Incorporating Local Texture Adversarial Branch and Hybrid Attention for Image Super-Resolution

Na Zhang[1], Hanhao Yao[2], Qingqi Zhang[3], Xiaoan Bao[4], Biao Wu[5], Xiaomei Tu[6]
School of Artificial Intelligence, Zhejiang Sci-Tech University, China[1,2,4,5]
Yamaguchi University, Japan[3]
ZGUC, China[6]

***Abstract*—In the field of image Super-Resolution reconstruction (SR), traditional SR techniques such as regression-based methods and CNN-based models fail to retain texture details in the reconstructed images. Conversely, Generative Adversarial Networks (GANs) have significantly enhanced the visual quality of image reconstruction through their adversarial training architecture. However, existing GANs still exhibit limitations in capturing local details and efficiently utilizing features. To address these challenges, we have proposed a super-resolution reconstruction method leveraging local texture adversarial and hybrid attention mechanisms. Firstly, a Local Texture Sampling Module (LTSM) is designed to precisely locate small regions with strong texture features within an image, and a local discriminator then performs pixel-by-pixel evaluation on these regions to enhance local texture details. Secondly, a hybrid attention module is integrated into the generator's residual module to improve feature utilization and representativeness. Finally, we conducted extensive experiments to validate the effectiveness of our method. The results demonstrate that our method surpasses other super-resolution reconstruction methods in terms of PSNR and SSIM on four benchmark datasets. Furthermore, our method visually generates high-resolution images with richer details and more realistic textures.**

***Keywords*—Super-resolution reconstruction; generative adversarial network; hybrid attention; local texture sampling**

I. Introduction

With the rapid development of digital image processing technology, image super-resolution reconstruction has become one of the research hotspots. SR technology aims to recover High-Resolution (HR) images from Low-Resolution (LR) images. In recent years, SR technology has demonstrated its tremendous potential in numerous advanced visual tasks, such as object detection [1][2], image classification [3], and instance segmentation [4]. Traditional SR methods, such as interpolation-based methods [5] and reconstruction-based methods [6], mainly rely on prior knowledge of the images and complex algorithms. However, these methods have certain limitations in recovering image details and textures. Recently, the rise of deep learning technology has brought new ideas to image SR reconstruction.

Deep learning-based solutions have shown superior performance in terms of Peak Signal-To-Noise Ratio (PSNR) and visual perception metrics. For example, Dong et al. [7]. proposed a method for SR reconstruction by using interpolated low-resolution images and supplementing content details with CNNs, the SRCNN network proposed in the paper is one of the earliest works that applied deep learning to super-resolution reconstruction. The EDSR network [8], the champion solution of the NTIRE2017 super-resolution challenge, streamlined the network by removing some unnecessary structures in the residual structure and proposed a multi-scale deep super-resolution system, which performs well under different super-resolution scales. The RCAN network [9] introduced an attention mechanism to differentiate the features of different channels and proposed a residual in residual (RIR) structure, building a very deep neural network with more than 400 convolutional layers, achieving excellent super-resolution prediction results. Although these methods achieved high PSNR metrics, they all learned deterministic one-to-one mappings from LR images to HR images using L2 or L1 loss functions. Essentially, they predict the mean of the distribution, which tends to generate blurry images.

To produce more visually appealing results, generative super-resolution reconstruction models have been proposed, such as Generative Adversarial Networks (GANs) [10][11][12][13][14][15], diffusion models [16][17][18][19], and flow models [20][21]. Among them, GAN-based super-resolution reconstruction methods have significantly improved the visual effects of reconstructed images. Despite the significant achievements of GANs in the field of image super-resolution, some challenges and areas for improvement remain. For instance, how to better utilize the feature information in images, improve the generalization ability and stability of the model, and more accurately restore local details of images are current research focuses. To address these issues, researchers have proposed various improvement strategies, such as introducing attention mechanisms to enhance the model's focus on important features, adjusting network structures to improve feature extraction efficiency, and adopting new regularization techniques to stabilize the training process.

In response to the issues of insufficient feature utilization and blurred local texture details in existing image super-resolution reconstruction methods, we propose a super-resolution reconstruction method based on local texture adversarial and hybrid attention generative adversarial networks. The hybrid attention module is added to the residual modules of the generator to enhance the model's utilization of features and the representativeness of each feature. To improve the quality of detail textures in generated images, we introduces LTSM, which can accurately locate patches with strong texture features based on the edge texture intensity of each local region in the input image, serving as a key reference for further processing. Along with the LTSM, a local discriminator is also employed, whose structure is identical to that of the global discriminator. Furthermore, the local discriminator

needs to make pixel-by-pixel judgments on the patches with the strongest texture information extracted by LTSM, making the local texture details of the reconstructed images more realistic and greatly enhancing the visual perception of the images.

The remainder of this paper is structured as follows: Section II provides a comprehensive review of related work, highlighting existing advancements and limitations in GAN-based and hybrid attention mechanisms for image super-resolution. Section III details the proposed method, including the generator and LTSM. Section IV presents extensive experimental results and analysis, comparing the proposed approach with state-of-the-art models on benchmark datasets. Finally, Section V concludes the paper, summarizing the contributions and discussing potential directions for future research.

II. Related Works

A. Single Image Super-Resolution Based on GANs

The goal of single SR is to enhance the resolution of a LR image to produce a corresponding HR image. Typically, LR images are obtained through a degradation process involving blurring and down-sampling. In classical image super-resolution reconstruction, bicubic down-sampling is widely used to simulate this degradation. By using it as a benchmark, different SR methods can be evaluated and directly compared, thereby verifying the effectiveness of new SR methods.

GANs [22] provide a principled approach to enhance the generator's ability to produce realistic images through adversarial training between the generator and discriminator. To improve perceptual quality, Johnson et al. [23] proposed a perceptual loss. Ledig et al. [10] introduced SRGAN, which performed adversarial training alongside the SRResNet generator, marking the first use of GANs for image super-resolution reconstruction. Subsequent improvements to generator architectures include Wang et al. [13], who proposed ESRGAN with a Residual-in-Residual Dense Block (RRDB) architecture, which has become a standard backbone for many state-of-the-art GAN-based super-resolution methods. Later, Rakotorinira et al. [24] enhanced ESRGAN with additional noise injection, presenting ESRGAN+ Zhang et al. [14] introduced a Ranker that learns perceptual metrics in RankSRGAN. For discriminator improvements, the relativistic discriminator concept proposed by RelativisticGAN [25] and the multi-discriminator strategy used by MPD-GAN [26] have provided greater training stability and better reconstruction image quality for GAN-based super-resolution methods.

Although the aforementioned GAN-based single image SR methods have made significant improvements in PSNR metrics and visual effects compared to interpolation- and regression-based methods, there is still considerable room for improvement in reconstructing local and highly textured regions of images.

B. Hybrid Attention Mechanism

In recent years, Transformer-based methods [27][28] have demonstrated remarkable performance in image restoration tasks, particularly in image SR and denoising. Despite these breakthroughs, attribution analysis reveals that existing networks have limited spatial utilization of input information. This indicates that the potential of transformers in current networks has not been fully exploited.

To better reconstruct images and activate more input pixels, Chen et al. proposed Hybrid Attention Transformer (HAT) [29]. HAT not only incorporates a channel attention mechanism to enhance the interaction efficiency between features but also introduces a window-based self-attention mechanism, further improving the model's ability to handle multi-scale features. This method effectively combines the global and local advantages of transformers, enhancing its performance in image restoration tasks. HAT can more meticulously focus on important features within the image, providing richer and more precise information for image restoration. However, Transformer-based SR reconstruction models often produce images with less realistic textures, whereas GANs can generate more visually appealing images. In terms of subjective visual effects, SR models based on GANs typically achieve better results. Additionally, GANs can combine various loss functions to adjust outputs, allowing them to perform well in different scenarios. Nevertheless existing models failed to effectively combined the advantages of transformer and GANs based approaches.

C. Existing Solutions and Limitations

Despite notable progress in image super-resolution (SR) research, existing methods still face critical limitations that hinder their effectiveness in addressing texture and detail reconstruction challenges. Table I summarizes the key limitations of prominent SR approaches and their unsuitability for the problem at hand.

The above limitations highlight the gaps in existing SR methods, particularly their inability to balance global structure preservation with detailed texture restoration. These shortcomings directly impact the visual realism and structural fidelity of reconstructed images. To address these challenges, the proposed approach introduces:

1) Hybrid Attention Residual Blocks (HARB): Combines window-based self-attention and channel attention to capture both global and local features, improving feature utilization and structural preservation.

2) Local Texture Sampling Module (LTSM): Targets high-frequency texture-rich regions for focused adversarial learning, enhancing detail realism and mitigating blurriness.

3) Dual adversarial branches: Integrates global and local discriminators to balance structural consistency and texture enhancement.

By addressing these gaps, the proposed method is particularly suited for generating high-resolution images with enhanced texture detail and structural fidelity, overcoming the limitations of existing approaches.

III. Proposed Method

The most distinctive feature of HR images is their intricate local texture patterns, which represent the distribution of local pixels. Specifically, high-frequency pixels concentrate around local edges, while low-frequency pixels smoothly spread adjacent to these edges. This separation pattern between high

TABLE I. LIMITATIONS IN EXISTING METHODS

Method	Advantages	Limitations	Suitability Issues for Current Problem
SRCNN[7]	Simple, pioneering CNN-based SR	Limited ability to capture high-frequency textures and complex details	Over-smooth outputs; insufficient texture restoration in high-resolution demands.
RCAN[8]	Channel attention for feature focus	High computational cost; limited enhancement of localized textures	Struggles with highly textured regions critical for detailed reconstructions.
SRGAN[9]	First GAN-based SR approach	Artifacts in outputs; struggles with preserving structural consistency	Insufficient detail fidelity in texture-rich and edge-dense areas.
ESRGAN[13]	Improved GAN design	Lacks mechanisms for enhancing localized texture details	Unable to accurately enhance localized, high-frequency textures.
SwinIR[27]	Transformer-based SR model	Effective global attention; high computational demand	Limited capability in generating realistic textures in local image regions.

and low-frequency elements starkly contrasts with LR images where high-frequency elements are either not distinctly separated or missing altogether. To address this, this study extends the framework of GANs by adding a patch-level learning branch. This branch adaptively applies adversarial learning to different local regions based on their edge characteristics, thereby enhancing the model's ability to capture texture patterns in HR images. Furthermore, to address the issue of insufficient feature utilization by existing models, we introduce a hybrid attention residual block in place of the original dense residual blocks within the generator. These hybrid attention blocks combine window-based multi-head self-attention mechanisms with channel attention mechanisms. This approach aims to activate more effective pixels for SR reconstruction tasks, thereby improving the utilization of features in input images. By integrating these advancements, the proposed method enhances the capability of GAN-based models to accurately reconstruct HR images while preserving and enhancing intricate local texture patterns.

A. Network Structure Overview

The network structure is depicted in Fig. 1, where I^{HR} represents the HR image;I^{LR} represents the LR image that obtained by bicubic interpolation and downsampling from;I^{SR} denotes the super-resolution image reconstructed by the generator. The high-resolution image are first input into the generator based on hybrid attention blocks to output super-resolution image, I^{HR} and I^{SR} are simultaneously input into the global discriminator, which outputs a grayscale image of the same height and width as the input image to determine whether the input image is a real high-resolution image or a super-resolution image generated by the generator, I^{HR} will also be sent into a pre-trained VGG-19 network, where the perceptual loss is calculated on the feature maps output from the middle convolutional layers of the network.

To better capture texture patterns that are more noticeable in local areas, a local adversarial learning branch is added. In this branch, LTSM is proposed, which constrains adversarial learning only in the local regions with the highest intensity. The LTSM takes mini-batches of I^{HR} and I^{SR} as input and outputs the top N patches I^{HR}_{patch} and I^{SR}_{patch} with the highest pixel intensities from these two mini-batches respectively. A local discriminator, which has the same structure as the global discriminator, simultaneously we established local discriminator to differentiate between the patches from I^{HR}_{patch} and I^{SR}_{patch} forming local adversarial learning, which has the same structure as the global discriminator. This promotes the generator to produce more realistic local texture details.

B. Generator

The existing model structure does not fully utilize the input features, leading to a loss of detail in the reconstructed images. we integrated Hybrid Attention Residual Blocks (HARB) into the Residual in Residual Dense Block(RRDB) structure. Specifically, an Hybrid Attention block(HAB) is embedded before the output of the RRDB module. The HAB combines channel attention and window-based multi-head self-attention in a parallel manner. Channel attention leverages global information, and self-attention has strong representation capabilities, ensuring that the network activates more effective pixels and extracts more input feature information. The structure of HARB is shown in Fig. 2. In the generator, only the last six RRDB blocks are replaced with HARB blocks.

The overall network structure of the generator is shown in Fig. 3. Since Batch Normalization (BN) layers can easily cause unwanted artifacts in SR reconstructed images, the entire generator structure does not use BN layers. All convolutional layers use LeakyReLU as the activation function, which addresses the zero-gradient issue for negative values, stabilizes model training, and accelerates model convergence. In the generator, a convolutional layer is first used to extract edge information from LR images, which is then fed into m RRDB blocks. The dense residual blocks and residual scaling techniques used in the RRDB blocks help train deeper network models, further improving the network's ability to capture semantic information. The intermediate feature maps produced by the RRDB blocks are then fed into n HARB blocks. These blocks use window-based multi-head self-attention to capture long-range dependencies in the sequence while focusing on important parts of the input feature information by paying attention to channel information. The upsampling part of the generator consists of two consecutive PixelShuffle [30], each of which doubles the resolution of the feature maps. Finally, two convolutional layers adjust the channels to output the SR reconstructed results.

C. Local Texture Sampling Module

In GANs, images reconstructed by the generator often exhibit blurriness and lack of detail. To improve the quality

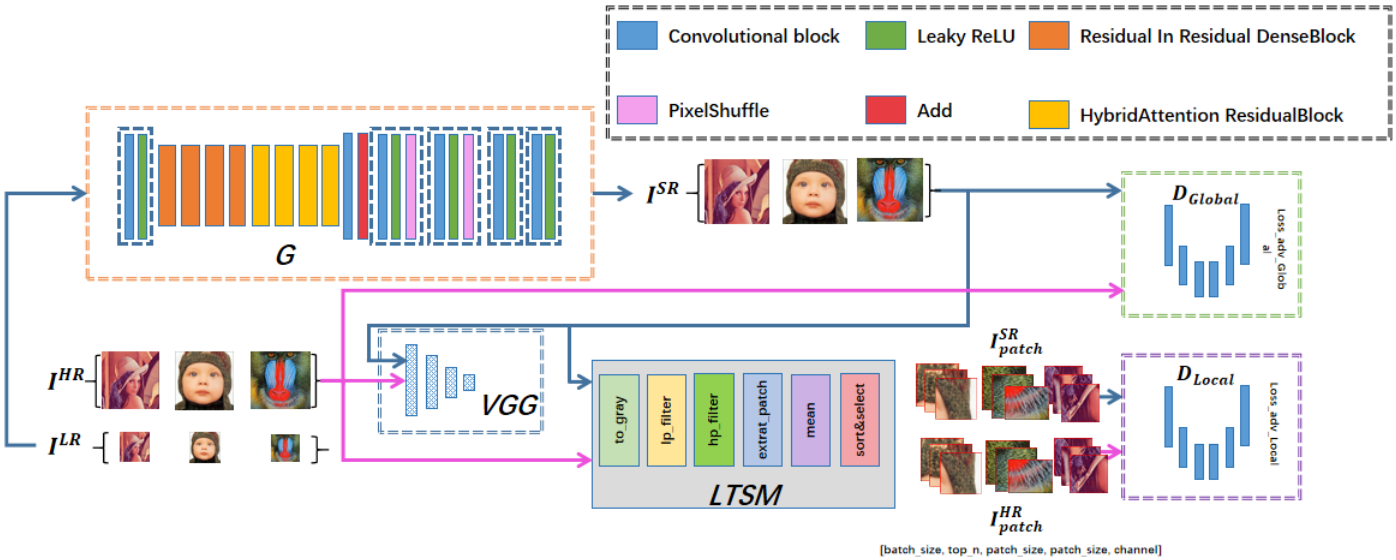

Fig. 1. The overall architecture of the proposed method, which consists of a global adversarial branch and a local adversarial branch. The LTSM is applied in the local branch to enhance the model's learning of texture details.

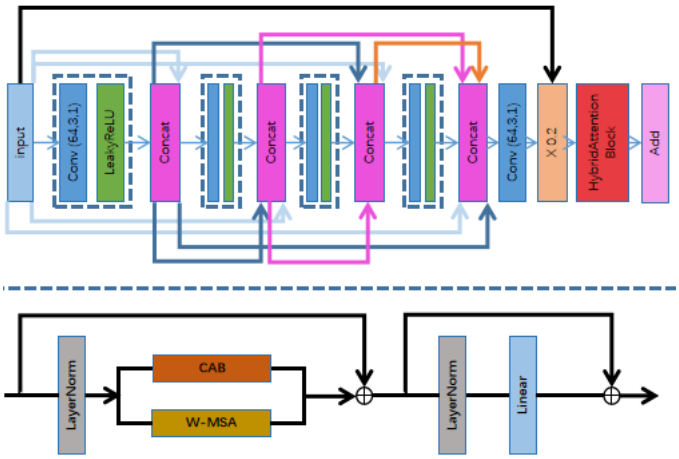

Fig. 2. The overall structure of the HARB is shown above the dashed line, consisting of Dense Residual Blocks and a Hybrid Attention Block (HAB). Below the dashed line is the structure of the HAB, which is composed of Channel Attention and Window-based Self-Attention mechanisms.

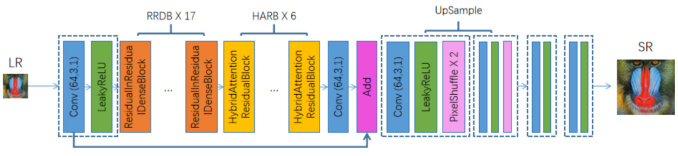

Fig. 3. The core of the generator consists of 17 RRDB and 6 HARB, after these blocks, an upsampling process using PixelShuffle is applied to increase the resolution of the image, followed by additional convolutional layers to produce the final HR image.

of local texture details in generated images, we proposes the LTSM. The LTSM is designed to extract local texture features from images. It uses an improved Sobel operator to calculate the edge strength of each local region in the input image and evaluates the texture features of these local regions based on their edge strength. The specific details of the LTSM are shown in Fig. 4. Specifically, in the preprocessing part of the LTSM, the input tensors I^{HR} and I^{SR} are first converted into ndarray format. Then, based on the hyperparameter $patchSize$, each group of images is divided into M patches I^{HR}_{patch} and I^{SR}_{patch}, here $M = Batch\_Size \times (\left|\frac{\mathrm{H}}{\mathrm{patchSize}}\right| + 1) \times (\left|\frac{\mathrm{W}}{\mathrm{patchSize}}\right| + 1)$ H and W represent the height and width of the image, respectively. At the same time, we also obtains a list of coordinates $patchCoordinates$ corresponding to the top-left corner of each patch in the original image. These segmented patches are processed through a guided-filter [31] $\mathcal{F}_{gf}$, which filters out noise from the image while retaining as much edge information as possible.

$$I_i^{patch} = \mathcal{F}_{gf}(I_i^{patch}, I_i^{patch}), i = 1, ..., Batch\_Size. \quad (1)$$

The denoised images are then fed into the improved Sobel operator, where the resulting four scores are squared and summed. Finally, the square root of the summed result is calculated, and the average value is taken to obtain the edge pixel intensity scores for all patches in an image. These scores serve as the keys for patch selection, and their values are calculated as follows:

$$\mathrm{Key_{j,k}} = \mathrm{Mean}(\sqrt[2]{\sum_{i=1}^{4}(I_{j,k}^{Patch} \otimes K_i)^2}), \quad (2)$$
$$\mathrm{j} = 1 \ldots Batch\_Size, \mathrm{k} = 1, \ldots, M$$

The symbol $\otimes$ represents the convolution operation. The Key values for all patches in a batch are calculated and sorted accordingly. Finally, the top N patches with the highest edge pixel intensity scores and their corresponding patch coordinates are obtained. Based on these coordinates, the correspond-

ing tensor patches are extracted from the original input tensors I^{SR} and I^{HR} preserving the original gradient information of the tensors. The architecture of LTSM is shown in Fig. 4:

D. Loss Function

1) Pixel-wise loss: traditional image super-resolution (SR) reconstruction methods are mostly based on the L2 pixel-level loss function mean-square error (MSE). Although this achieves a high PSNR value, using MSE tends to drive the solution towards a pixel-averaged result, which is overly smooth and perceptually poor. Therefore, in the pre-training phase, we only uses L1 loss to accelerate the convergence of the model. The pixel-wise loss is defined as shown in Eq. (3):

$$L_1 = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|G(I^{LR})_{(i,j)} - I^{HR}_{(i,j)}\right| \tag{3}$$

where G represents the generator.

2) Perception loss: we uses a pre-trained VGG-19 network to extract features. The perceptual loss is calculated using the feature maps before the LeakyReLU activation, as these feature maps contain more detailed information compared to the more sparse features after activation, providing stronger supervision. Features are extracted from the conv1-2, conv2-2, conv3-4, conv4-4, and conv5-4 layers, and the perceptual loss from each layer is weighted and summed to obtain the final perceptual loss. The perceptual loss is defined as shown in Eq. (4):

$$L_{percep} = \left\|\varphi(G(\mathrm{I^{LR}})) - \varphi(\mathrm{I^{HR}})\right\|_1 \tag{4}$$

Where $\varphi(\cdot)$ represents the pre-trained VGG-19 network.

3) Global adversarial loss: The global adversarial loss aims to capture global content feature information. The super-resolution images generated by the generator are input into the global discriminator to obtain a score for each pixel. Compared to the traditional VGG-style discriminator, which outputs a scalar for loss calculation, the discriminator is based on the idea of a U-Net style discriminator. The discriminator's loss is defined as the average decision of all pixels. Pixel-level loss calculation can make the texture details of the reconstructed image more precise. The least squares loss function (LSGAN) [32] is used instead of the cross-entropy loss function to achieve better training stability. The global adversarial loss function is defined as shown in Eq. (7):

$$\begin{aligned} L^D_{Global} = {} & E_{I^{HR}}[(D_{Global}(I^{HR}) - 1)^2] \\ & + E_{I^{LR}}[(D_{Global}(G(I^{LR})))^2] \end{aligned} \tag{5}$$

$$L^G_{Global} = E_{I^{LR}}[(D_{Global}(G(I^{LR})) - 1)^2] \tag{6}$$

$$L_{advGlobal} = L^D_{Global} + L^G_{Global} \tag{7}$$

4) Local adversarial loss: The local adversarial loss constrains adversarial training in small local regions with the highest edge texture intensity in the image, better promoting the generator to capture local texture features of high-resolution images. The output of the local discriminator is the average decision of all pixels in these small regions. The local adversarial loss is defined as shown in Eq. (10):

$$\begin{aligned} L^D_{Local} = {} & E_{h^p_i \sim I^{HR}_{patch}}\left[\frac{1}{N}\sum_{i=1}^{n}(D_{Local}(h^p_i) - 1)^2\right] + \\ & E_{l^p_i \sim I^{LR}_{patch}}\left[\frac{1}{N}\sum_{i=1}^{n}(D_{Local}(l^p_i))^2\right] \end{aligned} \tag{8}$$

$$L^G_{Local} = E_{l^p_i \sim I^{LR}_{patch}}\left[\frac{1}{N}\sum_{i=1}^{n}(D_{Local}(l^p_i) - 1)^2\right] \tag{9}$$

$$L_{advLocal} = L^D_{Local} + L^G_{Local} \tag{10}$$

Here h^p_i and l^p_i are the i-th small regions extracted by the LTSM from the high-resolution image I^{HR} and the super-resolution image I^{SR}. Since LTSM extracts the top N small regions with the highest edge texture intensity from each input image, the local adversarial loss is calculated by summing the loss over these N regions and then taking the average.

5) Pre-training and training loss function: The pre-training loss and training loss are based on the aforementioned loss functions. In the pre-training phase, only the generator is trained. The generator's pre-training loss is defined as shown in Eq. (11):

$$L_{pre} = L_1 \tag{11}$$

The training phase includes both generator loss and discriminator loss. The total loss function of the generator is defined as shown in Eq. (12):

$$L_G = \gamma_1 L_1 + \gamma_2 L^G_{Global} + \gamma_3 L^G_{Local} + \gamma_4 L_{percep} \tag{12}$$

where the weights of the generator's loss functions are $\gamma_1 = 0.08$,$\gamma_2 = 0.04$,$\gamma_3 = 0.02$,$\gamma_4 = 1$.

IV. Experimental Results and Analysis

A. Experimental Setup

The experiments were conducted on two NVIDIA GeForce RTX 3090 GPUs. The experiment used 800 HR images from the DIV2K dataset [33] and the corresponding LR images, obtained by bicubic interpolation with a scaling factor of 4, as the training dataset. The test sets are four standard datasets commonly used in the field of image super-resolution reconstruction: Set5, Set14, BSD100 and Urban100. The experiment used PSNR and SSIM as evaluation metrics. The settings and hyperparameter selection for the model during the training process are as follows: During training, the DIV2K dataset was randomly cropped into 128x128 images and subjected to random rotation and random flipping. The batch size for each input was 64. The number of RRDB blocks m was 16, and the number of HARB blocks n was 6. In the pre-training phase, only the PSNR-oriented pixel-wise loss defined in Eq. (3) was used to update the generator. The pre-training phase consisted of a total 6.25×10^4 iterations, with an initial learning rate of 2×10^{-4} The learning rate was halved after every 1.25×10^4 iterations.After the pre-training phase, the official training phase used Exponential Moving Average(EMA) to stabilize the training, with a weighting factor $\beta = 0.999$, In the official training phase, the initial learning rate for the generator was 1×10^{-4}, and the initial learning rate for the discriminator

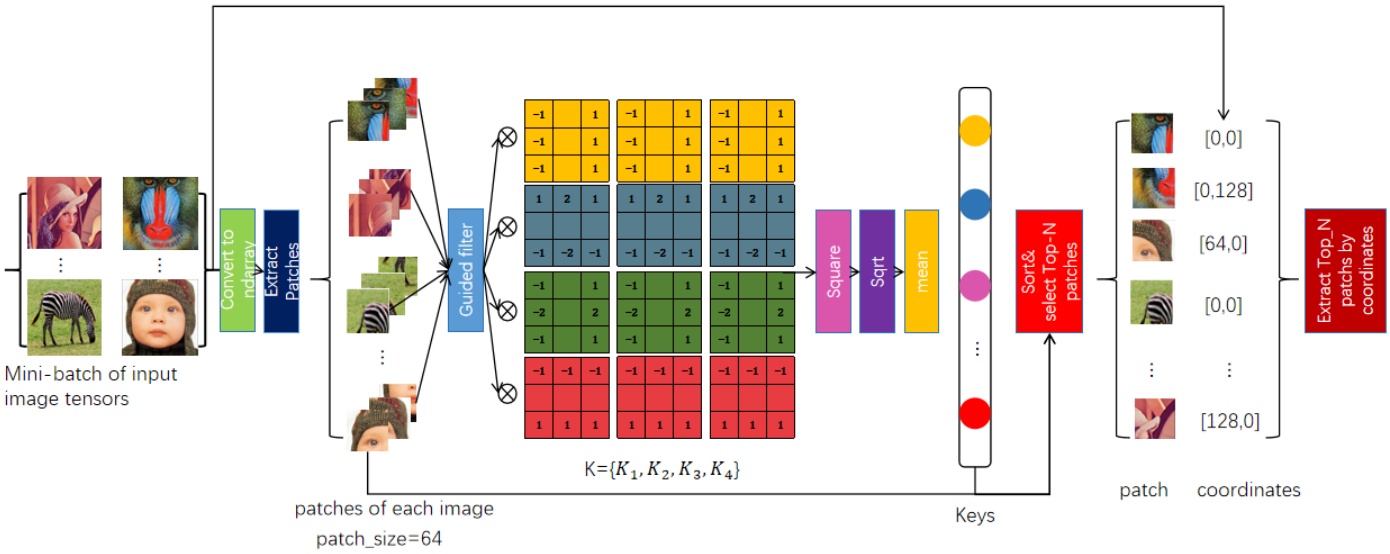

Fig. 4. Details of Local Texture Sampling Module (LTSM) which adaptively extracts local image patches with most salient texture features from each mini-batch of input images.

was 4×10^{-4}, The official training phase consisted of 7.5×10^{4} iterations. with the learning rates for both the generator and the discriminator halved after every 1.25×10^{4} iteration. Adam optimizer was used for all training phases, where $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1\times10^{-8}$.

B. The Effects of Hybrid Attention Residual Blocks

The generator is implemented based on Hybrid Attention Residual Blocks. To validate the effectiveness of these blocks in extracting more feature information and activating more effective pixels, this section conducts experiments on the pre-trained generator and compares the changes in PSNR values. As shown in Table II, increasing the number of Hybrid Attention Residual Blocks from 8 to 16 resulted in PSNR improvements of 0.22 dB and 0.13 dB on the Set5 and Set14 test sets, respectively. Further increasing the blocks from 16 to 24 resulted in a PSNR improvement of 0.05 dB on the Set5 test set and 0.01 dB on the Set14 test set. However, with more than 16 Hybrid Attention Residual Blocks, the generator's network parameters became excessively large. Therefore, we ultimately used 16 Hybrid Attention Residual Blocks to construct the generator, ensuring a high PSNR value while keeping the network parameter size manageable.

TABLE II. THE EFFECTS OF HYBRID ATTENTION RESIDUAL BLOCKS

The Number of HARB	SET5 PSNR(dB)	SET14 PSNR(dB)
0	30.12	27.8
8	31.11	28.1
16	31.93	28.23
32	31.98	28.24

C. The Effects of Local Adversarial Branch and LTSM

To verify the effectiveness of the LTSM and its impact on the generator, this section conducts quantitative and qualitative comparisons based on PSNR metrics and the quality of super-resolution reconstructed images. Specifically, we compare models using only the global adversarial module, models without LTSM extracting patches but using all patches, and the complete model.

As shown in Table III, introducing the local adversarial branch results in improvements in PSNR and SSIM metrics on each dataset, indicating that the local adversarial branch effectively enhances the structural similarity of images. Furthermore, not using the LTSM and training with all patches led to decreases in both PSNR and SSIM metrics, further validating the importance of the LTSM in extracting critical texture information. Comparing the reconstruction results shown in Fig. 5, we can visually observe differences in detail preservation and edge handling among different models. The complete model using the local adversarial branch and LTSM excels in restoring edge textures, producing clearer images with richer details. In contrast, models using only the global adversarial module show blurrier edge handling, and those not using the LTSM exhibit deficiencies in detail representation. This visual improvement vividly reflects the contribution of the local adversarial branch and LTSM in enhancing the effectiveness of super-resolution reconstruction. By comparing the reconstruction results under different model configurations, this study concludes that the local adversarial branch and LTSM are crucial for enhancing the performance of super-resolution reconstruction. They not only improve quantitative evaluation metrics but also demonstrate significant visual improvements in qualitative analysis. These findings underscore the importance of considering local texture features in the design of super-resolution reconstruction models.

D. Comparison with Existing SR Models

1) Quantitative comparison: In this section, our model is compared with several existing super-resolution (SR) models. The models chosen for comparison include traditional bicubic interpolation, as well as several deep learning-based methods

TABLE III. EFFECTS OF LOCAL ADVERSARIAL BRANCH AND LTSM ON PSNR AND SSIM METRICS

Datasets	W/O Local Adversarial Branch		W/O LTSM(All patches)		ours(full model)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set5	32.30	0.9073	31.22	0.8987	**32.32**	**0.9110**
Set14	28.12	0.8025	27.92	0.7829	**28.15**	**0.8231**
BSD100	27.92	0.6882	27.12	0.6801	**28.31**	**0.7012**
Urban100	26.66	0.8029	26.85	0.8012	**27.18**	**0.8206**

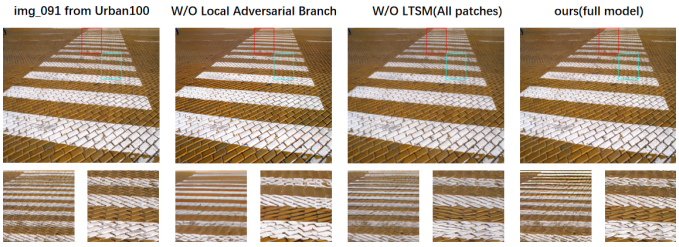

Fig. 5. Urban100 dataset img_091 Reconstruction Comparison, the results show that compared to model trained without the local adversarial branch and trained using all patches in the local branch, the model trained with the LTSM produces more realistic texture details.

such as SRCNN [7], RCAN [9], SRGAN [10], ESRGAN [13], and SwinIR [27]. Additionally, we incorporate the recently proposed Semantic-aware Discriminator (SeD) [34] into ESRGAN and SwinIR, a recent approach designed to enhance texture generation quality by leveraging semantic information. The improved versions of these models are denoted as ESRGAN+ and SwinIR+, respectively. Comparative experiments are conducted on four commonly used benchmark datasets: Set5 [35], Set14 [36], BSD100 [37], and Urban100 [38], with primary evaluation metrics being PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index).

In Table IV, our model demonstrates best performance across four commonly used datasets at all scales, especially achieves average improvement of 0.15 dB on PSNR compared to SwinIR+ [27][34] at $\times 4$ scale. From the SSIM results, it is evident that our model consistently achieves optimal performance across most datasets except for SSIM on BSD100($\times 4$). Compared to other models, our model shows an average SSIM improvement of 0.053 over SRCNN [7], 0.014 over RCAN [9], 0.047 over SRGAN [10], 0.056 over ESRGAN+ [13][34], and 0.003 over SwinIR+ [27][34]. These findings indicate that our model not only excels in image clarity (PSNR) but also performs exceptionally well in preserving image structure and details (SSIM).

In summary, through comparisons with various existing SR models, our proposed Generative Adversarial Network super-resolution reconstruction method based on local texture adversarial learning and hybrid attention demonstrates outstanding performance in both PSNR and SSIM metrics. This validates its effectiveness and superiority across different types of images.

2) Qualitative comparison: In terms of qualitative comparison, this study selected typical images from different datasets to visually assess the reconstruction results of various models. The specific results are shown in Fig. 6, 7 and 8.

Regarding image details and texture restoration, the proposed model demonstrates significant advantages Compared to other models, it preserves the details and textures of the original images better during reconstruction, the patches highlighted in red boxes represent critical regions for evaluating detail preservation and texture fidelity. For instance, in urban street scene images Img_014 and Img_087 from the Urban100 dataset, the proposed model not only reconstructs building edges and textures clearly but also presents more natural and realistic details. In contrast, other models like SRCNN [7], RCAN [9], and SRGAN [10] may exhibit blurriness or distortion in some details, which the proposed model effectively avoids.

Furthermore, on datasets like Set5 and Set14, the proposed model shows strong robustness and capability in restoring details in natural scenes and facial images. In Baboo from Set14, the proposed model performs a more natural reconstruction effect on facial details such as eyes and beard.

In comparison with existing SR models, the proposed model demonstrates superior performance in both quantitative metrics and qualitative effects. By integrating local texture adversarial learning and hybrid attention mechanisms, the proposed model not only enhances the accuracy of image reconstruction but also achieves a higher level of visual fidelity.

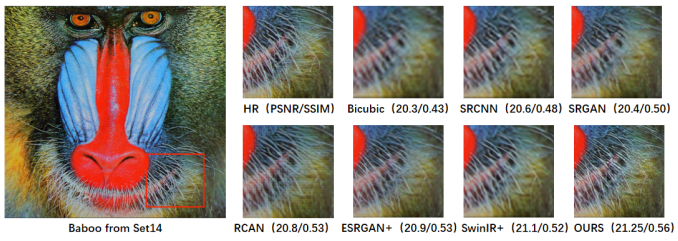

Fig. 6. Baboo from Set14 Reconstruction Comparison, the patches for comparison are marked with red boxes in the original images. PSNR/SSIM is calculated based on the patches to better reflect the performance difference.

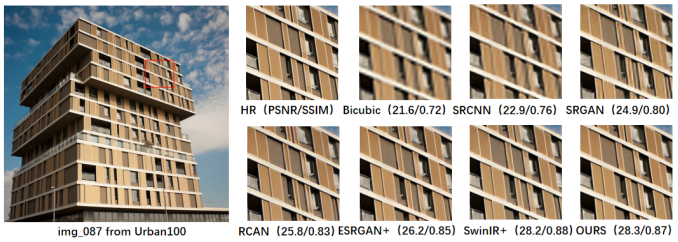

Fig. 7. Img_087 from Urban100 reconstruction comparison.

TABLE IV. QUANTITATIVE COMPARISONS (PSNR/SSIM) BEST PERFORMANCES ARE MARKED IN BOLD AND "+" INDICATES THAT METHODS INCORPORATE SED

Method	Scale	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	x2	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403
SRCNN[7]	x2	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
RCAN[9]	x2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384
SRGAN[10]	x2	36.86	0.9560	33.56	0.9156	32.12	0.8996	32.36	0.9196
ESRGAN+[13][34]	x2	37.45	0.9592	33.76	0.9175	32.30	**0.9075**	32.56	0.9315
SwinIR+[27][34]	x2	38.39	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9393
Ours	x2	**38.41**	**0.9652**	**34.22**	**0.9231**	**32.55**	0.9038	**33.47**	**0.9425**
Bicubic	x3	30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349
SRCNN[7]	x3	32.75	0.9090	29.28	0.8209	28.41	0.7863	26.24	0.7989
RCAN[9]	x3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702
SRGAN[10]	x3	33.20	0.9101	29.89	0.8322	29.13	0.7850	28.91	0.8577
ESRGAN+[13][34]	x3	34.75	0.9223	30.44	0.8455	29.30	0.8128	28.99	0.8679
SwinIR+[27][34]	x3	34.89	0.9312	30.89	0.8503	29.35	0.8124	29.29	0.8744
Ours	x3	**34.93**	**0.9388**	**30.90**	**0.8521**	**29.40**	**0.8155**	**29.47**	**0.8782**
Bicubic	x4	28.40	0.7854	26.09	0.7486	24.98	0.6935	23.12	0.6577
SRCNN[7]	x4	29.07	0.8504	26.64	0.7602	26.90	0.7101	23.98	0.7213
RCAN[9]	x4	30.83	0.8878	26.75	0.7889	27.77	0.7236	25.92	0.7985
SRGAN[10]	x4	29.40	0.8213	26.21	0.7428	27.1	0.7223	24.37	0.7802
ESRGAN+[13][34]	x4	30.46	0.8525	26.86	0.7905	27.85	0.6528	26.15	0.7328
SwinIR+[27][34]	x4	32.25	0.9012	28.12	0.7914	28.29	**0.7311**	26.71	0.8164
Ours	x4	**32.32**	**0.9110**	**28.15**	**0.8231**	**28.31**	0.7012	**27.18**	**0.8206**

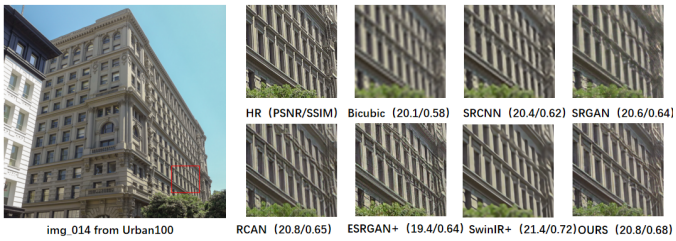

Fig. 8. Img_014 from Urban100 reconstruction comparison.

V. CONCLUSION

We had proposed a GAN based method for image SR reconstruction, leveraging local texture adversarial and hybrid attention residual block. LTSM is introduced to compute edge intensity in local image regions, this module effectively addresses issues of blurriness and detail loss commonly observed in traditional GAN-generated images. Additionally, a generator equipped with HARB is incorporated to enhance the utilization of input features during generation. This approach ensures better preservation of image structure and details, thereby improving overall image quality in reconstruction. Experimental validation on multiple standard datasets (Set5, Set14, BSD100, and Urban100) demonstrates superior performance in terms of PSNR and SSIM metrics, surpassing various existing super-resolution methods and exhibiting notable advantages in image detail and texture restoration. However it is not without limitations. One key limitation is the computational cost associated with the HARB and LTSM, which may limit its deployment in real-time applications or on devices with constrained resources. Additionally, while the LTSM effectively enhances local texture details, its reliance on patch selection based on edge intensity might overlook non-edge regions with critical texture information, leading to potential gaps in detail preservation in less textured areas.

For future work, we will explore optimizing the computational efficiency of the proposed framework by leveraging lightweight architectures or pruning techniques. Furthermore, incorporating adaptive mechanisms for selecting texture-rich regions, beyond edge intensity, could enhance the model's ability to generalize across diverse image types. Extending the current method to handle multi-frame super-resolution or domain-specific applications, such as medical imaging or satellite imagery, could also provide new directions for further exploration.

ACKNOWLEDGMENT

This work was supported by the Key Research and Development Program of Zhejiang Province (2020C03094), and the General Scientific Research Project of the Department of Education of Zhejiang Province (Y202250677, Y202250706, Y202250679).

REFERENCES

[1] Z. Cui, Y. Zhu, L. Gu, G. J. Qi, X. Li, R. Zhang, Z. Zhang, and T. Harada, "Exploring resolution and degradation clues as self-supervised signal for low quality object detection," 2022.

[2] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" 2015.

[3] L. Zhou, G. Chen, M. Feng, and A. Knoll, "Improving low-resolution image classification by super-resolution with enhancing high-frequency content," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021.

[4] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] A. Singh and J. Singh, "Content adaptive single image interpolation based super resolution of compressed images," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, p. 3014, 2020.

[6] X. Ma, J. Zhang, T. Li, L. Hao, and H. Duan, "Super-resolution geomagnetic reference map reconstruction based on dictionary learning and sparse representation," *IEEE Access*, vol. 8, pp. 84 316–84 325, 2020.

[7] 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.

[8] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[9] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[14] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Super resolution generative adversarial networks with learning to rank," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7149–7166, 2021.

[15] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[16] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 021–10 030.

[17] H. Phung, Q. Dao, and A. Tran, "Wavelet diffusion models are fast and scalable image generators," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 199–10 208.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[19] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[20] Y. Kim and D. Son, "Noise conditional flow model for learning the super-resolution space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 424–432.

[21] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Learning the super-resolution space with normalizing flow," *ECCV, Srflow*, vol. 2, 2020.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[23] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[24] N. C. Rakotonirina and A. Rasoanaivo, "Esrgan+: Further improving enhanced super-resolution generative adversarial network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3637–3641.

[25] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.

[26] O.-Y. Lee, Y.-H. Shin, and J.-O. Kim, "Multi-perspective discriminators-based generative adversarial network for image super resolution," *IEEE Access*, vol. 7, pp. 136 496–136 510, 2019.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[28] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 12 312–12 321.

[29] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong, "Hat: Hybrid attention transformer for image restoration," *arXiv preprint arXiv:2309.05239*, 2023.

[30] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[31] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[32] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[33] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 126–135.

[34] B. Li, X. Li, H. Zhu, Y. Jin, R. Feng, Z. Zhang, and Z. Chen, "Sed: Semantic-aware discriminator for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 784–25 795.

[35] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[36] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*. Springer, 2012, pp. 711–730.

[37] 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*, vol. 2. IEEE, 2001, pp. 416–423.

[38] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.