StarSRGAN: Improving Real-World Blind Super-Resolution

Khoa D. Vo

Faculty of Information Technology (FIT)
University of Science, VNU.HCM
Ho Chi Minh City, Vietnam
20c11008@student.hcmus.edu.vn

Len T. Bui

Faculty of Information Technology (FIT)
University of Science, VNU.HCM
Ho Chi Minh City, Vietnam
btlen@fit.hcmus.edu.vn

Figure 1: Comparison between our models and some standard SRGAN models. (**Zoom in for best view**)

ABSTRACT

The aim of blind super-resolution (SR) in computer vision is to improve the resolution of an image without prior knowledge of the degradation process that caused the image to be low-resolution. The State of the Art (SOTA) model Real-ESRGAN has advanced perceptual loss and produced visually compelling outcomes using more complex degradation models to simulate real-world degradations. However, there is still room to improve the super-resolved quality of Real-ESRGAN by implementing recent techniques. This research paper introduces StarSR-GAN, a novel GAN model designed for blind super-resolution tasks that utilize 5 various architectures. Our model provides new SOTA performance with roughly 10% better on the MANIQA and AHIQ measures, as demonstrated by experimental comparisons with Real-ESRGAN. In addition, as a compact version, StarSRGAN Lite provides approximately 7.5 times faster reconstruction speed (real-time upsampling from 540p to 4K) but can still keep nearly 90% of image quality, thereby facilitating the development of a real-time SR experience for future research. Our codes are released at https://github.com/kynthesis/StarSRGAN.

Keywords

Blind super-resolution, adaptive degradation, dual perceptual loss, multi-scale discriminator, dropout degradation

1 INTRODUCTION

The field of image super-resolution (SR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) counterpart, which is a challenging task due to the many-to-one mapping involved. The computer vision research community has devoted significant attention to SR, and deep learning algorithms have

been successfully applied to this problem. These techniques utilize neural networks to train an end-to-end mapping function, such as the SRCNN [Don16], which involves deep convolutional neural networks (CNNs) that generate high signal-to-noise ratio (PSNR) values. Still, the output is often excessively smoothed and needs more high-frequency features.

Researchers suggest using generative adversarial networks (GANs) [Goo14] for image SR tasks to overcome these limitations. A super-resolution GAN comprises a generator network and a discriminator network. The generator takes LR images as input and aims to create images that resemble the original HR image. At the same time, the discriminator distinguishes between "fake" and "real" HR images. However, these methods

assume an ideal bicubic downsampling kernel, making them unsuitable for real-world situations.

In contrast, blind SR aims to recover images with unknown and complicated degradations. Existing techniques can be classified as explicit or implicit modelling based on the underlying degradation process. Detailed modelling approaches frequently utilize the conventional degradation model, including blur, downsampling, noise, and JPEG compression. However, real-world degradations need to be simplified to be described by a simple combination of these factors. Current research has focused on emulating a more practical degradation process [Wan21] or designing an improved generator [Wan19]. The performance of the discriminator, which guides the generator to generate superior images, has received relatively little attention. Therefore, it is essential to acknowledge the importance of the discriminator, much like a loss function.

In this paper, we intend to further enhance the perceptual quality of super-resolved images by extending the robust Real-ESRGAN [Wan21] algorithm:

- We utilize the Star Residual-in-Residual Dense Block (StarRRDB), which was inspired by ESRGAN+ [Rak20] and had a higher capacity than the RRDB employed by Real-ESRGAN [Wan21].

- We obtain a performance breakthrough for our SR models by combining the Multi-scale Attention U-Net Discriminator with the present StarRRDB-based generator inspired by A-ESRGAN [Wei21].

- We replaced the standard high-order Real-ESRGAN degradation model with a DASR-inspired [Lia22] efficient Adaptive Degradation Model.

- We use ResNet Loss in addition to VGG Loss, and this Dual Perceptual Loss approach, inspired by ESRGAN-DP [Son22], acquires more sophisticated perceptual characteristics.

- We attempt to apply Dropout Degradation technique, inspired by RDSR [Kon21] to improve the generalization ability from appropriate usage of dropout benefits.

- We construct StarSRGAN Lite, a CNN-oriented compact version that can reconstruct images about 7.5 times faster than StarSRGAN and Real-ESRGAN.

Due to several modifications, our StarSRGAN achieves higher visual performance than Real-ESRGAN, making it more applicable to real-world applications.

2 RELATED WORK

2.1 Super-Resolution Methods

In blind image SR problems, deep CNNs are frequently used before implementing the GAN architecture. These techniques have achieved a remarkable peak signal-to-noise ratio (PSNR) due to the robust modelling capability of CNNs. However, since these PSNR-based approaches use pixel-wise specified losses such as mean squared error (MSE), the output is often overly smoothed, necessitating additional high-frequency information. In practice, most methods assume a bicubic downsampling kernel, which may not be effective for real-world images. Furthermore, current studies aim to incorporate reinforcement learning or GAN before image restoration.

Blind SR has been widely researched, with numerous studies focusing on degradation prediction and conditional restoration. The two processes can be performed separately or in tandem, often iteratively. These techniques rely on predefined degradation models, which may only consider synthetic degradations and fail to perform well with real-world images. Furthermore, inaccurate degradation models can result in unwanted artifacts in the reconstructed images.

Recently, SOTA research has proposed a perceptually-driven approach to improve GANs by more accurately simulating the perceptual loss between images. For instance, ESRGAN [Wan19] and ESRGAN+ [Rak20] have introduced a viable perceptual loss function and generator networks based on RRDB that can convincingly create HR images. Another method, Real-ESRGAN [Wan21], has introduced a high-order degradation model to make even more realistic images, achieving impressive results on the NIQE benchmark. However, these methods depend on a computationally complex backbone network and cannot handle images with varying levels of degradation. Therefore, DASR [Lia22] has introduced a degradation-adaptive framework to address this issue, creating an effective and efficient network for real-world SR challenges.

Our work has incorporated several benefits from various designs to produce a comprehensive solution.

2.2 Degradation Models

Blind SR approaches often rely on the classical degradation model, which may not fully represent the complex degradation in real-world images. Recent technique, such as Real-ESRGAN [Wan21], incorporate a broader range of degradation types and parameters into the modelling process to address this issue. These approaches increase the model's ability to enhance the perceptual quality of challenging LR inputs. However, the sampling of degradation parameters in these methods can be imbalanced, which limits their ability to gen-
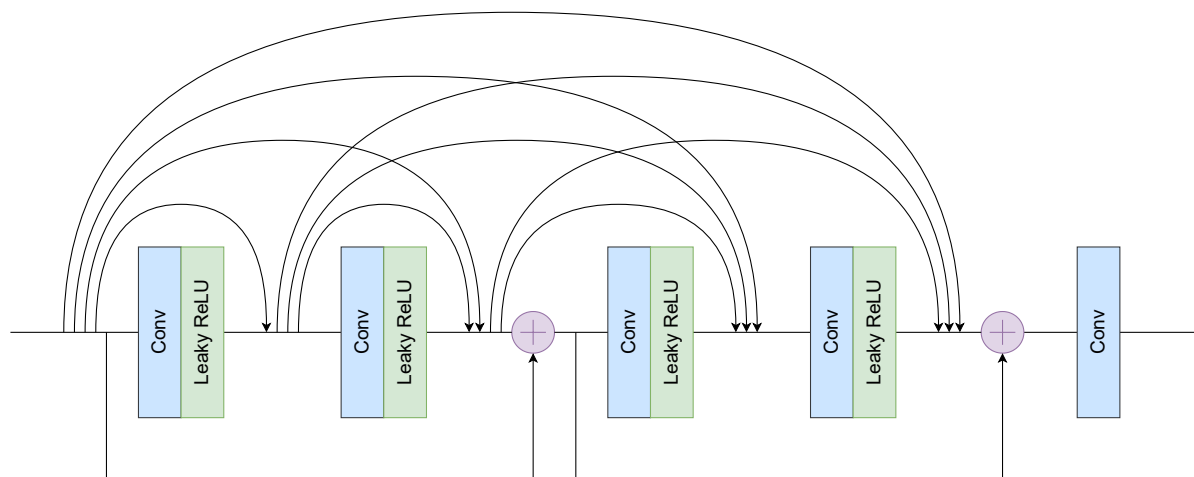
Figure 2: In Star Dense Block, residuals are added every two layers.

erate fine features, particularly for inputs with average degradations.

This study implements an adaptive degradation model that balances the degradation space by dividing it into three frequency-balanced levels. This balanced space optimises the model at different levels and provides a more accurate representation of real-world images.

2.3 Discriminator Models

There have been significant efforts to improve the discriminator model to synthesize high-quality HR images. Two critical challenges in achieving photorealistic HR images are the need for a broad receptive field for the discriminator to distinguish between the synthesized and ground truth images, which requires either a deep neural network or a big convolution kernel. Additionally, a single discriminator may struggle to provide input on global and local characteristics, leading to potential incoherence in the synthesized image, such as distorted wall textures.

pix2pixHD [Wan17] proposed a unique multiple discriminator architecture to address these challenges. The first discriminator takes downsampled synthetic images as input, has a broader receptive field with fewer parameters, and focuses on comprehending the global perspective. The second discriminator uses the entire synthesized image to learn the image's specifics. Another study [Sch20] employed a U-Net-based discriminator architecture to GAN-based blind SR challenges, preserving the global coherence of synthesized images while offering per-pixel input to the generator.

Our discriminator model combines the advantages of both designs, enabling the discriminator to learn edge representations, enhance training stability, and provide per-pixel feedback to the generator.

2.4 Image Quality Assessment Methods

In recent years, GAN has been widely utilized for restoring low-quality images (e.g., deblur, denoise, super-resolution). Some researchers have focused on assessing images using Image Quality Assessment (IQA) methods. Some synthetic textures look natural due to the GAN approach, making them difficult for humans to distinguish yet easy for machines to detect.

Attention-based Hybrid Image Quality Assessment Network (AHIQ) [Lao22] aims to quantify the human perception of image quality and has the potential for generalizing unknown and complex samples, notably GAN-based distortions. AHIQ won first place in Full-Reference (FR) Track for the NTIRE2022 Perceptual Image Quality Assessment Challenge [Nti22].

Multi-dimension Attention Network for No-Reference Image Quality Assessment (MANIQA) [Yan22] uses the multi-dimensional attention network for perceptual assessment. MANIQA placed first in the No-Reference (NR) Track of the NTIRE2022 Perceptual Image Quality Assessment Challenge [Nti22].

Since no real-world GT exists for SR images, NR-IQA is preferable to FR-IQA for SR visual comparison. Nonetheless, in this study, we use both AHIQ and MANIQA, as well as some classical metrics like PSNR, SSIM, NIQE, and LPIPS. The objective is to determine if StarSRGAN models can achieve new performance levels in many measures.

2.5 Perceptual Loss Methods

Since the groundbreaking SRCNN [Don16] was presented, applying deep learning to tackle the SR problem has garnered increasing interest. In addition to a significant boost in visual quality, it also features a greater variety of optimizations and enhancements.

Further research [Joh16] claimed that smoothing the reconstructed image was as simple as improving the MSE

or PSNR of the pixel space ratio between the GT image and the reconstructed image. In order to enhance the reconstruction effect, the perceptual loss was suggested to minimize the feature space error between the GT image and the rebuilt image.

Based on the concept of SRGAN [Led17], ESRGAN [Wan19] employed a range of strategies to enhance the texture features further. In terms of perceptual loss, they recommended using the output before the activation of the convolutional layer to gain additional feature information, with the error of the feature space before activation being the object to be reduced.

Discussing perceptual loss is essential for enhancing the reconstruction outcomes, particularly the realism of details. This work uses a unique dual perceptual loss function as a combination of ResNet [Kai16] loss and VGG [Sim15] loss to achieve the reduction of unnatural artifacts produced by the perceptual-driven technique.

3 PROPOSED METHODS

3.1 Adaptive Degradation Model

Recently, a high-order degradation model has been proposed to generate LR images more closely approximate real-world conditions. The model executes the same degradation operation multiple times and has advanced from simple bicubic down-sampling to include shuffling and second-order pipelines.

In this research, we incorporate numerous image degradation procedures, including blurring (both isotropic and anisotropic Gaussian blur), resizing (both down-sampling and up-sampling with area, bilinear, and bicubic operations), noise corruption (both additive Gaussian and Poisson noise), and JPEG compression.

Inspired by the DASR [Lia22] approach, our architecture is designed to be adaptive to a wide range of real-world inputs and handle a subspace of images with different degradation levels. We divide the entire degradation space D into three levels: $[D_1, D_2, D_3]$, with one of these randomly chosen to produce LR-HR image pairs during training. The probability distribution for selecting the levels is [0.3, 0.3, 0.4]. D_1 and D_2 use first-order degradation with small and large parameter ranges, while D_3 uses second-order degradation. We use isotropic and anisotropic Gaussian kernels for the blur operation with a probability of [0.65, 0.35]. If an isotropic blur kernel is supplied, we set $\sigma_1 = \sigma_2$. In the second degradation stage of D_3, we skip the blur operation with a 20% probability and use sinc kernel filtering with an 80% probability, following the approach used in Real-ESRGAN. We scale the image to the appropriate LR size, a quarter of its original size.

3.2 Network Architecture

StarSRGAN. The core block of ESRGAN enables the network to be very scalable and more straightforward

to train. The Star Dense Block we suggested is replacing the Dense Block to increase the network's capacity. Figure 2 depicts an extra level of residual learning within the Dense Blocks to expand the capacity without increasing complexity. After two layers, a residual is added to each block. This novel architecture produces images with improved perceptual quality by utilizing feature exploitation and exploration.

StarSRGAN Lite. The lightweight version of StarSRGAN focuses on delivering faster reconstruction times while maintaining acceptable visual quality. Like its predecessor, the model seeks to capitalize on the Multi-scale Discriminator, Attention U-Net Discriminator, Dual Perceptual Loss, Dropout Degradation, and Adaptive Degradation. ESPCN [Shi16] influences the network design, a super-resolution CNN which brings asymptotic real-time performance. Figure 3 depicts the architecture of StarSRGAN Lite.
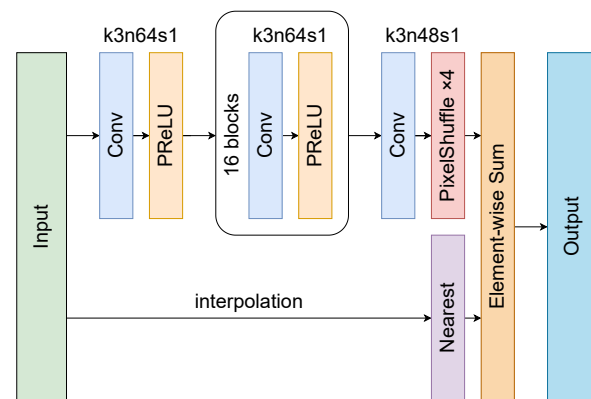

Figure 3: Architecture of StarSRGAN Lite with corresponding kernel size (k), number of filters (n), and stride (s) indicated for each convolutional layer.

3.3 Attention U-Net Discriminator

Taking inspiration from A-ESRGAN [Wei21], we have developed an Attention U-Net Discriminator structure, depicted in Figure 4, that aims to enhance the quality of the reconstructed image while increasing the efficiency of image reconstruction. The structure comprises a downsampling encoding module, an upsampling decoding module, and multiple Attention Blocks. The Attention and Concatenation Blocks are designed following the A-ESRGAN architecture. To perform semantic segmentation of 2D images, we adapted the Attention Gate, initially scheduled for 3D medical images as described in [Okt18]. Furthermore, we incorporated Spectral Normalization regularization [Miy18] to stabilize the training process.

3.4 Multi-scale Discriminator

StarSRGAN has a Multi-scale Discriminator architecture consisting of two identical U-Net discriminators.
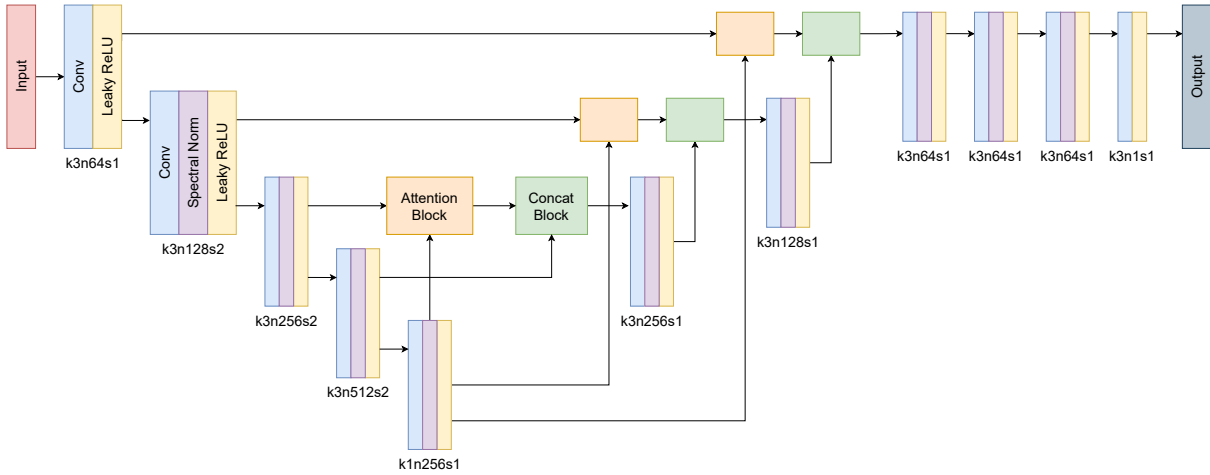
Figure 4: Architecture of Attention U-Net Discriminator with corresponding kernel size (k), number of feature maps (n), and stride (s) indicated for each convolutional layer.

The first discriminator D_1 accepts an original scale image as input, and the second discriminator D_2 accepts a $2\times$ downsampled image as input.

The output of the U-Net discriminator is a $W \times H$ matrix, with each member representing the probability that the pixel it represents is True. We utilize the Sigmoid function to normalize the output and the binary cross-entropy loss to determine the overall loss of one discriminator. Assuming C is the output matrix, we define $D = \sigma(C)$, x_r is 'real' data, and x_f is 'fake' data.

Consequently, we define the loss of one discriminator as

$$L_D = \sum_{w=1}^{W} \sum_{h=1}^{H} (-E_{x_r}[\log(D_{(x_r,x_f)}[w,h])] \\ - E_{x_f}[1 - \log(D(x_f,x_r)[w,h])]) \qquad (1)$$

Because we have multi-scale discriminators, we will sum the loss of all the discriminators to get the overall loss as

$$L_{Total} = \lambda_1 L_{D_{normal}} + \lambda_2 L_{D_{sampled}} \qquad (2)$$

where λ_1 and λ_2 are coefficients. We can also derive the generator loss from a single discriminator as

$$L_G = \sum_{w=1}^{W} \sum_{h=1}^{H} (-E_{x_r}[1 - \log(D_{(x_r,x_f)}[w,h])] \\ - E_{x_f}[\log(D(x_f,x_r)[w,h])]) \qquad (3)$$

where x_f represents the output of the generator $G(x_i)$.

3.5 Dual Perceptual Loss

By training a deep neural network, we address the SR problem. According to the theory given by Dong et al. [Don16], the following is the goal of optimization:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} L\left(G\left(I_i^{LR};\theta\right), I_i^{HR}\right), \qquad (4)$$

Where $I_i^{LR} \in \mathbf{R}^{C \times H \times W}$ and $I_i^{HR} \in \mathbf{R}^{C \times H \times W}$ represent the $i\text{-}th$ LR and HR sub-image pairings, respectively, in the training set. $G\left(I_i^{LR};\theta\right)$ is the representation of the up-sampling network. θ is the parameter to be optimized within the neural network. L is the loss function that can be represented as:

$$L = \lambda l_{content} + \eta l_{adversarial} + \gamma l_{percep} \qquad (5)$$

where $l_{content}$ is the content loss of the pixel-wise 1-norm distance between images reconstructed by the generator and GT images, $l_{adversarial}$ is the loss derived from the mentioned Multi-scale Discriminator, l_{percep} is the Dual Perceptual Loss we implement. λ, η, and γ are the coefficients of balancing different loss terms.

SRGAN [Led17] proposed defining the VGG loss based on the ReLU activation layer of the pre-trained VGG-19 network. ESRGAN [Wan19] redefined the VGG loss after the convolutional layer and before the activation layer to gain additional feature information. This study uses the VGG loss specified in [Wan19], as the L1 Norm function is utilized to determine the Manhattan Distance between the reconstructed image features and the GT image features:

$$l_{VGG/i,j} = \frac{1}{C_{i,j}W_{i,j}H_{i,j}} \sum_{z=1}^{C_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \\ \left| \Phi_{i,j}\left(G_{\theta_G}\left(I^{LR}\right)\right)_{x,y,z} - \Phi_{i,j}\left(I^{HR}\right)_{x,y,z} \right| \qquad (6)$$

where $\Phi_{i,j}$ represents features acquired by the $j\text{-}th$ convolution (before activation) in the VGG network before the $i\text{-}th$ max-pooling layer. In the VGG network, $C_{i,j}$, $W_{i,j}$, and $H_{i,j}$ are the dimensions of their respective feature spaces.

Based on the concepts of SRGAN [Led17], the ResNet loss is defined based on the ReLU activation layer of

Table 1: Quantitative comparison with SOTA methods on common test datasets using standard IQAs for SR (4×). LPIPS/NIQE ↓: the lower, the better. PSNR/SSIM/MANIQA/AHIQ ↑: the higher, the better. Note that MANIQA and AHIQ are the current SOTA IQA methods. The best and second performance are marked in **red** and blue. **Discussion**: The unusual city textures on Urban100 are generally intricate and should consider further fine-tuning for particular usage. Models using real-world emulated data synthesis failed on Manga109 (Japanese comic) dataset, as prior predictions. Thanks to the Adaptive Degradation Model have helped StarSRGAN solve this problem partially. StarSRGAN has a slightly lower performance on Set14, which contains many drawing pictures.

Method	DIV2K						Set5						Set14					
	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ
Original HR	80.00	1.0000	0.0000	3.2342	0.6542	0.6289	80.00	1.0000	0.0000	3.2122	0.6338	0.6353	80.00	1.0000	0.0000	5.7318	0.5807	0.6402
SRGAN	25.71	0.7076	0.1808	3.0079	0.5256	0.4847	30.64	0.8361	0.1080	4.3067	0.5084	0.5359	19.94	0.4688	0.2566	2.5155	0.5106	0.4865
ESRGAN	24.94	0.6827	0.1471	3.1900	0.5616	0.4938	29.92	0.8258	0.1117	3.9433	0.5396	0.5461	18.60	0.4209	0.2039	2.3251	0.5329	0.5203
Real-ESRGAN	24.23	0.6646	0.2637	3.1461	0.5728	0.5112	26.06	0.7806	0.1901	3.6883	0.5794	0.5524	18.00	0.4186	0.3578	3.6056	0.5979	0.5233
A-ESRGAN	23.12	0.6032	0.3110	2.9722	0.5398	0.4217	23.95	0.6980	0.2060	2.8405	0.6046	0.4730	18.12	0.3476	0.4927	2.7213	0.6017	0.5245
FeMaSR	22.10	0.6178	0.2236	4.9289	0.5383	0.4426	24.22	0.7202	0.1943	4.4306	0.6101	0.5054	17.49	0.3872	0.3314	3.6338	0.5884	0.4968
Swin2SR	25.65	0.7164	0.4282	5.8883	0.4346	0.4319	28.53	0.8629	0.2855	6.9146	0.5123	0.5681	19.94	0.4756	0.2556	2.6782	0.5689	0.5307
StarSRGAN	25.53	0.7232	0.1365	3.2513	0.6151	0.5529	29.60	0.8576	0.0905	3.3224	0.6195	0.6160	19.22	0.5167	0.2035	3.2480	0.5932	0.5661
StarSRGAN Lite	24.36	0.6708	0.2588	4.3514	0.5016	0.4128	28.19	0.7995	0.1731	3.7280	0.5249	0.5309	17.91	0.4572	0.2780	4.6240	0.5284	0.5011

Method	BSD100						Urban100						Manga109					
	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ
Original HR	80.00	1.0000	0.0000	3.3586	0.8206	0.6761	80.00	1.0000	0.0000	1.8932	0.7743	0.6395	80.00	1.0000	0.0000	3.0512	0.7282	0.7055
SRGAN	20.39	0.6455	0.2679	2.8920	0.6461	0.3503	18.98	0.6557	0.1617	1.6539	0.6303	0.49381	22.58	0.6822	0.1398	2.4046	0.5991	0.5039
ESRGAN	20.26	0.6187	0.2391	2.5170	0.6684	0.3795	18.06	0.6223	0.1435	2.4639	0.6753	0.5204	22.70	0.6735	0.1239	2.4314	0.6266	0.5252
Real-ESRGAN	19.35	0.6127	0.3183	4.6826	0.6815	0.4029	17.75	0.5966	0.2205	4.9107	0.7068	0.4643	20.26	0.6262	0.2203	4.6752	0.6883	0.4366
A-ESRGAN	18.52	0.5729	0.3242	3.4259	0.6583	0.3374	17.77	0.5546	0.2643	1.9287	0.6461	0.4133	20.26	0.5567	0.2757	3.1648	0.6239	0.4398
FeMaSR	19.38	0.6044	0.3193	4.3513	0.6486	0.3586	16.78	0.5580	0.2274	4.2489	0.6234	0.4429	19.28	0.5960	0.2127	3.6849	0.5838	0.4522
Swin2SR	20.97	0.6577	0.4307	3.8591	0.6103	0.3986	20.01	0.6782	0.3591	4.6336	0.5821	0.4864	22.29	0.7138	0.2724	4.5726	0.5663	0.4616
StarSRGAN	20.35	0.6496	0.2302	2.8900	0.7127	0.4508	19.94	0.7167	0.1045	1.9064	0.7299	0.5538	24.61	0.8065	0.1456	3.0859	0.6943	0.5626
StarSRGAN Lite	20.36	0.5736	0.2912	3.1389	0.6124	0.3396	20.50	0.6736	0.2941	3.6585	0.6035	0.4790	22.46	0.7195	0.1919	3.1200	0.6128	0.4994

the 50-layer pre-trained ResNet network presented in [Kai16]. Since the ResNet network structure differs from the VGG network, each feature space is specified by a unique block. ResNet-50 architecture is divided into four stages, each containing several bottleneck layers. The extracted perceptual features use the output value of the bottleneck layer at each stage, and the ResNet loss can also be expressed as:

$$l_{RES/m,n} = \frac{1}{C_{m,n} W_{m,n} H_{m,n}} \sum_{z=1}^{C_{m,n}} \sum_{x=1}^{W_{m,n}} \sum_{y=1}^{H_{m,n}} \\ \left| \beta_{m,n} \left(G_{\theta_G} \left(I^{LR} \right) \right)_{x,y,z} - \beta_{m,n} \left(I^{HR} \right)_{x,y,z} \right| \quad (7)$$

where $\beta_{m,n}$ represents features obtained by the n-th bottleneck layer (after activation) at the m-th stage. $C_{m,n}$, $W_{m,n}$, and $H_{m,n}$ are the dimensions of their respective feature spaces in the ResNet network.

Finally, the Dual Perceptual Loss l_{percep} function under the two perceptual losses is expressed as:

$$l_{DP} = l_{VGG} + \frac{1}{\mu} \zeta_{l_{VGG}, l_{RES}} l_{RES}, \quad (8)$$

where the ResNet loss l_{RES} is weighted dynamically and the weight value is $\frac{1}{\mu} \zeta_{l_{VGG}, l_{RES}}$, μ is a nonzero constant. The $\zeta_{l_{VGG}, l_{RES}}$ can be expressed as:

$$\zeta_{l_{VGG}, l_{RES}} = \frac{l_{VGG} + c}{l_{RES} + c} \quad (9)$$

Where c is a small positive constant when its job only prevents the denominator from becoming zero, therefore, $\frac{1}{\mu} \zeta_{l_{VGG}, l_{RES}}$ is only a value that fluctuates with the ratio of l_{VGG} to l_{RES}. Consequently, $\frac{1}{\mu} \zeta_{l_{VGG}, l_{RES}}$ is regarded as the weight value under the ResNet loss, which can only alter the update range of network parameters and not the update direction.

3.6 Dropout Degradation

In high-level vision tasks, dropout is intended to reduce the overfitting problem. However, it is rarely used in low-level vision tasks such as image SR. As a traditional regression problem, SR behaves differently for high-level tasks and is sensitive to the dropout process. RDSR [Kon21] dropout research demonstrates that appropriate dropout utilization benefits SR networks and improves generalizability. In our study, we employ this approach primarily for observational purposes.

In particular, we add the dropout layer before the final output layer. According to the results of our experiments, this technique improves network performance in a multi-degradation condition.

4 EXPERIMENTS

4.1 Implementation

To better compare the functionality of various mechanisms, including: Adaptive Degradation Model (Adapt-Deg), Attention U-Net Discriminator (Attn-Unet), Multi-scale Discriminator (Multi-Disc), and

Table 2: Upsampling benchmark and inference performance of Real-ESRGAN and StarSRGAN models.
System specification: Ubuntu 22.04 LTS, AMD Ryzen 5 5600X, NVIDIA GeForce RTX 3080 Ti, CUDA 12.1

Method	Upsampling Benchmark (FPS)										Performance			
	360p to 1080p		480p to 1440p		540p to 4K		720p to 5K		1080p to 8K		Image Quality		Inference Time	
	Python	C++	Python	C++	Python	C++	Python	C++	Python	C++	NR-IQA	FR-IQA	PyTorch	NCNN
Real-ESRGAN	2.85	6.94	1.63	3.97	1.29	3.14	0.73	1.78	0.30	0.73	100%	100%	100%	100%
StarSRGAN	2.59	6.42	1.56	3.81	1.24	3.02	0.69	1.72	0.28	0.69	107%	112%	94%	95%
StarSRGAN Lite	21.32	53.68	11.67	28.43	9.36	22.79	5.44	13.59	2.48	6.14	89%	93%	739%	753%

Table 3: Quantitative comparison of StarSRGAN variations on common test datasets for SR (average IQA).
LPIPS/NIQE ↓: the lower, the better. PSNR/SSIM/MANIQA/AHIQ ↑: the higher, the better. Note that MANIQA
and AHIQ are the current SOTA IQA methods. The best and second performance are marked in **red** and blue.
Discussion: The dropout degradation technique needs to be more predictable when implement on SR models.
Some models even have better NIQE than the HR. The IQA may not be accurately sufficient for blind SR tasks.

Method	Adapt-Deg	Attn-Unet	Multi-Disc	Dual-Loss	Drop-Out	PSNR	SSIM	LPIPS	NIQE	MANIQA	AHIQ
Original HR						80.00	1.0000	0.0000	3.4135	0.6986	0.6543
Real-ESRGAN						20.94	0.6166	0.2618	4.1181	0.6378	0.4818
StarSRGAN V1						23.43	0.6893	0.1361	2.5619	0.6320	0.5073
StarSRGAN V2	✓					24.56	0.7152	0.1220	2.7378	0.6422	0.5154
StarSRGAN V3	✓	✓				24.38	0.7065	0.1226	2.7281	0.6582	0.5442
StarSRGAN V4	✓	✓	✓			24.43	0.7185	0.1211	2.8669	0.6590	0.5333
StarSRGAN V5	✓	✓	✓	✓		23.21	0.7117	0.1518	2.9507	0.6608	0.5504
StarSRGAN+ V1					✓	23.12	0.6769	0.1361	2.6444	0.6368	0.5483
StarSRGAN+ V2	✓				✓	23.86	0.7144	0.1475	2.4271	0.6458	0.5681
StarSRGAN+ V3	✓	✓			✓	23.60	0.7035	0.1187	2.3328	0.6421	0.5705
StarSRGAN+ V4	✓	✓	✓		✓	23.60	0.7031	0.1206	2.3468	0.6403	0.5673
StarSRGAN+ V5	✓	✓	✓	✓	✓	24.75	0.7142	0.1255	2.2946	0.6444	0.5357
StarSRGAN Lite V1						19.97	0.5452	0.3684	4.0249	0.5317	0.3410
StarSRGAN Lite V2	✓					22.49	0.6524	0.2340	3.8435	0.5344	0.4287
StarSRGAN Lite V3	✓	✓				21.68	0.6385	0.2373	3.4966	0.5546	0.4489
StarSRGAN Lite V4	✓	✓	✓			22.08	0.6373	0.2350	3.6288	0.5578	0.4732
StarSRGAN Lite V5	✓	✓	✓	✓		22.30	0.6490	0.2479	3.7701	0.5639	0.4605
StarSRGAN Lite+ V1					✓	18.40	0.5688	0.2960	3.6561	0.5161	0.3842
StarSRGAN Lite+ V2	✓				✓	20.30	0.5609	0.2472	3.7053	0.5635	0.4518
StarSRGAN Lite+ V3	✓	✓			✓	18.73	0.5693	0.2491	3.9238	0.5614	0.4487
StarSRGAN Lite+ V4	✓	✓	✓		✓	19.65	0.5646	0.2363	4.9272	0.5813	0.4569
StarSRGAN Lite+ V5	✓	✓	✓	✓	✓	19.93	0.5714	0.2532	5.1208	0.5197	0.4221

Dual Perceptual Loss (Dual-Loss), we build 5 different StarSRGAN models corresponding models, including:

- StarSRGAN (V1): nearest similar to Real-ESRGAN but using the novel Star Dense Block.

- StarSRGAN V1 + Adapt-Deg (V2): using the Apdative Degradation Model instead of the High-order Degradation Model.

- StarSRGAN V2 + Attn-Unet (V3): using the Attention U-Net Discriminator instead of U-Net Discriminator.

- StarSRGAN V3 + Multi-Disc (V4): using the Multiscale Discriminator instead of Single Discriminator.

- StarSRGAN V4 + Dual-Loss (V5): using the Dual Perceptual Loss instead of Single Perceptual Loss.

Similarly, we also have 5 StarSRGAN Lite models (from V1 to V5), representing corresponding variations of StarSRGAN in compact architecture.

To observe the benefit of the dropout degradation (Drop-Out) technique, we also conducted a separate similar experiment with dropout layers, denoted by the plus sign (e.g. StarSRGAN+, StarSRGAN Lite+).

We trained 5 StarSRGAN, 5 StarSRGAN+, 5 StarSRGAN Lite, and 5 StarSRGAN Lite+ models on DIV2K [Nti17a] and Flickr2K [Nti17b] datasets. The training HR size is set to 256. We train our models on an NVIDIA A100 GPU with a batch size of 32 by using Adam optimizer. We train the StarSRNet models for 2000K iterations with a learning rate 2×10^{-4} while training the StarSRGAN models for 1000K iterations with a learning rate 1×10^{-4}. We also adopt the Exponential Moving Average (EMA) for more reg-
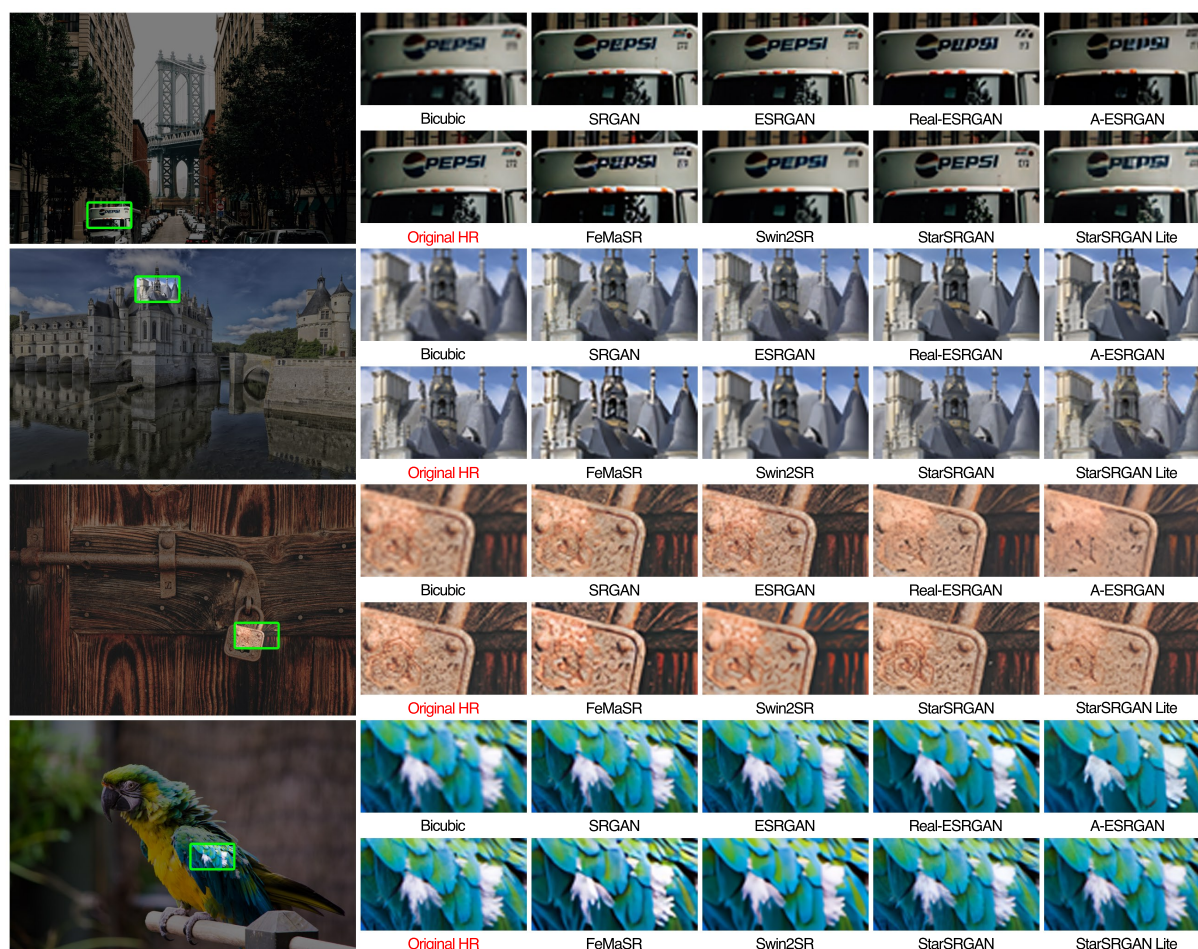
Figure 5: Visual comparisons of standard SR models with StarSRGAN models in 4×. (**Zoom in for best view**)
Discussion: The logo on the truck and the carving of the lock were recovered precisely on StarSRGAN. The image quality reconstructed by StarSRGAN Lite stays caught up with other models. All models deliver pleasant details and textures on the parrot feathers and the bell tower. Artifacts are also hard to be spotted on every model.

ular training and better performance. StarSRGAN and StarSRGAN Lite are trained with a combination of L1 Loss, Perceptual Loss, and GAN loss, with weights $[1,1,0.1]$, respectively. Models which implement the Multi-scale Discriminator are composed of two discriminators, D_{normal} and $D_{sampled}$, which have the input of $1\times$ and $2\times$ down-sampled images as the input. The weight for GAN loss of D_{normal} and $D_{sampled}$ is $[1,1]$. Our implementation is based on the BasicSR [Wan22].

4.2 Datasets

Previous studies have typically evaluated blind image SR models using synthetic LR images manually degraded from HR images. However, these images may not accurately represent the LR images resulting from real-world degradation processes, which often involve complex combinations of multiple degradation processes. Additionally, no publicly available datasets contain LR images from real-world sources. As an alternative, we have used real-world images scaled up by a factor of 4 for testing purposes. We employ

real-world images from 5 classical benchmarks to evaluate our approach, including Set5, Set14, BSD100, Urban100, Manga109, and the modern DIV2K Validation dataset [Nti17a]. These datasets contain images from diverse categories, such as portraits, landscapes, and structures. A reliable SR model should perform well on most of these standard datasets.

4.3 Compared Methods

We examine the proposed StarSRGAN and StarSRGAN Lite models with the SRGAN [Led17], ESRGAN [Wan19], Real-ESRGAN [Wan21], A-ESRGAN [Wei21], FeMaSR [Che22], and Swin2SR [Con22] models. The architecture of StarSRGAN V1 is the nearest similar to the architecture of Real-ESRGAN. More specifically, Residual Dense Block has been replaced with the novel Star Residual Dense Block, which can help evaluate the effectiveness of StarSRGAN even with a slight modification. On the other hand, StarSRGAN Lite models aim to reduce the reconstruction time of super-resolution. Therefore,
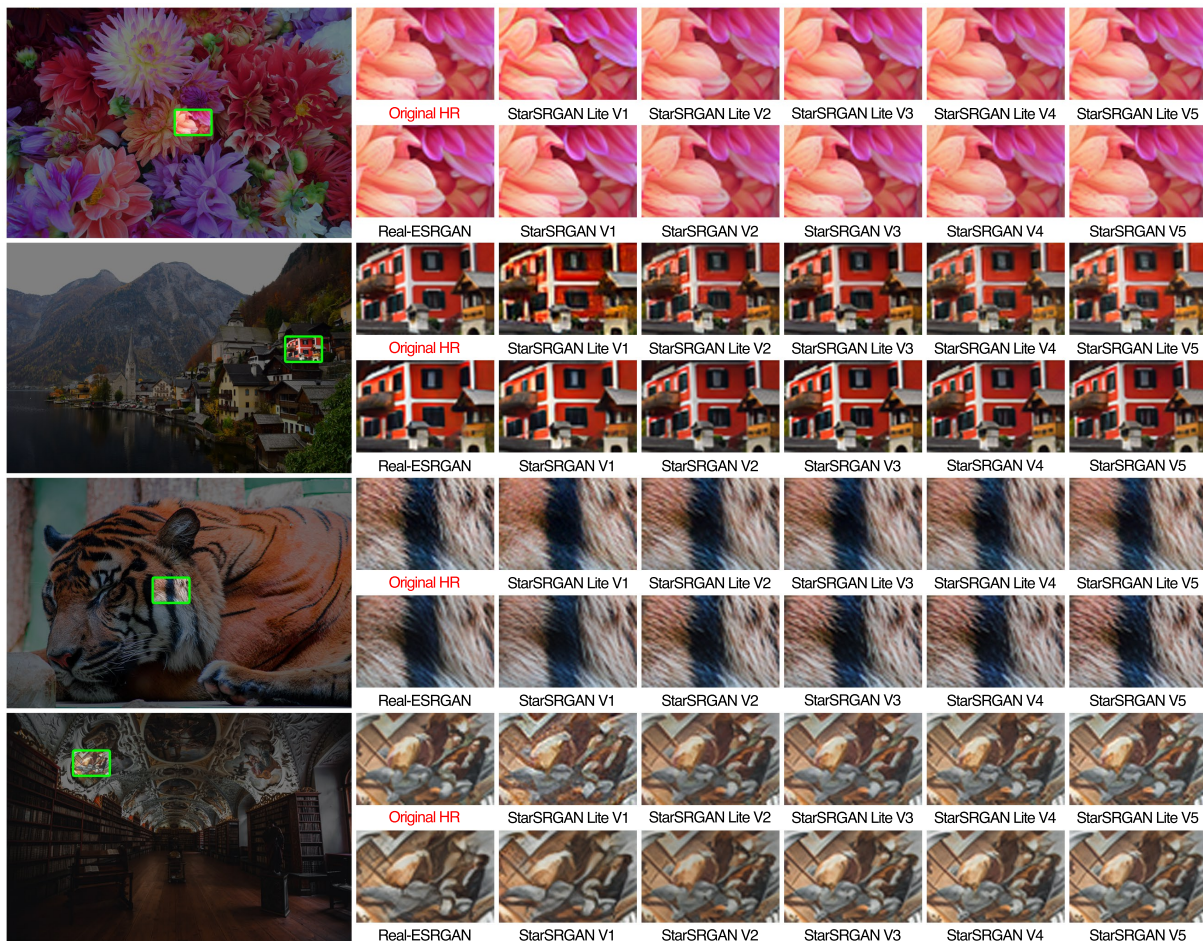
Figure 6: Visual comparisons of StarSRGAN variations with an upsampling scale of $4\times$. (**Zoom in for best view**) **Discussion**: If we pay close attention, we can observe that the StarSRGAN V5 delivered more tiny details on the petals. The chimneys and the windows also look more similar to the HR. Every model reconstructed the fur of the tiger with outstanding quality. The coat color was also better restored on StarSRGAN V5 than on StarSRGAN V1 or Real-ESRGAN. The images enhanced by StarSRGAN Lite models are acceptable compared to its predecessor.

in addition to comparing perceptual quality, we also compare reconstruction time in frame rate (expressed in frames per second or FPS) between the Real-ESRGAN and StarSRGAN, and StarSRGAN Lite models.

4.4 Experiment Results

Table 1 compares StarSRGAN with other SR models on several standard test datasets for SR. The results indicate that the classical IQAs like PSNR, SSIM, and NIQE are no more suitable for SR evaluation. Previous research claims that to achieve high PNSR, the model tends to generate over-smooth results. The baseline model, SRGAN, usually obtains a better SSIM index than more advanced models. Moreover, some models even have a better NIQE score than the HR, making this measure the most unpredictable IQA in this work. On the other hand, the measurements that come from MANIQA and AHIQ are anticipated and reasonable.

Models with real-world emulated data synthesis perform poorly on illustration. Fortunately, StarSRGAN

with Adaptive Degradation Model has partially solved this issue. Unnatural city-featured textures like windows, roads, and bricks tend to be more complicated to reconstruct than other textures. Fine-tuning models on extra training data is promising to sort out the problem.

Through upsampling benchmark results shown in Table 2, we can find that StarSRGAN has traded off its inference time to achieve better image quality compared to Real-ESRGAN. In the opposite direction, StarSRGAN Lite sacrifices its image quality to gain impressive performance in reconstructed time. Specifically, the lightweight architecture has brought real-time performance with more than 20 FPS when upscaling from 540p to 4K with C++ optimized executable file.

Table 3 shows that even our most straightforward variation, StarSRGAN V1, outperforms the Real-ESRGAN method in most metrics. Variations applied Dropout Degradation technique are not steady enough, and further research should be conducted. From visual com-
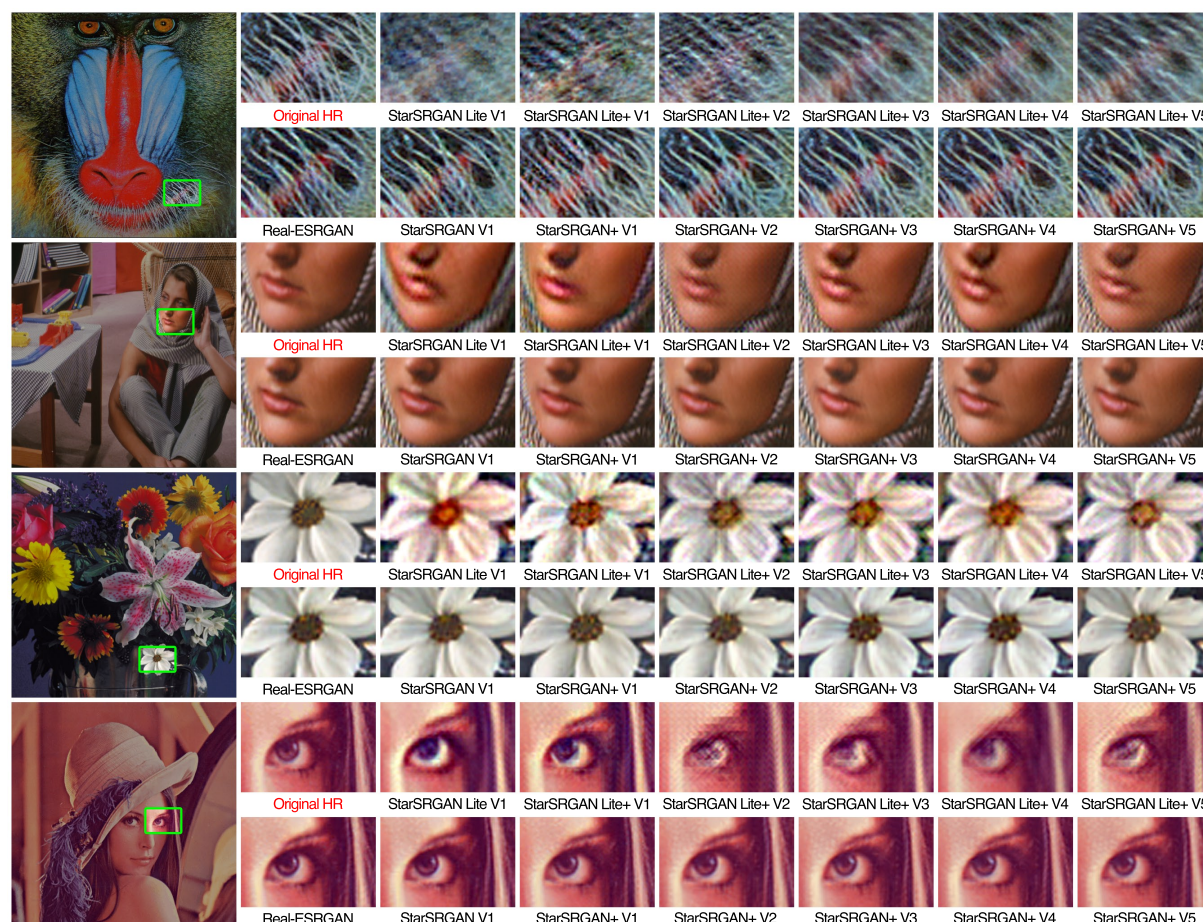
Figure 7: Further observation of StarSRGAN+ models applying Dropout technique. (**Zoom in for best view**)
Discussion: We can easily observe that Dropout Degradation brings unstable performance on StarSRGAN Lite models. On the contrary, the technique adequately integrated with StarSRGAN without apparent conflicts.

parison (some examples are shown in Figure 6 and Figure 7), we observe that our methods can recover sharper edges and restore better texture details. Although, it is hard for human-eye to distinguish because both Real-ESRGAN and StarSRGAN bring excellent perceptual quality. Note that the StarSRGAN V5 models have been selected for comparison in Table 1 and Table 2.

5 CONCLUSIONS

The present study introduces two novel GAN-based models, StarSRGAN and StarSRGAN Lite, for blind SR tasks. StarSRGAN integrates advancements from 5 previous research works and yields new SOTA performance, surpassing the leading SR method, Real-ESRGAN, by 10% on both SOTA No-Reference IQA and Full-Reference IQA methods (MANIQA and AHIQ). StarSRGAN Lite, a lightweight version of StarSRGAN, also inherits improvements from its predecessor and offers real-time inference performance, processing upsampled frames from 540p to 4k at over 20 FPS when executed on a C++ optimized executable file. Several directions for further enhancement of

StarSRGAN architectures are recommended. For instance, retraining the models with newly released datasets such as DIV8K or Unsplash. Additionally, other activation functions such as SiLU and GELU, could be a better alternative for the familiar ReLU. Super-resolving only interested objects and disregarding unnecessary regions like the background could improve StarSRGAN inference performance. Applying Video SR techniques and leveraging spatiotemporal data could also be a promising direction for further research. Currently, StarSRGAN models support only the $4\times$ upscale factor, and other upscale factors such as $2\times$, $8\times$, and $16\times$ are also necessary. Another approach is employing an image classifier to distinguish between real-life and unreal images and choosing the most optimized for each case. Better batch size and more iterations could be explored with more robust hardware. In conclusion, these directions could facilitate the development of future research.

6 REFERENCES

[Don16] Dong, C. et al. (2016) "Image Super-Resolution Using Deep Convolutional Networks", IEEE TPAMI 2016, 38(2), pp. 295-307. doi: 10.1109/TPAMI.2015.2439281.

[Joh16] Johnson, J. et al. (2016) "Perceptual Losses for Real-Time Style Transfer and Super-Resolution", ECCV 2016, pp. 694-711. doi: 10.1007/978-3-319-46475-6_43

[Led17] Ledig, C. et al. (2017) "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network", IEEE/CVF CVPR 2017, pp. 105-114. doi: 10.1109/CVPR.2017.19.

[Wan19] Wang, X. et al. (2019) "ESRGAN: Enhanced super-resolution generative adversarial networks", Springer 2019, pp. 63-79. doi: 10.1007/978-3-030-11021-5_5.

[Goo14] Goodfellow, I. et al. (2014) "Generative Adversarial Networks", NeurIPS 2014, 3. doi: 10.1145/3422622.

[Wan21] Wang, X. et al. (2021) "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data", IEEE/CVF IC-CVW 2021, pp. 1905-1914. doi: 10.1109/IC-CVW54120.2021.00217.

[Rak20] Rakotonirina, N. C. and Rasoanaivo, A. (2020) "ESRGAN+: Further Improving Enhanced Super-Resolution Generative Adversarial Network", IEEE ICASSP 2020, pp. 3637-3641. doi: 10.1109/ICASSP40776.2020.9054071.

[Wei21] Wei, Z. et al. (2021) "A-ESRGAN: Training Real-World Blind Super-Resolution with Attention U-Net Discriminators" arXiv 2021 [eess.IV]. doi: 10.48550/arXiv.2112.10046.

[Lia22] Liang, J. Zeng, H. and Zhang, L. (2022) "Efficient And Degradation-Adaptive Network For Real-World Image Super-Resolution", ECCV 2022, pp. 574-591. doi: 10.1007/978-3-031-19797-0_33.

[Son22] Song, J. et al. (2022) "Dual Perceptual Loss for Single Image Super-Resolution Using ESRGAN", arXiv 2022 [eess.IV]. doi: 10.48550/arXiv.2201.06383.

[Sch20] Schonfeld, E., et al. (2020) "A U-Net Based Discriminator for Generative Adversarial Networks", arXiv 2020 [cs.CV]. doi: 10.48550/arXiv.2002.12655.

[Wan17] Wang, T. et al. (2017) "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs", arXiv 2017 [cs.CV]. doi: 10.48550/arXiv.1711.11585.

[Yan22] Yang, S. et al. (2022) "MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment", arXiv 2022 [cs.CV]. doi: 10.48550/arXiv.2204.08958.

[Lao22] Lao, S. et al. (2022) "Attentions Help CNNs See Better: Attention-Based Hybrid Image Quality Assessment Network" arXiv 2022 [cs.CV]. doi: 10.48550/arXiv.2204.10485.

[Nti22] Gu, J. et al. (2022) "NTIRE 2022 Challenge on Perceptual Image Quality Assessment", IEEE/CVF CVPRW 2022, pp. 950-966. doi: 10.1109/CVPRW56347.2022.00109.

[Kai16] He, K. et al. (2016) "Deep Residual Learning for Image Recognition", IEEE CVPR 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90.

[Sim15] Simonyan, K. and Zisserman, A. (2015) "Very deep convolutional networks for large-scale image recognition", ICLR 2015, pp. 1-14. doi: 10.48550/arXiv.1409.1556.

[Shi16] Shi, W. et al. (2016) "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network", IEEE CVPR 2016, pp. 1874-1883. doi: 10.1109/CVPR.2016.207.

[Nti17a] Agustsson, E. and Timofte, R. (2017) "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study", IEEE CVPRW 2017, pp. 1122-1131. doi: 10.1109/CVPRW.2017.150.

[Nti17b] Timofte, R. et al. (2017) "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results", IEEE CVPRW 2017, pp. 1110-1121. doi: 10.1109/CVPRW.2017.149.

[Wan22] Wang, X. et al. (2022) "BasicSR: Open Source Image and Video Restoration Toolbox". GitHub 2022. https://github.com/XPixelGroup/BasicSR.

[Okt18] Oktay, O. et al. (2018) "Attention U-Net: Learning Where to Look for the Pancreas" arXiv 2018 [cs.CV]. doi: 10.48550/arXiv.1804.03999.

[Miy18] Miyato, T. et al. (2018) "Spectral Normalization for Generative Adversarial Networks", ICLR 2018. doi: 10.48550/arXiv.1802.05957.

[Kon21] Kong, X. et al. (2021) "Reflash Dropout in Image Super-Resolution." arXiv 2021 [cs.CV]. doi: 10.48550/arXiv.2112.12089.

[Che22] Chen, C. et al. (2022) "Real-World Blind Super-Resolution via Feature Matching with Implicit High-Resolution Priors", ACM 2022, pp. 1329-1338. doi: 10.1145/3503161.3547833.

[Con22] Conde, M. V. et al. (2023) "Swin2SR: SwinV2 Transformer for Compressed Image Super-Resolution and Restoration", Springer 2023, pp. 669-687. doi: 10.1007/978-3-031-25063-7_42