution, detail richness, and complexity, thereby enabling a comprehensive examination of our model's generalization ability and adaptability.

For evaluation metrics, we have adopted two widely recognized and commonly used image quality assessment indicators: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR primarily measures the fidelity of the reconstructed image by calculating the pixel differences between the original and reconstructed images to assess image quality. On the other hand, SSIM focuses on evaluating the structural similarity between images by comparing their brightness, contrast, and structural information to quantify image quality. The comprehensive application of these two indicators provides us with a comprehensive and objective evaluation system to accurately measure the model's performance in image reconstruction and detail restoration.

4.3. Comparative experiments

We will comprehensively evaluate our method on five standard datasets: Set5, Urban100, BSD100, General100, and DIV2K100. To verify its performance, we will compare it with state-of-the-art methods such as SRGAN, SFTGAN[16], ESRGAN, BSRGAN, SPSR[17], SROOE[18], and ResShift[19]. All experiments are conducted based on the officially provided code and pre-trained weights to ensure the fairness and accuracy of the evaluation.

Table 1. Quantitative results analysis

Metrics	Benchmark	SRGAN	SFTGAN	ESRGAN	SPSR	BSRGA	SROOE(T=0)	ResShift	OURS
Training Dataset		DIV2K	ImageNet+ OST	DIV2K+ OST	DIV2K	DF2K+FFHQ+ WED	DF2K	ImageNet	DIV2K
PSNR↑	Set5	29.92	30.057	30.138	30.397	27.6409	29.15	27.8995	31.3576
	BSD100	26.9218	25.5169	25.288	25.495	25.6155	26.45	25.1475	**27.0213**
	Urban100	24.41	24.338	24.365	24.804	23.3834	25.21	23.9597	**25.3210**
	General100	29.327	29.159	29.425	29.424	27.4969	**30.08**	28.0902	29.8702
	DIV2K100	28.165	28.085	28.175	28.182	27.3121	29.33	26.511	**29.4500**
SSIM↑	Set5	0.8478	0.8483	0.8523	0.8626	0.8066	0.8453	0.8235	**0.8671**
	BSD100	0.7117	0.653	0.65	0.658	0.652	**0.7416**	0.6475	0.7032
	Urban100	0.7302	0.7235	0.7341	0.7673	0.6918	**0.8020**	0.726	0.7756
	General100	0.8074	0.8060	0.8095	0.8277	0.7752	0.8662	0.7939	**0.8683**
	DIV2K100	0.7745	0.7707	0.7759	0.7951	0.7569	0.8413	0.738	**0.8445**

The quantitative results are presented in Table 1, with bolded fields representing the optimal values for each evaluation metric. The data in the table clearly shows that our method demonstrates significant advantages in terms of the two key image quality evaluation metrics: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). Although our method slightly underperforms ESRGAN in terms of PSNR on the Set5 dataset, upon deeper analysis, we found that this is primarily due to the relatively smaller amount of training data used by our method. However, when compared across other datasets, our method exhibits clear superiority, further validating its generalization ability and robustness across different datasets. In terms of the SSIM metric, our method also performs impressively. Notably, on the Set5 dataset, despite using less training data, our method still exhibits a certain advantage in SSIM values. This fully demonstrates our method's capability in preserving image structure and recovering details. Nevertheless, we also observe that our SSIM values are relatively lower on the BSD100 and Urban100 validation sets. We conducted an in-depth analysis of this phenomenon and speculate that the possible reason is the lack of image data similar to these two validation sets in our training dataset. Therefore, in future work, we will focus on optimizing the training dataset by incorporating more diverse image data to enrich the training content, thereby improving the model's performance on these two validation sets.

Table 2. Quantitative comparison of DIV2KRK

DataSet		DIV2KRK											
Method		Bicubic	EDSR	RCAN	DBPN	IKC	DANv1	DANv2	AdaTarget	KOALAnet	DCLS	SRDDGAN	OUR
× 2	PSNR	28.73	29.17	29.20	29.13	-	32.56	32.58	-	31.89	32.75	-	**35.95**
	SSIM	0.80	0.82	0.82	0.82	-	0.89	0.90	--	0.89	0.91	-	**0.92**
× 4	PSNR	25.33	25.64	25.66	25.58	27.70	27.55	28.74	28.42	27.77	28.99	27.89	**29.45**
	SSIM	0.68	0.69	0.69	0.69	0.77	0.76	0.79	0.79	0.76	0.79	0.79	**0.85**

Table 2 presents quantitative results on the DIV2KRK test set, comparing the performance of different methods for 2x and 4x super-resolution tasks. We first compare our method with current state-of-the-art blind super-resolution methods such as IKC, DANv1[20], DANv2[21], AdaTarget[22], KOALAnet[23], and DCLS[24]. Secondly, we compare it with classic methods based on bicubic interpolation, including EDSR[25], RCAN[3], and DBPN. Finally, we also include comparisons with the latest GAN-based super-resolution method, SRDDGAN. The experimental results demonstrate that our proposed method significantly outperforms other compared methods in terms of both PSNR and SSIM, showcasing its superior performance in super-resolution tasks. It achieves remarkable improvements in both detail restoration and overall visual effects.

In the paper, we conducted an in-depth qualitative analysis on three images from the SET5 dataset: BABY, Butterfly, and head, using a total of seven different methods for comparison, including Bicubic, SRGAN, SFTHAN, ESRGAN, SPSR, BSRGAN, and ResShift. As shown in Figure 3, the analysis results reveal the performance differences among various methods in image super-resolution reconstruction. Taking the BABY image as an example, our method is compared with the current mainstream ResShift method. Although ResShift achieves certain results in image reconstruction, it generates more noise in the image, affecting visual quality. In contrast, although there is still a gap between our method and high-quality HR (High-Resolution) images in terms of processing effects, our method demonstrates significant advantages in detail preservation and noise reduction compared to other mainstream methods. We observed the analysis results for the Butterfly image. In the segment extracted from the HR image, there are a small number of noise points in the original image itself. However, after processing with our method, the image quality is significantly improved, even surpassing the visual perception of the original image in some aspects. Meanwhile, the BSRGAN method experiences over-sharpening during processing, resulting in slight color shifts that affect the naturalness and authenticity of the image. In contrast, our method is closer to the HR image in terms of detail processing and color restoration, demonstrating higher reconstruction quality.

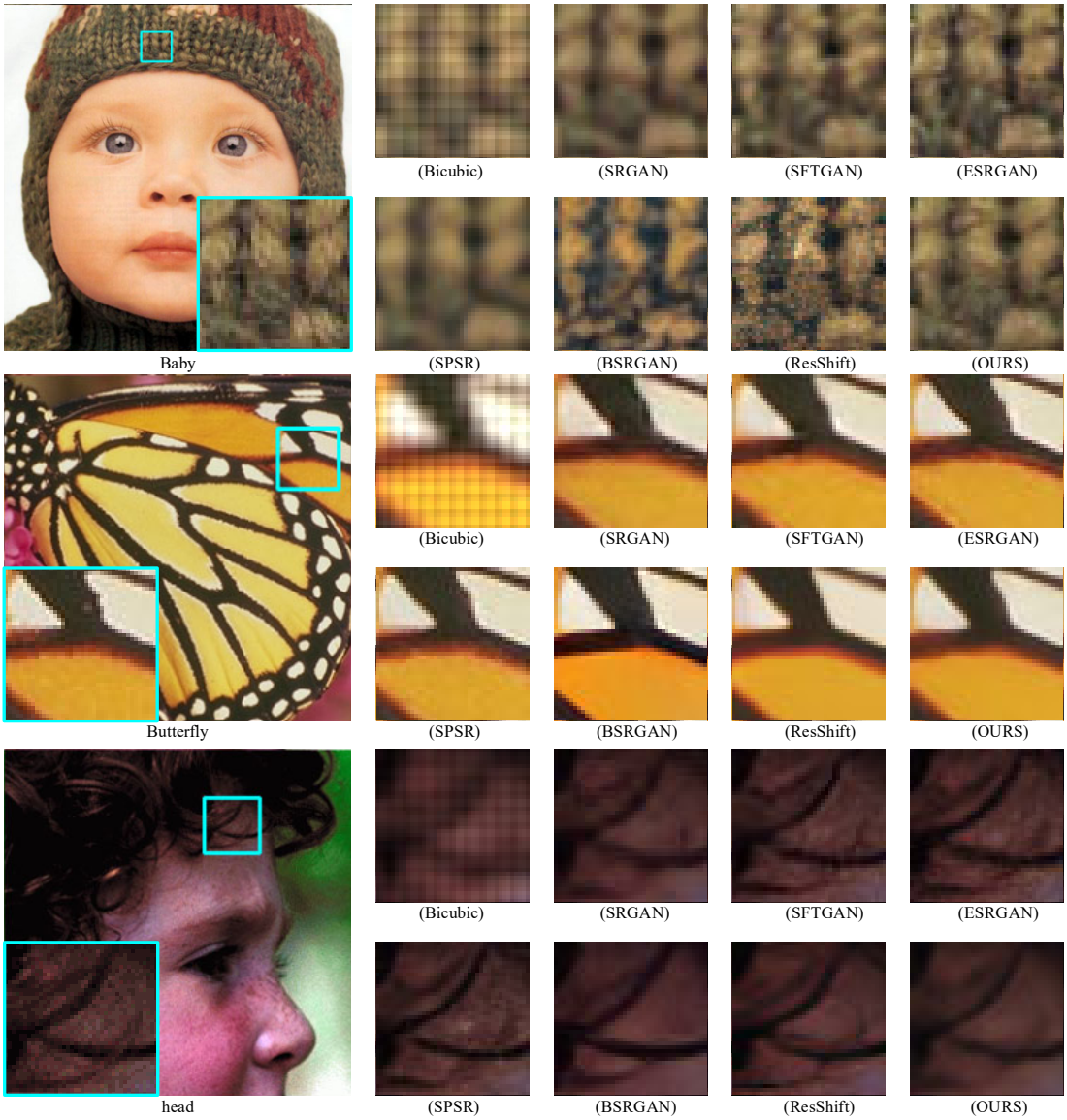

Figure 3. Qualitative analysis chart

4.4. Ablation studies

To investigate the impact of the effectiveness of each component in the proposed network architecture, we progressively modify the baseline model and compare the differences among them. Based on the principles of ablation experiments, we construct three different networks to illustrate the importance of each module. (1) We test the effectiveness using Swin Transformer as the encoder structure and LIIF as the decoder model structure. (2) We test the model effectiveness using the SRGAN generator as the encoder structure and LIIF as the decoder model structure. (3) We test the model effectiveness using Group Propagation Vision Transformer as the encoder structure and LIIF as the decoder model structure. The results of the ablation experiments are shown in Table 3.

Table 3. Analysis of ablation experiments

Methods	Dataset	PSNR	SSIM
Swin Transformer +LIIF	Set5	29.4916	0.8579
SRGAN+LIIF	Set5	27.3845	0.7372
GP-VIT+LIIF	Set5	31.3576	0.8671

5. Conclusion

This paper mainly studies the implicit function-based super-resolution reconstruction method utilizing Group Propagation Vision Transformer (GP-ViT). We deeply explore the architecture of the GP-ViT model and its efficiency in global information exchange. By integrating the continuous representation of implicit functions, we achieve continuous modeling of image features, enabling the reconstruction of richer and more coherent high-resolution images in detail. GP-ViT significantly reduces the computational complexity and memory consumption of the model through a group propagation mechanism, while enhancing the local feature extraction capability. This allows the algorithm to excel in preserving global information and recovering local details. Comparative experiments with current mainstream super-resolution algorithms, such as IKC, DAN, EDSR, RCAN, and GAN-based methods, demonstrate that our method achieves high levels on both PSNR and SSIM key indicators, especially exhibiting significant advantages in preserving global information and suppressing artifacts. Despite significant advancements in super-resolution technology in recent years, it still faces challenges such as training instability, insufficient high-frequency detail

recovery, and limited adaptability to complex scenes. Our research provides a new approach to addressing these issues by combining GP-ViT and implicit function representation, balancing computational efficiency while improving reconstruction quality, offering valuable insights for the future development of super-resolution technology.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang, 'Learning a Deep Convolutional Network for Image Super-Resolution', in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 184–199. doi: 10.1007/978-3-319-10593-2_13.

[2] J. Kim, J. K. Lee, and K. M. Lee, 'Accurate Image Super-Resolution Using Very Deep Convolutional Networks', presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654. Accessed: Feb. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Kim_Accurate_Image_Super-Resolution_CVPR_2016_paper.html

[3] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, 'Image Super-Resolution Using Very Deep Residual Channel Attention Networks', in Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Berlin, Heidelberg: Springer-Verlag, Sep. 2018, pp. 294–310. doi: 10.1007/978-3-030-01234-2_18.

[4] H. Chen et al., 'Pre-Trained Image Processing Transformer', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12299–12310. Accessed: Feb. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Pre-Trained_Image_Processing_Transformer_CVPR_2021_paper.html

[5] C. Ledig et al., 'Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 105–114. doi: 10.1109/CVPR.2017.19.

[6] X. Wang et al., 'ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks', in Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 63–79. doi: 10.1007/978-3-030-11021-5_5.

[7] K. Zhang, W. Zuo, and L. Zhang, 'Learning a Single Convolutional Super-Resolution Network for Multiple Degradations', presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3262–3271. Accessed: Feb. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Learning_a_Single_CVPR_2018_paper.html

[8] J. Gu, H. Lu, W. Zuo, and C. Dong, 'Blind Super-Resolution With Iterative Kernel Correction', in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, pp. 1604–1613. doi: 10.1109/CVPR.2019.00170.

[9] C. Yang, J. Xu, S. D. Mello, E. J. Crowley, and X. Wang, 'GPViT: A High Resolution Non-Hierarchical Vision Transformer with Group Propagation', Apr. 25, 2023, arXiv: arXiv:2212.06795. doi: 10.48550/arXiv.2212.06795.

[10] Y. Chen, S. Liu, and X. Wang, 'Learning Continuous Image Representation with Local Implicit Image Function', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 8624–8634. doi: 10.1109/CVPR46437.2021.00852.

[11] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. A. Morel, 'Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding', presented at the British Machine Vision Conference (BMVC), 2012. Accessed: Jul. 29, 2024. [Online]. Available: https://inria.hal.science/hal-00747054

[12] J.-B. Huang, A. Singh, and N. Ahuja, 'Single image super-resolution from transformed self-exemplars', in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 5197–5206. doi: 10.1109/CVPR.2015.7299156.

[13] Y. Matsui et al., 'Sketch-based manga retrieval using manga109 dataset', Multimed Tools Appl, vol. 76, no. 20, pp. 21811–21838, Oct. 2017, doi: 10.1007/s11042-016-4020-z.

[14] D. Martin, C. Fowlkes, D. Tal, and J. Malik, 'A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics', in Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Jul. 2001, pp. 416–423 vol.2. doi: 10.1109/ICCV.2001.937655.

[15] E. Agustsson and R. Timofte, 'NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study', in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jul. 2017, pp. 1122–1131. doi: 10.1109/CVPRW.2017.150.

[16] X. Wang, K. Yu, C. Dong, and C. Change Loy, 'Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform', in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 606–615. doi: 10.1109/CVPR.2018.00070.

[17] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, 'Structure-Preserving Super Resolution With Gradient Guidance', in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, pp. 7766–7775. doi: 10.1109/CVPR42600.2020.00779.

[18] J. Park, S. Son, and K. M. Lee, 'Content-Aware Local GAN for Photo-Realistic Super-Resolution', in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2023, pp. 10551–10560. doi: 10.1109/ICCV51070.2023.00971.

[19] Z. Yue, J. Wang, and C. C. Loy, 'ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting', Advances in Neural Information Processing Systems, vol. 36, pp. 13294–13307, Dec. 2023.

[20] zhengxiong luo, Y. Huang, S. Li, L. Wang, and T. Tan, 'Unfolding the Alternating Optimization for Blind Super Resolution', in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 5632–5643. Accessed: Jul. 29, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/3d2d8ccb37df977cb6d9da15b76c3f3a-Abstract.html

[21] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, 'End-to-end Alternating Optimization for Blind Super Resolution', May 14, 2021, arXiv: arXiv:2105.06878. doi: 10.48550/arXiv.2105.06878.

[22] Y. Jo, S. Wug Oh, P. Vajda, and S. Joo Kim, 'Tackling the Ill-Posedness of Super-Resolution through Adaptive Target Generation', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 16231–16240. doi: 10.1109/CVPR46437.2021.01597.

[23] S. Y. Kim, H. Sim, and M. Kim, 'KOALAnet: Blind Super-Resolution using Kernel-Oriented Adaptive Local Adjustment', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 10606–10615. doi: 10.1109/CVPR46437.2021.01047.

[24] Z. Luo, H. Huang, L. Yu, Y. Li, H. Fan, and S. Liu, 'Deep Constrained Least Squares for Blind Image Super-Resolution',

in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 17621–17631. doi: 10.1109/CVPR52688.2022.01712.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, 'Enhanced Deep Residual Networks for Single Image Super-Resolution', presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society, Jul. 2017, pp. 1132–1140. doi: 10.1109/CVPRW.2017.151.