# Intensify Perception Transformer Generative Adversarial Network for Image Super-Resolution

Yuzhen Chen[1,2,3], Gencheng Wang[1,2,3], Rong Chen[1,2,3(✉)], and Zi Hui[1,2,3]

[1] Xizang Minzu University, Xianyang 712000, Shaanxi, China
{wgc,cr2008}@xzmu.edu.cn
[2] Key Laboratory of Optical Information Processing and Visualization Technology of Tibet Autonomous Region, Xianyang 712000, Shaanxi, China
[3] Xizang Cyberspace Governance Research Center, Xianyang 712000, Shaanxi, China

**Abstract.** Generative adversarial networks (GANs) are widely used for image super-resolution (SR) and have recently attracted increasing attention due to their potential to generate rich details. However, generators are usually based on convolutional neural networks, which lack global modeling capacity and limit the performance of the network. To address this problem, we propose a hierarchical partitioned Transformer block to extract features at different scales, which alleviates the loss of information and helps global modelling. We then design a Transformer in residual block to reconstruct more natural structural textures in SR results. Finally, we integrate the intensify perception Transformer network with an existing discriminator network to form the intensify perception Transformer generative adversarial network (IPTGAN). We conducted experiments on several benchmark datasets, RealSR dataset and PIRM self-validation dataset to verify the generalization ability of our IPTGAN. The results show that our IPTGAN exhibits better visual quality and significantly less complexity compared to several state-of-the-art GAN-based image SR methods.

**Keywords:** Image super-resolution · GAN · Transformer · Moderated self-attention · Intensify perception

## 1 Introduction

Image super-resolution (SR), which aims to generate a high-resolution (HR) image from a given low-resolution (LR) image by attempting to recover the missing information, is a low-level computer vision (CV) task. Since the pioneering work of SRCNN [5], deep convolutional neural networks (CNNs) have brought prosperous development to the field of image SR. Peak signal-to-noise ratio (PSNR) has been used as a measure for various SR networks, but the PSNR metric fundamentally diverges from the subjective evaluation of human observers. As a result, PSNR-oriented methods tend to produce smoother results without

H. Lu et al. (Eds.): ICIG 2023, LNCS 14358, pp. 310–321, 2023.
https://doi.org/10.1007/978-3-031-46314-3_25

sufficient high-frequency details. To address this issue, several perceptual-driven methods have been proposed to improve the visual quality of SR results. For instance, the perceptual loss [9] is proposed to optimize SR methods in a feature space instead of a pixel space.

Generative adversarial network (GAN) consists of a generator network responsible for generating SR images and a discriminator network that tries to distinguish between SR images and real HR images. Through the competition of the generator and the discriminator, the GAN is encouraged to favor images that look more real. The original GAN [4,7] used a fully-connected network and was limited to generating small images. One milestone in achieving visually pleasing results is SRGAN [12]. The basic block of SRGAN is built with residual blocks and optimized using perceptual loss. With these techniques, SRGAN significantly improves the overall visual quality of reconstructions compared to PSNR-oriented methods. DCGAN [19] was the first to scale up GAN using CNN, which allowed for stable training at higher resolutions and with deeper generator. ESRGAN [21], as a representative work, proposed a practical perceptual loss as well as a residual in residual dense block (RRDB) to produce SR images with convincing visual quality. Since then, using CNNs as GAN backbone in CV has become a common practice. However, CNNs have a limited receptive field, which makes it inefficient to process long-range dependencies without passing through sufficient layers. This can result in a loss of feature information and fine details, leading to high computational costs and optimization difficulties.

Recently, Transformers have demonstrated effectiveness in global modeling and have been applied to various CV tasks, such as image classification, object detection, semantic segmentation and SR. It is important to note that while Transformer-based networks generally have higher computational complexity compared to CNNs, the utilization of self-attention in Transformers greatly enhances the expressive power of the model. The self-attention enables network to model dependencies effectively, allowing it to focus on comprehensive information. Taking inspiration from the above, we propose a perceptual-driven Transformer-based GAN, called the intensify perception Transformer generative adversarial network (IPTGAN), to address the aforementioned limitations and drawbacks. First, we improve the network structure by introducing the Transformer and residual connections to enhance the information flow and better learn the features of the data. We further introduce hierarchization and partition into different size strategies to the Transformer, allowing for a flexible receptive field and enabling global modeling. Additionally, we propose a moderated self-attention (MSA) enabling the network to learn more information. The contributions of our work can be summarized as follows:

- We propose a Transformer in residual block (TRB) that enables the network to capture more pixel information, resulting in improved result quality. The TRB is efficient and extensible, it can be easy to integrate into SR networks.
- We propose a intensify perception Transformer network (IPTNet), which is a generator with excellent scalability. It can be combined with existing discriminators to form GANs, achieving excellent SR results.

- We propose IPTGAN, a perception-driven yet powerful GAN, to efficiently address the SISR problem. IPTGAN performs well not only on benchmark datasets but also on RealSR dataset and PIRM self-validation dataset, achieving superior visual results compared to several state-of-the-art GAN-based methods. Furthermore, IPTGAN requires significantly fewer parameters than ESRGAN, making it more practical in real-world applications.

## 2 Related Work

### 2.1 GAN-Based SR Methods

GANs are a class of generative models that are learned through a minimax optimization game between a generator network and a discriminator network. The GANs have been proven to be competitive in learning mappings among manifolds and thus improving local textures. SRGAN [12] was the first to introduce GAN into SR, where the generator was composed of residual blocks. To enhance the results, SRGAN employed perceptual and adversarial losses for training. EnhanceNet [20] and SRFeat [18] utilized multiple loss terms or discriminators to improve performance. ESRGAN [21] further improved the performance of SRGAN by proposing RRDB, removing batch normalization layers and employing the relativistic discriminator [10]. Although the RRDB has demonstrated effectiveness, it still has a significant number of parameters, resulting in considerable computational costs. BSRGAN [23] is a blind SR method that performs SR by designing a complex degradation process that mimics real-world conditions. BSRGAN incorporates a pixel alignment technique to correct spatial distortion and ensure pixel-level matching between the SR image and the HR image. However, it still faces the issue of over-smoothing in SR images.

### 2.2 Transformer-Based SR Methods

Transformer was initially developed for natural language processing, researchers have found that the self-attention in the Transformer effectively models dependencies among data. ViT [6] was the first to introduce the Transformer into CV by achieving highly competitive ImageNet classification results, treating an image as a sequence of $16 \times 16$ visual words. Swin Transformer [15] adopts a similar idea to ViT, introducing shifted window mechanism to enhance performance. However, it has a high computational complexity, especially for large input images. SwinIR [13] inherits the Swin Transformer for SR task and achieves impressive results. However, it also inherits many components that were designed for high-level CV task, making them redundant and fragmented for SR. Swin Transformer V2 [14] improves upon the Swin Transformer by using larger window sizes and a new data adaptive training strategy. However, it involves a more complex training process and requires additional time and computational costs. Although Swin Transformer and Swin Transformer V2 did not specifically introduce the Transformer into GANs and SR tasks, they are indeed representative works in CV.

# 3 Methods

As previously stated, our main aim is to enhance the overall perceptual quality of the SR images. The IPTGAN follows the same principle as the traditional GAN, where the competition between the two networks enables the generator to produce images that are more realistic and closer to the ground truths. The IPTNet is illustrated in Fig. 1.

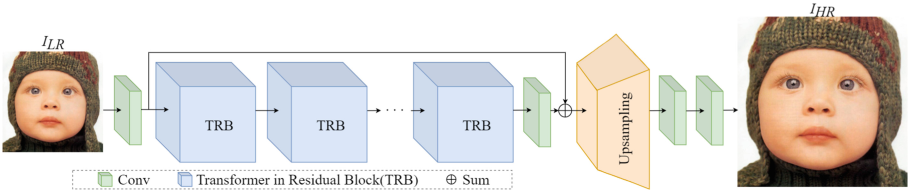

**Fig. 1.** The architecture of the proposed IPTNet for IPTGAN.

## 3.1 Generator

The IPTNet comprises three main components: a shallow feature extraction module, a deep feature extraction module and an image reconstruction module. The shallow feature extraction module includes a convolutional layer, while the deep feature extraction module consists of multiple TRBs and a convolutional layer. The reconstruction module consists of a upsampling layer and two convolutional layers.

**Transformer in Residual Block.** The design of the TRB is inspired by the RRDB, which has become a classical algorithm in this field by combining multiple convolutional layers with a dense connection to achieve a deep network structure. However, the limited receptive field of convolution makes it inefficient to process long-range dependencies without passing through sufficient layers. Additionally, training a deep enough network presents significant computational and time costs. As shown in Fig. 2(a), the TRB consists of three hierarchical partitioned Transformer blocks (HPTBs) and a residual connection. To avoid imposing unnecessary burdens on the TRB, we have removed the dense connection, as each HPTB already has two residual connections. By leveraging the strong modeling capabilities of the Transformer, we can achieve better results with significantly fewer parameters and layers.

**Hierarchical Partitioned Transformer Block.** HPTB adopts the classic Transformer framework. The input features will pass-through layer normalization (LayerNorm), hierarchical partitioned moderated self-attention (HPMS) shown in Fig. 2(b), LayerNorm and multi-layer perceptron (MLP) in sequence.

Convolutions and previous Transformers extract features at a fixed size. However, the fixed size is not directly related to the image contents, and it may lead to the loss of pixel information at the edges of the split blocks. To address this issue, we divide the input channels into $k$ groups and then into blocks of different sizes for MSA calculation. Additionally, shifted window machine shifts the window in a diagonal direction, then extracting the shifted features. This approach further mitigates the loss of pixel information and facilitates communication among surrounding pixels, resulting in improved image generation.

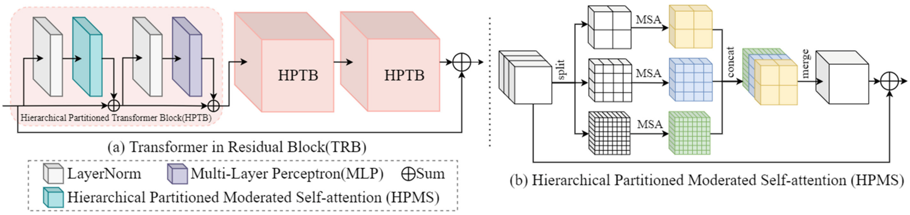
Hierarchical Partitioned Transformer Block(HPTB)
HPTB
HPTB
(a) Transformer in Residual Block(TRB)
LayerNorm
Multi-Layer Perceptron(MLP)
⊕Sum
Hierarchical Partitioned Moderated Self-attention (HPMS)
split
MSA
MSA
MSA
concat
merge
(b) Hierarchical Partitioned Moderated Self-attention (HPMS)

**Fig. 2.** The process of Transformer in residual block. (a) is the Transformer in residual block. (b) is the hierarchical partitioned moderated self-attention.

**Moderated Self-attention.** The conventional self-attention in the Transformer employs dot-product to measure the similarity between the query vector ($q$) and the key vector ($k$) of a pixel pair. However, this method often produces extreme values that will lead to attention being disproportionately focused on a few pixels, resulting in suboptimal reconstructions. In contrast, cosine similarity is naturally normalized. By leveraging cosine for self-attention, more eased values can be obtained, offering a more accurate measurement of similarity between vectors. Unlike the commonly used SoftMax, which is more suitable for classification tasks, our approach generates attention values that are better suited for SR tasks. The proposed MSA calculation can be expressed as follows:

$$f_{msa} = \frac{q \cdot k}{||q|| \times ||k||} \cdot v/\tau \tag{1}$$

In (1), where the $\tau$ is a learnable scalar and the $v$ is the value vector. By this method, the values of MSA are distributed more evenly so that more information can be noticed and learned.

Self-attention can be time-consuming, especially when dealing with input feature of large size. To address this issue and enhance network training efficiency, HPTB incorporates the shared attention mechanism derived from ELAN. This approach enables network to calculate self-attention only for specific HPMSs, while the subsequent HPMSs at the same scale can directly reuse the pre-computed attention values. Consequently, network eliminate two reshapes and one convolution operations for each self-attention calculation. Although this method results in a slight reduction in SR performance, the impact is negligible in light of the substantial reduction in computational costs and time required.

### 3.2 Discriminator

It is well-known that pixel-wise PSNR-oriented SR methods often result in over-smoothed results and fail to adequately recover high-frequency details. The discriminator is trained to discriminate between the generated SR image ($I_{SR}$) and the HR image ($I_{HR}$). We adopt the relativistic discriminator introduced in ESRGAN, which differs from the standard discriminator in SRGAN. Instead of estimating the probability of $I_{SR}$ being real and natural like the standard discriminator, the relativistic discriminator aims to predict the relative realism between $I_{SR}$ and $I_{HR}$. This utilization of the relativistic discriminator enables the generation of sharper edges and more realistic texture details.

### 3.3 Losses

In order to ensure consistency between the content of $I_{SR}$ and $I_{HR}$, our IPTGAN is trained using a combination of multiple loss functions, which can be formulated as follows:

$$L_G = L_p + \lambda L_G^{Ra} + \eta L_1 \tag{2}$$

where $L_1 = E_{I_{LR}}||I_{HR} - I_{SR}||_1$ denotes the content loss, measuring the 1-norm distance between $I_{HR}$ and $I_{SR}$. The $L_p$ represents the perceptual loss proposed by ESRGAN, while the $\lambda$ and the $\eta$ are coefficients used to balance the different loss terms. In (2), $L_G^{Ra}$ is defined as:

$$L_G^{Ra} = -E_{I_{HR}}[log(1 - D_{Ra}(I_{HR}, I_{SR}))] - E_{I_{SR}}[log(D_{Ra}(I_{HR}, I_{SR}))] \tag{3}$$

where $D_{Ra}$ refers to the standard discriminator with the relativistic average discriminator [10]. The $E_{I_{HR}}[\cdot]$ and $E_{I_{SR}}[\cdot]$ represents the operation of averaging over all real and fake data within the mini-batch, respectively.

## 4 Experiments

### 4.1 Training Details

Following ESRGAN, all experiments are performed with a scaling factor of ×4 between $I_{LR}$ and $I_{HR}$. We obtain the $I_{LR}$ by bicubic the $I_{HR}$. For training data, we utilize the DIV2K dataset [1], which comprises 800 high-quality images. We train the IPTGAN in RGB channels and augment the training dataset with random horizontal flips and 90° rotations. We evaluate the IPTGAN on several benchmark datasets: Set14 [22], BSD100 [16], Urban100 [8] and Manga109 [17]. We further test our IPTGAN on RealSR dataset [3] and PIRM self-validation dataset [2].

The IPTNet is trained using the perceptual loss with $\lambda = 5 \times 10^{-3}$ and $\eta = 1 \times 10^{-2}$. The learning rate is set to $1 \times 10^{-4}$ and halved at $[50k, 100k, 200k, 300k]$ iterations. The window sizes of HPMS are set to $4 \times 4$, $8 \times 8$ and $16 \times 16$. The shared attention is set to $n = 1$, i.e., only calculate the first HPMS. We use Adam [11] and alternately update the generator and discriminator networks until the model converges. The IPTGAN is implemented using PyTorch on NVIDIA 3080Ti GPU.

## 4.2 Quantitative Evaluation

As shown in Table 1, we compared our IPTGAN with three state-of-the-art GAN-based SR methods, namely SRGAN [12], BSRGAN [23] and ESRGAN [21], using several public benchmark datasets. Remarkably, despite having significantly fewer parameters, IPTGAN consistently outperformed all other methods in terms of PSNR, structure similarity index measure (SSIM) and perceptual index (PI). We further compared the IPTNet and IPTGAN on several benchmark datasets, the result is shown in Table 2.

**Table 1.** PSNR/SSIM/PI comparisons of IPTGAN and several state-of-the-art GAN-based SR methods on several benchmarks at ×4.

| Model | Param (K) | FLOPs (G) | Set14 (PSNR/SSIM/PI) | BSD100 (PSNR/SSIM/PI) | Urban100 (PSNR/SSIM/PI) | Manga109 (PSNR/SSIM/PI) |
|---|---|---|---|---|---|---|
| SRGAN | 1547 | 231 | 24.21/0.6349/1.31 | 23.68/0.5990/1.32 | –/– | –/– |
| BSRGAN | 16697 | 1835 | 23.60/0.6295/1.46 | 23.91/0.6084/1.50 | 21.55/0.6467/1.63 | 22.84/0.7529/1.43 |
| ESRGAN | 16697 | 1859 | **24.17**/0.6440/1.25 | 23.45/0.5975/1.27 | 21.99/0.6707/1.40 | 24.93/0.7838/1.02 |
| IPTGAN | 8212 | 596 | 23.61/**0.6468**/**1.23** | **23.94**/**0.6263**/**1.20** | **22.52**/**0.6984**/**1.34** | **25.03**/**0.8061**/**0.95** |

**Table 2.** PSNR/SSIM/PI comparisons of IPTGAN and IPTNet on several benchmarks at ×4.

| Model | Set14 (PSNR/SSIM/PI) | BSD100 (PSNR/SSIM/PI) | Urban100 (PSNR/SSIM/PI) | Manga109 (PSNR/SSIM/PI) |
|---|---|---|---|---|
| IPTNet | **24.57**/0.6451/1.41 | **24.98**/0.6237/1.45 | **23.47**/0.6955/1.58 | **26.21**/0.8012/1.29 |
| IPTGAN | 23.61/**0.6468**/**1.23** | 23.94/**0.6263**/**1.20** | 22.52/**0.6984**/**1.35** | 25.03/**0.8061**/**0.95** |

## 4.3 Qualitative Results

We compared our IPTGAN with SRGAN [12], BSRGAN [23] and ESRGAN [21] on several benchmark datasets. Since SRGAN was not evaluated on Urban100, we present the comparison graphs separately in Fig. 3 and Fig. 4. As shown in these figures, the IPTGAN generates more natural and realistic effects such as stairs, cactus and tiger stripes. The restored images of wolves, fences and holes exhibit better overall visual consistency with the ground truth. In contrast, other methods tend to produce images that are either too smooth or too sharp. Results demonstrate that IPTGAN achieves competitive performance under the same scaling factor. The PI values also show the SR images generated by IPTGAN outperform other methods. Noteworthy, the parameters of IPTGAN are significantly less and achieved SR results that are more consistent with the ground truth and human visual effects.

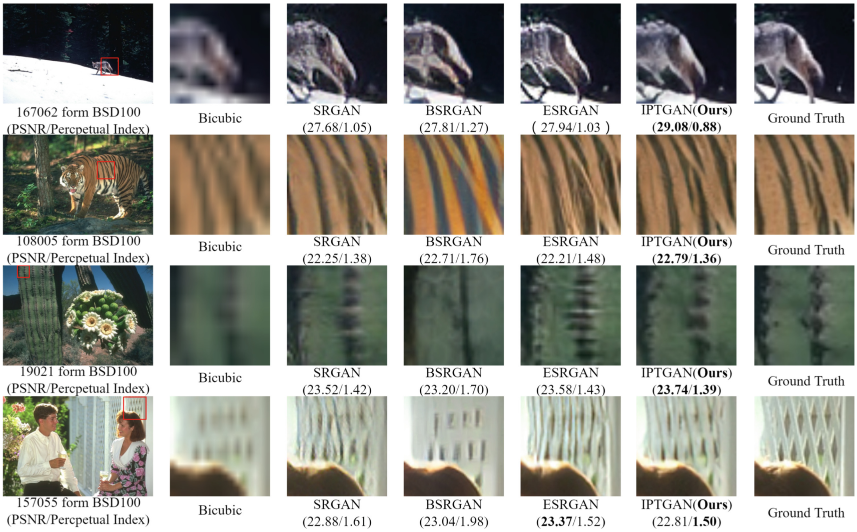

**Fig. 3.** Visual comparison of IPTGAN with other GAN-based SR methods at ×4. The best values are in **bold faces. Please zoom in for the best view**.

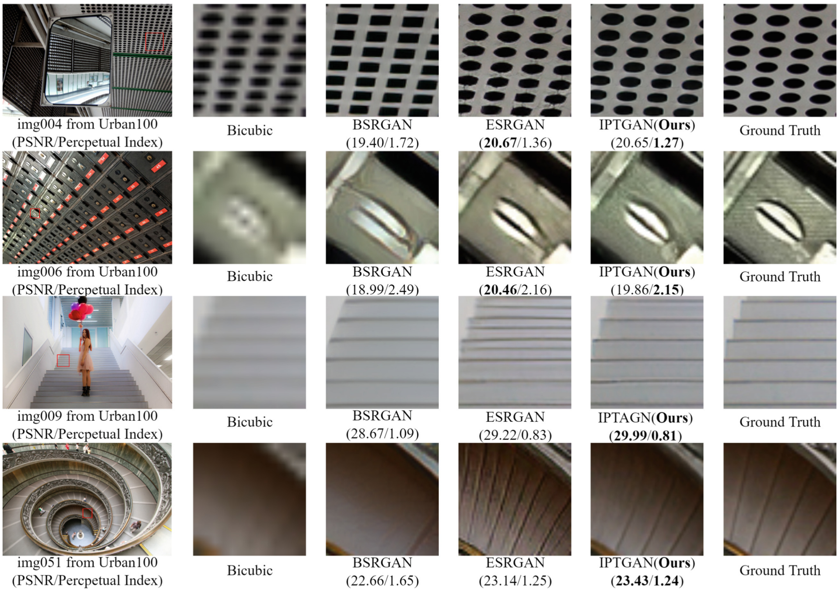

**Fig. 4.** Visual comparison of IPTGAN with other GAN-based SR methods on Urban100 dataset ×4. The best values are in **bold faces. Please zoom in for the best view**.

### 4.4 Ablation Study

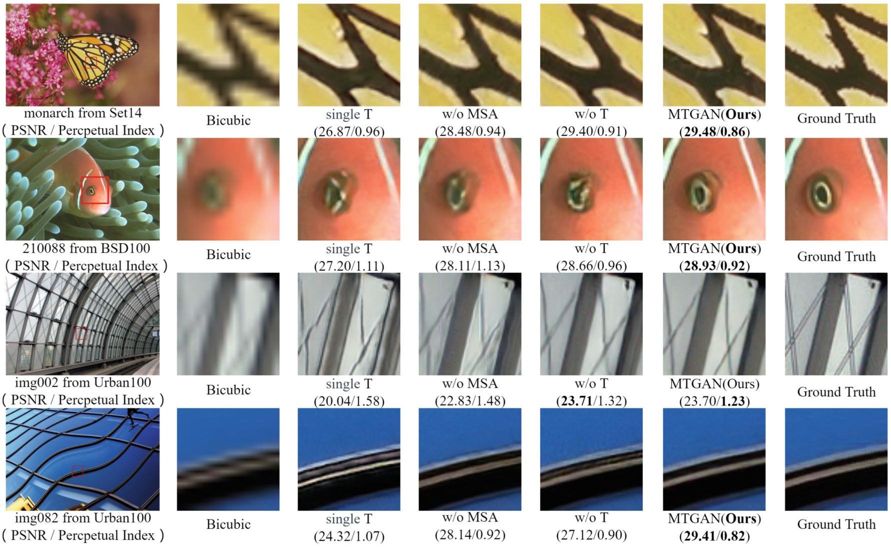

**Fig. 5.** Overall visual comparison of the effects of each design in IPTGAN in the ablation study at ×4. The best values are in **bold faces. Please zoom in for the best view**.

**Table 3.** The ablation experimental results of IPTGAN on several benchmarks at ×4.

| Model | Set14<br>(PSNR/SSIM/PI) | BSD100<br>(PSNR/SSIM/PI) | Urban100<br>(PSNR/SSIM/PI) | Manga109<br>(PSNR/SSIM/PI) |
|---|---|---|---|---|
| single T | 22.48/0.5934/1.37 | 22.67/0.5704/1.37 | 20.57/0.6193/1.66 | 22.84/0.7435/1.18 |
| w/o MSA | 23.40/0.6291/1.33 | 23.60/0.6076/1.32 | 21.74/0.6644/1.56 | 24.29/0.7861/ 1.07 |
| w/o T | 23.56/0.6349/1.27 | 23.78/0.6172/1.27 | 22.26/0.6821/1.49 | 24.77/0.7938/1.00 |
| IPTGAN | **23.61/0.6468/1.23** | **23.94/0.6263/1.20** | **22.52/0.6984/1.35** | **25.03/0.8061/0.95** |

To demonstrate the effectiveness of our design, we conducted several ablation studies. The Fig. 5 and Table 3 illustrate the impact of each design component in IPTGAN. As expected, when self-attention employs dot product for calculations (w/o MSA), the resulting SR images exhibit blurring and artifacts. Similarly, when the two LayerNorm layers and MLP of HPTB are removed (w/o T), the SR image becomes oversharpened in certain regions compared to the ground truth. While the single T indicates simply stacking HPTBs, the reconstructed results still lack naturalness. This is because the mere stacking of HPTBs fails to achieve the effect obtained by a set of three HPTBs and a residual connection. The SR images produced by IPTGAN are visually pleasing, displaying more natural textures and edges without noticeable artifacts.

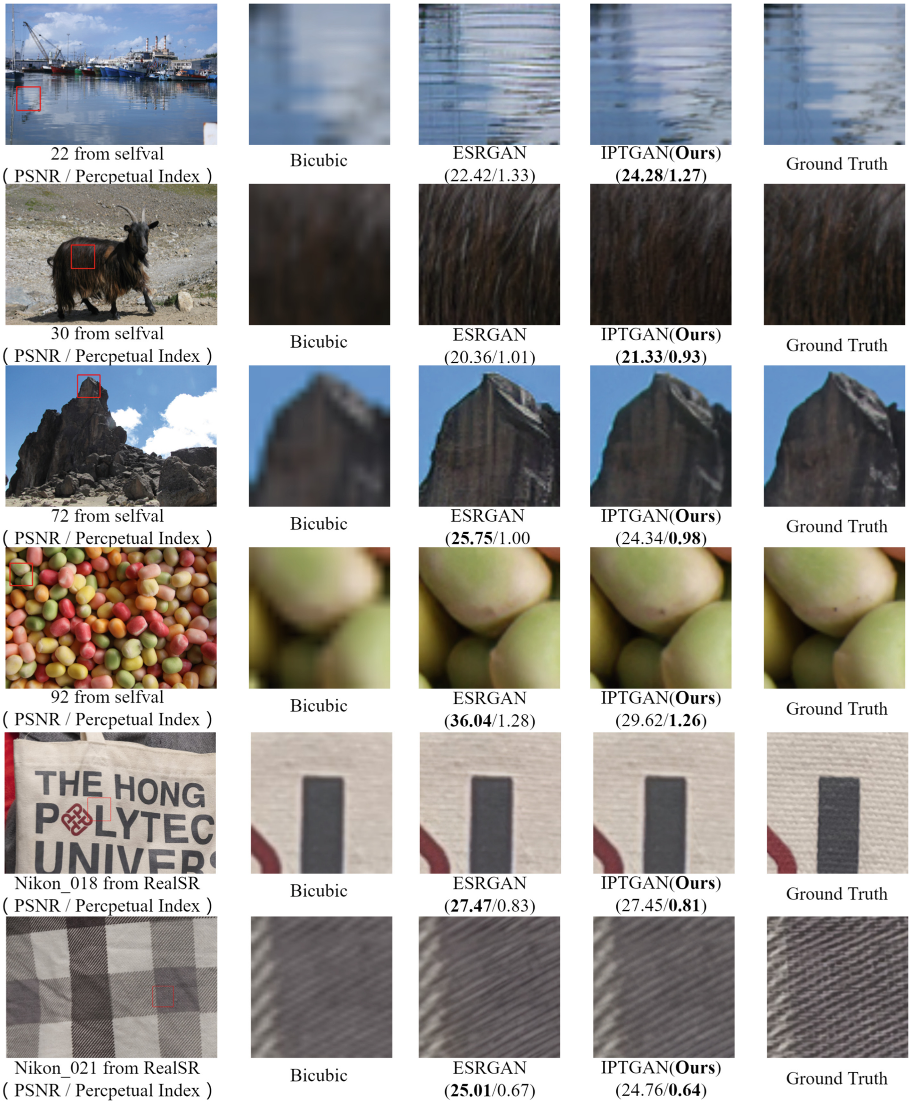

**Fig. 6.** Visual comparison of IPTGAN with ESRGAN on RealSR dataset and PIRM self-validation dataset at ×4. The best values are in **bold faces. Please zoom in for the best view**.

### 4.5 Generalization Ability

We evaluated IPTGAN on RealSR dataset and PIRM self-validation dataset, which are used to evaluate the performance of SR methods in real-world scenarios [2]. These datasets consist of images from various scenes, including natural landscapes, urban buildings and portraits. Since ESRGAN is also tested on the

PIRM self-validation dataset, in this section we only show comparisons of the visual effects with ESRGAN as shown in Fig. 6. Our IPTGAN successfully reconstructs softer lines in wool, lake surfaces and textile, aligning with the subjective evaluation of human observers and exhibiting better consistency with the ground truth compared to ESRGAN. By demonstrating SR results of these challenging datasets, IPTGAN showcases its enhanced generalization ability and adaptability to process real-world images.

## 5 Conclusion

In this paper, we propose a Transformer-based SR generator that leverages both the hierarchization and partition into different size strategies, as well as the moderated self-attention, to enhance pixel-to-pixel communication and make more information available for learning. The proposed TRB further improves the performance of the IPTNet. Experimental results show that IPTGAN surpasses several state-of-the-art GAN-based SR methods in several benchmark datasets while utilizing significantly fewer parameters. The generalization ability test demonstrates the abilities of IPTGAN. Although IPTGAN exhibits promising results, we acknowledge a limitation where the SR image may lack sufficient high-frequent details in specific areas. Our future research will focus on enhancing the network's capability to generate high-frequent details.

## References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: dataset and study. In: CVPRW (2017)
2. Blau, Y., Mechrez, R., Timofte, R., et al.: The 2018 PIRM challenge on perceptual image super-resolution. In: ECCV (2018)
3. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: a new benchmark and a new model. In: International Conference on Computer Vision, ICCV, pp. 3086–3095 (2019)
4. Denton, E.L., Chintala, S., Szlam, A., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: NeurIPS (2015)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: ICLR (2021)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: NeurIPS (2014)
8. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015)
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43

10. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. In: ICLR (2019)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
12. Ledig, C., Theis, L., Huszar, F., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
13. Liang, J., Cao, J., et al.: Swinir: image restoration using swin transformer. In: ICCVW (2021)
14. Liu, Z., Hu, H., Lin, Y., et al.: Swin transformer V2: scaling up capacity and resolution. In: CVPR (2022)
15. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV (2021)
16. Martin, D.R., Fowlkes, C.C., Tal, D., et al.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
17. Matsui, Y., et al.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl. **76**, 21811–21838 (2017)
18. Park, S., Son, H., Cho, S., Hong, K., Lee, S.: Srfeat: single image super-resolution with feature discrimination. In: ECCV (2018)
19. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
20. Sajjadi, M.S.M., Schölkopf, B., Hirsch, M.: Enhancenet: single image super-resolution through automated texture synthesis. In: ICCV (2017)
21. Wang, X., Yu, K., Wu, S., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: ECCV (2018)
22. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces (2010)
23. Zhang, K., Liang, J., Gool, L.V., et al.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV (2021)