FouriScale: A Frequency Perspective on Training-Free High-Resolution Image Synthesis

Linjiang Huang[1,3⋆], Rongyao Fang[1⋆], Aiping Zhang[4], Guanglu Song[5]
Si Liu[6], Yu Liu[5], and Hongsheng Li[1,2,3] ✉

[1] CUHK MMLab [2] Shanghai AI Laboratory [3] CPII under InnoHK
[4] Sun Yat-Sen University [5] SenseTime Research [6] Beihang University
ljhuang524@gmail.com; {rongyaofang@link, hsli@ee}.cuhk.edu.hk

Abstract. In this study, we delve into the generation of high-resolution images from pre-trained diffusion models, addressing persistent challenges, such as repetitive patterns and structural distortions, that emerge when models are applied beyond their trained resolutions. To address this issue, we introduce an innovative, training-free approach FouriScale from the perspective of frequency domain analysis. We replace the original convolutional layers in pre-trained diffusion models by incorporating a dilation technique along with a low-pass operation, intending to achieve structural consistency and scale consistency across resolutions, respectively. Further enhanced by a padding-then-crop strategy, our method can flexibly handle text-to-image generation of various aspect ratios. By using the FouriScale as guidance, our method successfully balances the structural integrity and fidelity of generated images, achieving arbitrary-size, high-resolution, and high-quality generation. With its simplicity and compatibility, our method can provide valuable insights for future explorations into the synthesis of ultra-high-resolution images. The source code is available at https://github.com/LeonHLJ/FouriScale.

Keywords: Diffusion Model · Training Free · High-Resolution Synthesis

1 Introduction

Recently, Diffusion models [18, 34] have emerged as the predominant generative models, surpassing the popularity of GANs [12] and autoregressive models [9,31]. Some text-to-image generation models, which are based on diffusion models, such as Stable Diffusion (SD) [34], Stable Diffusion XL (SDXL) [30], Midjourney [27], and Imagen [35], have shown their astonishing capacity to generate high-quality and fidelity images under the guidance of text prompts. To ensure efficient processing on existing hardware and stable model training, these models are typically trained at one or a few specific image resolutions. For instance, SD models are often trained using images of 512×512 resolution, while SDXL models are typically trained with images close to 1024×1024 pixels.

⋆ Equal contribution. ✉ Corresponding author.

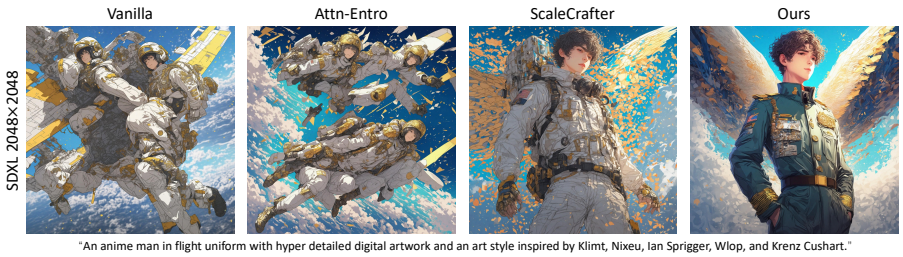


Fig. 1: Visualization of pattern repetition issue of high-resolution image synthesis (2048×2048) using SDXL [30]. Attn-Entro [23] fails to address this problem and ScaleCrafter [14] still struggles with this issue in image details. Our method successfully handles this problem and generates high-quality images without model retraining.

However, as shown in Fig. 1, directly employing pre-trained diffusion models to generate an image at a higher resolution will lead to repetitive patterns and unforeseen artifacts. Some studies [2, 22, 24] have attempted to create larger images by utilizing pre-trained diffusion models to stitch together overlapping patches into a panoramic image. Nonetheless, the absence of a global direction for the whole image restricts their ability to generate images focused on specific objects and fails to address the issue of repetitive patterns. [23] has explored adapting pre-trained diffusion models for generating images of various sizes by examining attention entropy. Nevertheless, ScaleCrafter [14] found that the key point of generating high-resolution images lies in the convolution layers. They introduce a re-dilation operation and a convolution disperse operation to enlarge kernel sizes of convolution layers, largely mitigating the problem. However, their conclusion stems from empirical findings, lacking a deeper exploration of this issue. Additionally, it needs an initial offline computation of a linear transformation between the original convolutional kernel and the enlarged kernel, falling short in terms of compatibility and scalability when there are variations in the kernel sizes of the UNet and the desired target resolution of images.

In this work, we present FouriScale, an innovative and effective approach that handles the issue through the perspective of frequency domain analysis, successfully demonstrating its effectiveness through both theoretical analysis and experimental results. FouriScale substitutes the original convolutional layers in pre-trained diffusion models by simply introducing a dilation operation coupled with a low-pass operation, aimed at achieving structural and scale consistency across resolutions, respectively. Equipped with a padding-then-crop strategy, our method allows for flexible text-to-image generation of different sizes and aspect ratios. Furthermore, by utilizing FouriScale as guidance, our approach attains remarkable capability in producing high-resolution images of any size, with integrated image structure alongside superior quality. The simplicity of FouriScale eliminates the need for any offline pre-computation, facilitating compatibility and scalability. We envision FouriScale providing significant contributions to the advancement of ultra-high-resolution image synthesis in future research.

2 Related Work

2.1 Text-to-Image Synthesis

Text-to-image synthesis [7, 19, 34, 35] has seen a significant surge in interest due to the development of diffusion probabilistic models [18, 38]. These innovative models generate data from a Gaussian distribution and refine it through a denoising process. With their capacity for high-quality generation, they have made significant leaps over traditional models like GANs [7, 12]. The Latent Diffusion Model (LDM) [34] integrates the diffusion process within a latent space, achieving astonishing results in the generation of realistic images, which boosts significant interest in the domain of generating via latent space [4, 15, 25, 29, 42]. These models are typically trained at one or a few specific image resolutions to ensure efficient processing on existing hardware and stable model training. For instance, Stabe Diffusion (SD) [34] is trained using 512×512 pixel images, while SDXL [30] models are typically trained with images close to 1024×1024 resolution, accommodating various aspect ratios simultaneously.

2.2 High-Resolution Synthesis via Diffusion Models

High-resolution synthesis has always received widespread attention. Prior works mainly focus on refining the noise schedule [6, 21], developing cascaded architectures [19, 35, 39] or mixtures-of-denoising-experts [1] for generating high-resolution images. Despite their impressive capabilities, diffusion models were often limited by specific resolution constraints. Some methods have tried to address these issues by accommodating a broader range of resolutions. For example, Any-size Diffusion [46] fine-tunes a pre-trained SD on a set of images with a fixed range of aspect ratios, similar to SDXL [30]. FiT [26] views the image as a sequence of tokens and adaptively padding image tokens to a predefined maximum token limit, ensuring hardware-friendly training and flexible resolution handling. However, these models require model training, overlooking the inherent capability of the pre-trained models to handle image generation with varying resolutions. Most recently, some methods [2, 22, 24] have attempted to generate panoramic images by utilizing pre-trained diffusion models to stitch together overlapping patches. [23] has explored adapting pre-trained diffusion models for generating images of various sizes by examining attention entropy. ElasticDiff [13] uses default resolution to guide the generation of arbitrary-size images. DemoFusion [10] adapts a cascaded fashion with a strategy of low- and high-resolution fusion to maintain global consistency. Recently, ScaleCrafter [14] finds that the key point lies in convolution layers. They present a re-dilation and a convolution disperse operation to expand convolution kernel sizes, which requires an offline calculation of a linear transformation from the original convolutional kernel to the expanded one. In contrast, we deeply investigate the issue of repetitive patterns and handle it through the perspective of frequency domain analysis. The simplicity of our method eliminates the need for any offline pre-computation, facilitating its compatibility and scalability.

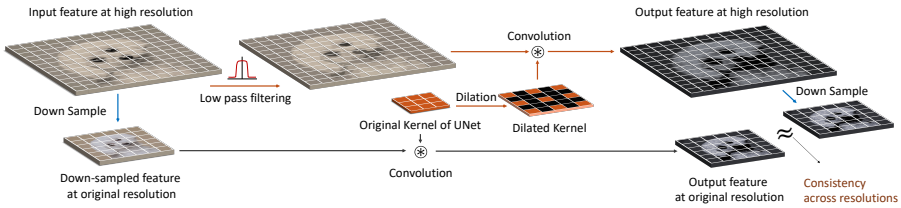


Fig. 2: The overview of FouriScale (orange line), which includes a dilation convolution operation (Sec. 3.2) and a low-pass filtering operation (Sec. 3.3) to achieve structural consistency and scale consistency across resolutions, respectively.

3 Method

Diffusion models, also known as score-based generative models [18, 38], belong to a category of generative models that follow a process of progressively introducing Gaussian noise into the data and subsequently generating samples from this noise through a reverse denoising procedure. The key denoising step is typically carried out by a U-shaped Network (UNet), which learns the underlying denoising function that maps from noisy data to its clean counterpart. The UNet architecture, widely adopted for this purpose, comprises stacked convolution layers, self-attention layers, and cross-attention layers. Some previous works have explored the degradation of performance when the generated resolution becomes larger, attributing to the change of the attention tokens' number [23] and the reduced relative receptive field of convolution layers [14]. Based on empirical evidence in [14], convolutional layers are more sensitive to changes in resolution. Therefore, we primarily focus on studying the impact brought about by the convolutional layers. In this section, we will introduce FouriScale, as shown in Fig. 2. It includes a dilation convolution operation (Sec. 3.2) and a low-pass filtering operation (Sec. 3.3) to achieve structural consistency and scale consistency across resolutions, respectively. With the tailored padding-then-cropping strategy (Sec. 3.4), FouriScale can generate images of arbitrary aspect ratios. By utilizing FouriScale as guidance (Sec. 3.5), our approach attains remarkable capability in generating high-resolution and high-quality images.

3.1 Notation

2D Discrete Fourier Transform (2D DFT). Given a two-dimensional discrete signal $F(m, n)$ with dimensions $M \times N$, the two-dimensional discrete Fourier transform (2D DFT) is defined as:

$$F(p, q) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} F(m, n) e^{-j2\pi\left(\frac{pm}{M} + \frac{qn}{N}\right)}. \tag{1}$$

2D Dilated Convolution. A dilated convolution kernel of the kernel $k(m, n)$, denoted as $k_{d_h, d_w}(m, n)$, is formed by introducing zeros between the elements of

the original kernel such that:

$$k_{d_h,d_w}(m,n) = \begin{cases} k(\frac{m}{d_h}, \frac{n}{d_w}) & \text{if } m \,\%\, d_h = 0 \text{ and } n \,\%\, d_w = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where d_h, d_w is the dilation factor along height and width, respectively, m and n are the indices in the dilated space. The % represents the modulo operation.

3.2 Structural Consistency via Dilated Convolution

The diffusion model's denoising network, denoted as ϵ_θ, is generally trained on images or latent spaces at a specific resolution of $h \times w$. This network is often constructed using a U-Net architecture. Our target is to generate an image of a larger resolution of $H \times W$ at the inference stage using the parameters of denoising network ϵ_θ without retraining.

As previously discussed, the convolutional layers within the U-Net are largely responsible for the occurrence of pattern repetition when the inference resolution becomes larger. To prevent structural distortion at the inference resolution, we resort to establishing structural consistency between the low resolution and high resolution, as shown in Fig. 2. In particular, for a convolutional layer Conv_k in the UNet with its convolution kernel k, and the high-resolution input feature map F, the structural consistency can be formulated as follows:

$$\text{Down}_s(F) \circledast k = \text{Down}_s(F \circledast k'), \tag{3}$$

where Down_s denotes the down-sampling operation with scale s[1], and $\circledast$ represents the convolution operation. This equation implies the need to customize a new convolution kernel k' for a larger resolution. However, finding an appropriate k' can be challenging due to the variety of feature map F. The recent ScaleCrafter [14] method uses structure-level and pixel-level calibrations to learn a linear transformation between k and k', but learning a new transformation for each new kernel size and new target resolution can be cumbersome.

In this work, we propose to handle the structural consistency from a frequency perspective. Suppose the input $F(x, y)$, which is a two-dimensional discrete spatial signal, belongs to the set $\mathbb{R}^{H_f \times W_f \times C}$. The sampling rates along the x and y axes are given by Ω_x and Ω_y correspondingly. The Fourier transform of $F(x, y)$ is represented by $F(u, v) \in \mathbb{R}^{H_f \times W_f \times C}$. In this context, the highest frequencies along the u and v axes are denoted as u_{max} and v_{max}, respectively. Additionally, the Fourier transform of the downsampled feature map $\text{Down}_s(F(x, y))$, which is dimensionally reduced to $\mathbb{R}^{\frac{H_f}{s} \times \frac{W_f}{s} \times C}$, is denoted as $F'(u, v)$.

Theorem 1. *Spatial down-sampling leads to a reduction in the range of frequencies that the signal can accommodate, particularly at the higher end of the*

[1] For simplicity, we assume equal down-sampling scales for height and width. Our method can also accommodate different down-sampling scales in this context through our padding-then-cropping strategy (Section 3.4).

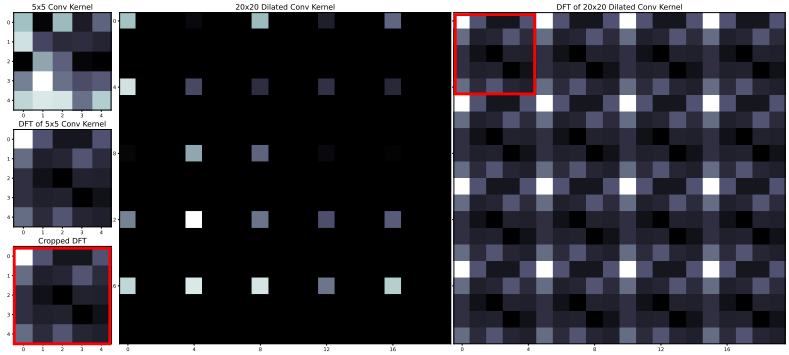


Fig. 3: We visualize a random 5×5 kernel for better visualization. The Fourier spectrum of its dilated kernel, with a dilation factor of 4, clearly demonstrates a periodic character. It should be noted that we also pad zeros to the right and bottom sides of the dilated kernel, which differs from the conventional use. However, this does not impact the outcome in practical applications.

spectrum. This process causes high frequencies to be folded to low frequencies, and superpose onto the original low frequencies. For a one-dimensional signal, in the condition of s strides, this superposition of high and low frequencies resulting from down-sampling can be mathematically formulated as

$$F'(u) = \mathbb{S}(F(u), F\left(u + \frac{a\Omega_x}{s}\right)) \mid u \in \left(0, \frac{\Omega_x}{s}\right), \tag{4}$$

where $\mathbb{S}$ dentes the superposing operator, Ω_x is the sampling rates in x axis, and $a = 1, \ldots, s-1$.

Lemma 1. *For an image, the operation of spatial down-sampling using strides of s can be viewed as partitioning the Fourier spectrum into $s \times s$ equal patches and then uniformly superimposing these patches with an average scaling of $\frac{1}{s^2}$.*

$$\mathrm{DFT}\left(\mathrm{Down}_s(F(x,y))\right) = \frac{1}{s^2} \sum_{i=0}^{s-1} \sum_{j=0}^{s-1} F_{(i,j)}(u,v), \tag{5}$$

where $F_{(i,j)}(u,v)$ is a sub-matrix of $F(u,v)$ by equally splitting $F(u,v)$ into $s \times s$ non-overlapped patches and $i, j \in \{0, 1, \ldots, s-1\}$.

The proof of Theorem 1 and Lemma 1 are provided in the supplementary material. They describe the shuffling and superposing [32,43,47] in the frequency domain imposed by spatial down-sampling. If we transform Eq. (3) to the fre-

quency domain and follow conclusion in Lemma 1, we can obtain:

$$\begin{aligned}
&\left(\frac{1}{s^2}\sum_{i=0}^{s-1}\sum_{j=0}^{s-1}F_{(i,j)}(u,v)\right)\odot k(u,v) \leftarrow \text{Left side of Eq. (3)} \\
&=\frac{1}{s^2}\sum_{i=0}^{s-1}\sum_{j=0}^{s-1}\left(F_{(i,j)}(u,v)\odot k(u,v)\right) \qquad (6)\\
&=\frac{1}{s^2}\sum_{i=0}^{s-1}\sum_{j=0}^{s-1}\left(F_{(i,j)}(u,v)\odot k'_{(i,j)}(u,v)\right), \leftarrow \text{Right side of Eq. (3)}
\end{aligned}$$

where $k(u,v)$, $k'(u,v)$ denote the fourier transform of kernel k and k', respectively, $\odot$ is element-wise multiplication. Eq. (6) suggests that the Fourier spectrum of the ideal convolution kernel k' should be the one that is stitched by $s \times s$ Fourier spectrum of the convolution kernel k. In other words, there should be a periodic repetition in the Fourier spectrum of k', the repetitive pattern is the Fourier spectrum of k.

Fortunately, the widely used dilated convolution perfectly meets this requirement. Suppose a kernel $k(m,n)$ with the size of $M \times N$, it's dilated version is $k_{d_h,d_w}(m,n)$, with dilation factor of (d_h,d_w). For any integer multiples of d_h, namely $p' = pd_h$ and integer multiples of d_w, namely $q' = qd_w$, the exponential term of the dilated kernel in the 2D DFT (Eq. (1)) becomes:

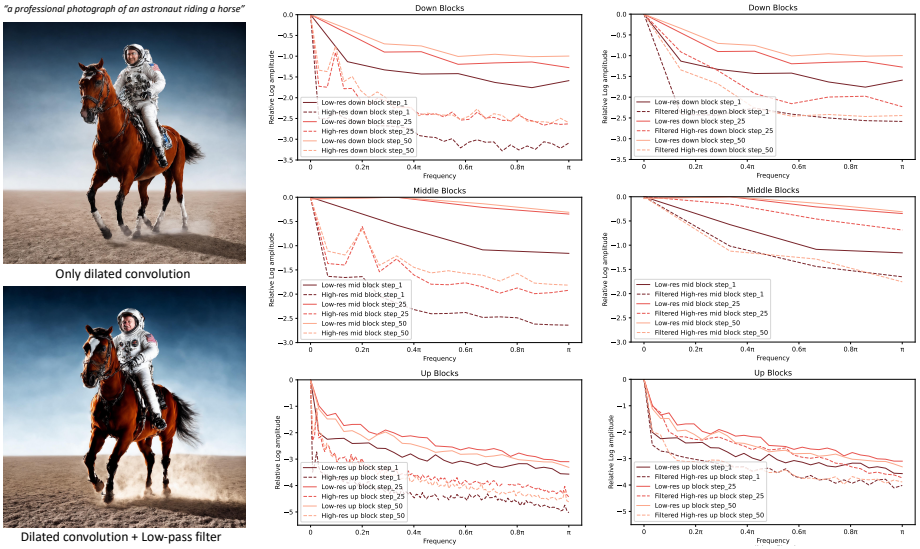
$$e^{-j2\pi\left(\frac{p'm}{d_h M}+\frac{q'n}{d_w N}\right)} = e^{-j2\pi\left(\frac{pm}{M}+\frac{qn}{N}\right)}, \qquad (7)$$

which is periodic with a period of M along the m-dimension and a period of N along the n-dimension. It indicates that a dilated convolution kernel parameterized by the original kernel k, with dilation factor of $(H/h, W/w)$, is the ideal convolution kernel k'. In Fig. 3, we visually demonstrate the periodic repetition of dilated convolution. We noticed that [14] also uses dilated operation. In contrast to [14], which is from empirical observation, our work begins with a focus on frequency analysis and provides theoretical justification for its effectiveness.

3.3 Scale Consistency via Low-pass Filtering

However, in practice, dilated convolution alone cannot well mitigate the issue of pattern repetition. As shown in Fig. 4a (top left), the issue of pattern repetition is significantly reduced, but certain fine details, like the horse's legs, still present issues. This phenomenon is because of the aliasing effect after the spatial down-sampling, which raises the distribution gap between the features of low resolution and the features down-sampled from high resolution, as presented in Fig. 4b. Aliasing alters the fundamental frequency components of the original signal, breaking its consistency across scales.

Here, we introduce a low-pass filtering operation, or spectral pooling [33] to remove high-frequency components that might cause aliasing, intending to construct scale consistency among resolutions. Let $F(m,n)$ be a two-dimensional



(a) Visual comparisons **(b)** without filtering **(c)** with filtering

Fig. 4: (a) Visual comparisons between the images created at a resolution of 2048 × 2048: with only the dilated convolution, and with both the dilated convolution and the low-pass filtering. (b)(c) Fourier relative log amplitudes of input features from three distinct layers from the down blocks, mid blocks, and up blocks of UNet, respectively, are analyzed. We also include features at reverse steps 1, 25, and 50. (b) Without the application of the low-pass filter. There is an evident distribution gap of the frequency spectrum between the low- and high-resolution. (c) With the application of the low-pass filter. The distribution gap is largely reduced.

discrete signal of resolution $M \times N$. Spatial down-sampling of $F(m, n)$, by factors s_h and s_w along the height and width respectively, alters the Nyquist limits to $M/(2s_h)$ and $N/(2s_w)$ in the frequency domain, corresponding to half the new sampling rates along each dimension. The expected low-pass filter should remove frequencies above these new Nyquist limits to prevent aliasing. Therefore, the optimal mask size (assuming the frequency spectrum is centralized) for passing low frequencies is $M/s_h \times N/s_w$. It ensures the preservation of all valuable frequencies within the downscaled resolution while preventing aliasing.

As illustrated in Fig. 4c, the application of the low-pass filter results in a closer alignment of the frequency distribution between high and low resolutions. This ensures that the left side of Eq. (3) produces a plausible image structure. Additionally, since our target is to rectify the image structure, low-pass filtering would not be harmful because it generally preserves the structural information of a signal, which predominantly resides in the lower frequency components [28,44].

Subsequently, the final kernel k^* is obtained by applying low-pass filtering to the dilated kernel. Given the periodic nature of the Fourier spectrum of the dilated kernel, the Fourier spectrum of the new kernel k^* is obtained by expanding

Fig. 5: Visual comparisons between the images generated at a resolution of 2048×1024 by SD 2.1: (a) without the application of padding-then-cropping strategy, and (b) with the application of padding-then-cropping strategy.

the original kernel k's spectrum with zero frequencies. Therefore, this expansion avoids introducing new frequency components into the new kernel k^*. In practice, we do not directly calculate the kernel k^* but replace the original Conv_k with the following equivalent operation to ensure computational efficiency:

$$\text{Conv}_k(F) \rightarrow \text{Conv}_{k'}(\text{iDFT}(H \odot \text{DFT}(F)), \tag{8}$$

where H denotes the low-pass filter. Fig. 4a (bottom left) shows that the combination of dilated convolution and low-pass filtering resolves pattern repetition.

3.4 Adaption to Arbitrary-size Generation

The derived conclusion applies only when the high-resolution and low-resolution images have identical aspect ratios. From Eq. (5) and Eq. (6), it becomes apparent that when the aspect ratios vary, meaning the dilation rates along the height and width are different, the well-constructed structure in the low-resolution image would be distorted and compressed, as shown in Fig. 5 (a). However, in real-world applications, the ideal scenario is for a pre-trained diffusion model to generate arbitrary-sized images.

We introduce a straightforward yet efficient approach, termed *padding-then-cropping*, to solve this problem. Fig. 5 (b) demonstrates its effectiveness. In essence, when a layer receives an input feature at a standard resolution of $h_f \times w_f$, and this input feature increases to a size of $H_f \times W_f$ during inference, our first step is to zero-pad the input feature to a size of $rh_f \times rw_f$. Here, r is defined as the maximum of $\lceil \frac{H_f}{h_f} \rceil$ and $\lceil \frac{W_f}{w_f} \rceil$, with $\lceil \cdot \rceil$ representing the ceiling operation. The padding operation assumes that we aim to generate an image of size $rh \times rw$, where certain areas are filled with zeros. Subsequently, we apply Eq. (8) to rectify the issue of repetitive patterns in the higher-resolution output. Ultimately, the obtained feature is cropped to restore its intended spatial size. This step is necessary to not only negate the effects of zero-padding but also control the computational demands when the resolution increases, particularly those arising from the self-attention layers in the UNet architecture. Taking computational efficiency into account, our equivalent solution is outlined in Algorithm 1.

Algorithm 1 Pseudo-code of FouriScale

Data: Input: $F \in \mathbb{R}^{C\times H_f\times W_f}$. Original size: $h_f \times w_f$.
Result: Output: $F_{conv} \in \mathbb{R}^{C\times H_f\times W_f}$
$r = max(\lceil \frac{H_f}{h_f} \rceil, \lceil \frac{W_f}{w_f} \rceil)$
$F_{pad} \leftarrow$ Zero-Pad$(F) \in \mathbb{R}^{C\times rh_f\times rw_f}$ ▷ Zero Padding
$F_{dft} \leftarrow$ DFT$(F_{pad}) \in \mathbb{C}^{C\times rh_f\times rw_f}$ ▷ Discrete Fourier transform
$F_{low} \leftarrow H \odot F_{dft}$ ▷ Low pass filtering
$F_{idft} \leftarrow$ iDFT(F_{low}) ▷ Inverse Fourier transform
$F_{crop} \leftarrow$ Crop$(F_{idft}) \in \mathbb{R}^{R\times H_f\times W_f}$ ▷ Cropping
$F_{conv} \leftarrow$ Conv$_{k'}(F_{crop})$ ▷ Dilation factor of k' is r

3.5 FouriScale Guidance

FouriScale effectively mitigates structural distortion. However, it would introduce certain artifacts and unforeseen patterns in the background, as depicted in Fig. 6 (b). We identify that the main issue stems from the application of low-pass filtering when generating the conditional estimation in classifier-free guidance [20], which leads to a ringing effect and loss of detail. To improve image quality and reduce artifacts, as shown in Fig. 6 (a), we develop a guided version of FouriScale for reference, aiming to align the output, rich in details, with it. Specifically, beyond the unconditional and conditional estimations from the FouriScale-modified UNet, we generate an extra conditional estimation. This one is subjected to identical dilated convolutions but utilizes milder low-pass filters to accommodate more frequencies. We substitute its attention maps of attention layers with those from the conditional estimation processed through FouriScale, in a similar spirit with image editing [5,11,16]. This strategy allows for the incorporation of correct structural information [40, 41, 45] derived from FouriScale to guide the generation, simultaneously mitigating the decline in image quality and loss of details. The final noise estimation is determined using both the unconditional and the newly conditional estimations following classifier-free guidance. As we can see in Fig. 6 (c), the aforementioned issues are largely mitigated.

3.6 Detailed Designs

Annealing dilation and filtering. Since the image structure is primarily outlined in the early reverse steps, the subsequent steps focus on enhancing the details, we implement an annealing approach for both dilation convolution and low-pass filtering. Initially, for the first S_{init} steps, we employ the ideal dilation convolution and low-pass filtering. During the span from S_{init} to S_{stop}, we progressively decrease the dilation factor and r (as detailed in Algorithm 1) down to 1. After S_{stop} steps, the original UNet is utilized to refine image details further.
Settings for SDXL. For Stable Diffusion XL [30] (SDXL), our observations reveal that using an ideal low-pass filter leads to suboptimal outcomes. Instead, a gentler low-pass filter, which modulates rather than completely eliminates high-frequency elements using a coefficient $\sigma \in [0, 1]$ (set to 0.6 in our method) delivers

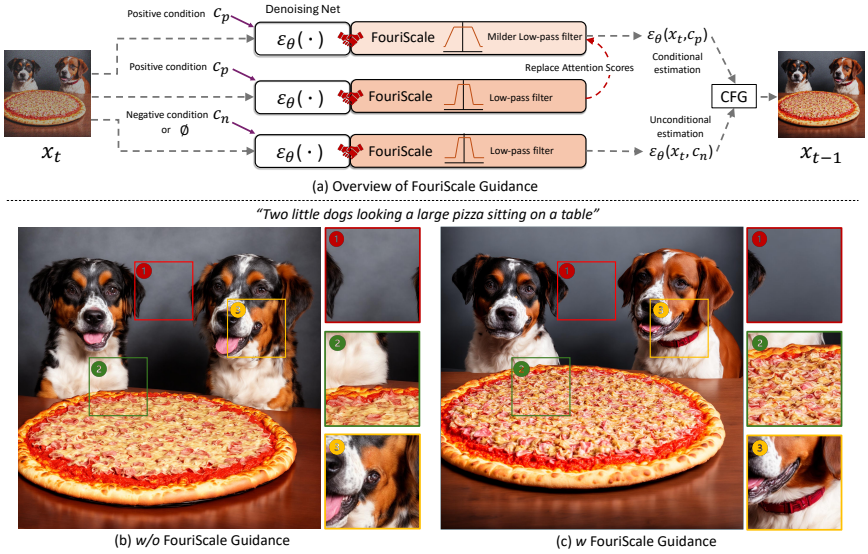


Fig. 6: (a) Overview of FouriScale guidance. CFG denotes Classifier-Free Guidance. (b)(c) Visual comparisons between the images created at 2048 × 2048 by SD 2.1: (b) without the application of FouriScale guidance, ❶ has unexpected artifacts in the background, ❷❸ are wrong details, (c) with the application of FouriScale guidance.

superior visual quality. This phenomenon can be attributed to SDXL's ability to handle changes in scale effectively, negating the need for an ideal low-pass filter to maintain scale consistency, which confirms the rationale of incorporating low-pass filtering to address scale variability. For SDXL, we calculate the scale factor r (refer to Algorithm 1) by determining the training resolution whose aspect ratio is closest to the one of target resolution.

4 Experiments

Experimental setup. Wo follow [14] to report results on three text-to-image models, including SD 1.5 [11], SD 2.1 [8], and SDXL 1.0 [30]. The tested resolutions are 4×, 6.25×, 8×, and 16× the pixel count of their respective training resolutions. For both SD 1.5 and SD 2.1 models, the training resolution is set at 512×512 pixels, while the inference resolutions are 1024^2, 1280^2, 2048×1024, and 2048^2. For SDXL, it is trained at resolutions close to 1024^2 pixels, with the higher inference resolutions being 2048^2, 2560^2, 4096×2048, and 4096^2. We default use FreeU [37] in all experimental settings.

Testing dataset and evaluation metrics. Following [14], we assess performance using the Laion-5B dataset [36], which comprises 5 billion pairs of images and their corresponding captions. For tests conducted at an inference resolution of 1024×1024, we select a subset of 30,000 images, each paired with randomly

Table 1: Quantitative comparisons among training-free methods. The best and second best results are highlighted in **bold** and underline. KID_r and KID_b are scaled by 10^2.

Resolution	Method	SD 1.5				SD 2.1				SDXL 1.0			
		$FID_r\downarrow$	$KID_r\downarrow$	$FID_b\downarrow$	$KID_b\downarrow$	$FID_r\downarrow$	$KID_r\downarrow$	$FID_b\downarrow$	$KID_b\downarrow$	$FID_r\downarrow$	$KID_r\downarrow$	$FID_b\downarrow$	$KID_b\downarrow$
4× 1:1	Vanilla	26.96	1.00	15.72	0.42	29.90	1.11	19.21	0.54	49.81	1.84	32.90	0.92
	Attn-Entro	26.78	0.97	15.64	0.42	29.65	1.10	19.17	0.54	49.72	1.84	32.86	0.92
	ScaleCrafter	23.90	0.95	11.83	0.32	25.19	**0.98**	13.88	**0.40**	49.46	1.73	36.22	1.07
	Ours	**23.62**	**0.92**	**10.62**	**0.29**	**25.17**	**0.98**	**13.57**	**0.40**	**33.89**	**1.21**	**20.10**	**0.47**
6.25× 1:1	Vanilla	41.04	1.28	31.47	0.77	45.81	1.52	37.80	1.04	68.87	2.79	54.34	1.92
	Attn-Entro	40.69	1.31	31.25	0.76	45.77	1.51	37.75	1.04	68.50	2.76	54.07	1.91
	ScaleCrafter	37.71	1.34	25.54	0.67	35.13	1.14	23.68	0.57	55.03	2.02	45.58	1.49
	Ours	**30.27**	**1.00**	**16.71**	**0.34**	**30.82**	**1.01**	**18.34**	**0.42**	**44.13**	**1.64**	**37.09**	**1.16**
8× 1:2	Vanilla	50.91	1.87	44.65	1.45	57.80	2.26	51.97	1.81	90.23	4.20	79.32	3.42
	Attn-Entro	50.72	1.86	44.49	1.44	57.42	2.26	51.67	1.80	89.87	4.15	79.00	3.40
	ScaleCrafter	35.11	1.22	29.51	0.81	41.72	1.42	35.08	1.01	106.57	5.15	108.67	5.23
	Ours	**35.04**	**1.19**	**26.55**	**0.72**	**37.19**	**1.29**	**27.69**	**0.74**	**71.77**	**2.79**	**70.70**	**2.65**
16× 1:1	Vanilla	67.90	2.37	66.49	2.18	84.01	3.28	82.25	3.05	116.40	5.45	109.19	4.84
	Attn-Entro	67.45	2.35	66.16	2.17	83.68	3.30	81.98	3.04	113.25	5.44	106.34	4.81
	ScaleCrafter	32.00	1.01	27.08	0.71	40.91	1.32	33.23	0.90	84.58	3.53	85.91	3.39
	Ours	**30.84**	**0.95**	**23.29**	**0.57**	**39.49**	**1.27**	**28.14**	**0.73**	**56.66**	**2.18**	**49.59**	**1.63**

chosen text prompts from the dataset. Given the substantial computational demands, our sample size is reduced to 10,000 images for tests at inference resolutions exceeding 1024×1024. We evaluate the quality and diversity of the generated images by measuring the Frechet Inception Distance (FID) [17] and Kernel Inception Distance (KID) [3] between generated images and real images, denoted as FID_r and KID_r. To show the methods' capacity to preserve the pre-trained model's original ability at a new resolution, we also follow [14] to evaluate the metrics between the generated images at the base training resolution and the inference resolution, denoted as FID_b and KID_b.

4.1 Quantitative Results

We compare our method with the vanilla text-to-image diffusion model (Vanilla), the training-free approach [23] (Attn-Entro) that accounts for variations in attention entropy, and ScaleCrafter [14], which modifies convolution kernels through re-dilation and linear transformation. We show the experimental results in Tab. 1. Compared to the vanilla diffusion models, our method obtains much better results. Attn-Entro fails at high resolutions due to ignoring cross-resolution structural consistency. Due to the absence of scale consistency, ScaleCrafter performs worse than our method on the majority of metrics. Additionally, we observe that ScaleCrafter often struggles to produce acceptable images for SDXL. Conversely, our method can generate images with plausible structures and rich details at various high resolutions, compatible with any pre-trained diffusion models

Besides, our method achieves better inference speed than ScaleCrafter [14]. Under the 16× setting for SDXL, ScaleCrafter takes 577 seconds per image, while our method averages 540 seconds on a single NVIDIA A100 GPU.

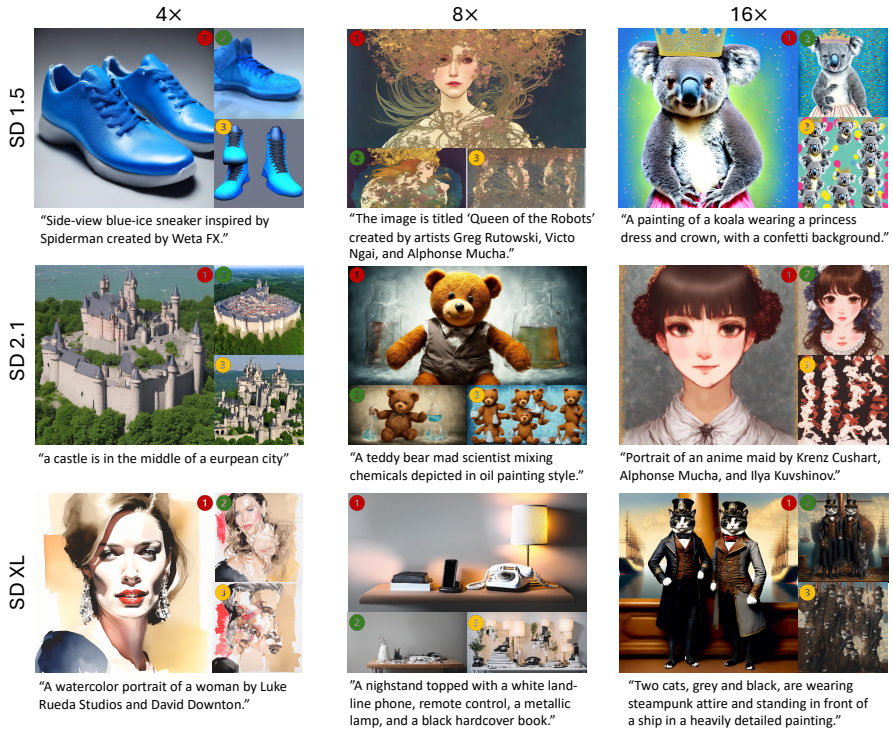


Fig. 7: Visual comparisons between ❶ ours, ❷ ScaleCrafter [14] and ❸ Attn-Entro [23], under settings of 4×, 8×, and 16×, employing three pre-trained diffusion models.

4.2 Qualitative Results

Fig. 7 presents a comprehensive visual comparison across various upscaling factors (4×, 8×, and 16×) with different pre-trained diffusion models. Our method demonstrates superior performance in preserving structural integrity and fidelity compared to ScaleCrafter [14] and Attn-Entro [23]. At 4× upscaling, FouriScale faithfully reconstructs fine details. In contrast, ScaleCrafter and Attn-Entro often exhibit blurring and loss of details. As we move to more extreme 8× and 16× upscaling factors, the advantages of FouriScale become even more pronounced. Our method consistently generates images with coherent global structures and locally consistent textures across diverse subjects. The compared methods still struggle with repetitive artifacts and distorted shapes.

4.3 Ablation Study

To validate the contributions of each component in our proposed method, we conduct ablation studies on the SD 2.1 model generating 2048 × 2048 images.

Method	FID_r
FouriScale	39.49
w/o guidance	43.75
w/o guidance & filtering	46.74

Table 2: Ablation studies on FouriScale components on SD 2.1 model under 16× 1:1 setting.

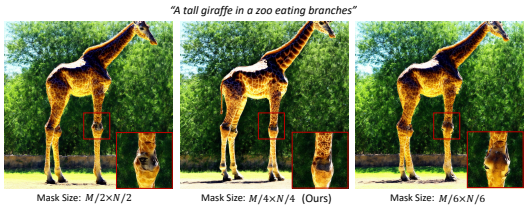


Fig. 8: Comparison of mask sizes for passing low frequencies generating 2048^2 images by SD 2.1. M, N denote height and width of target resolution.

First, we analyze the effect of using FouriScale Guidance as described in Sec. 3.5. We compare the default FouriScale which utilizes guidance versus removing the guidance and solely relying on the conditional estimation from the FouriScale-modified UNet. As shown in Tab. 2, employing guidance improves the FID_r by 4.26, demonstrating its benefits for enhancing image quality. The guidance allows incorporating structural information from the FouriScale-processed estimation to guide the generation. This balances between maintaining structural integrity and preventing loss of details.

Furthermore, we analyze the effect of the low-pass filtering operation described in Sec. 3.3. Using the FouriScale without guidance as the baseline, we additionally remove the low-pass filtering from all modules. As shown in Tab. 2, this further deteriorates the FID_r to 46.74. The low-pass filtering is crucial for maintaining scale consistency across resolutions and preventing aliasing effects that introduce distortions. Without it, the image quality degrades significantly.

A visual result of comparing the mask sizes for passing low frequencies is depicted in Fig. 8. The experiment utilizes SD 2.1 (trained with 512×512 images) to generate images of 2048×2048 pixels, setting the default mask size to $M/4 \times N/4$. We can find that the optimal visual result is achieved with our default settings. As the low-pass filter changes, there is an evident deterioration in the visual appearance of details, which underscores the validity of our method.

5 Conclusion and Limitation

We present FouriScale, a novel approach that enhances the generation of high-resolution images from pre-trained diffusion models. To address challenges such as repetitive patterns and structural distortions, FouriScale introduces a dilation operation and a low-pass filtering operation to improve structural and scale consistency among resolutions from the frequency perspective. Incorporating a padding-then-cropping strategy and FouriScale guidance further enhances the flexibility and quality of high-resolution generation, accommodating different aspect ratios while maintaining structural integrity and image fidelity. However, FouriScale still struggles with generating ultra-high-resolution samples, which typically exhibit unintended artifacts. Besides, its focus on operations within convolutions limits its applicability to purely transformer-based diffusion models.

Acknowledgement

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by Smart Traffic Fund PSRI/76/2311/PR, by RGC General Research Fund Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK.

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. arXiv preprint arXiv:2302.08113 (2023)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference on Learning Representations (2018)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR. pp. 22563–22575 (2023)
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
6. Chen, T.: On the importance of noise scheduling for diffusion models. arXiv preprint arXiv:2301.10972 (2023)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS **34**, 8780–8794 (2021)
8. Diffusion, S.: Stable diffusion 2-1 base. `https://huggingface.co/stabilityai/stable-diffusion-2-1-base/blob/main/v2-1_512-ema-pruned.ckpt` (2022)
9. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. NeurIPS **34**, 19822–19835 (2021)
10. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: Demofusion: Democratising high-resolution image generation with no $$$. In: CVPR. pp. 6159–6168 (2024)
11. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation. arXiv preprint arXiv:2306.00986 (2023)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS **27** (2014)
13. Haji-Ali, M., Balakrishnan, G., Ordonez, V.: Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In: CVPR. pp. 6603–6612 (2024)
14. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. arXiv preprint arXiv:2310.07702 (2023)
15. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
16. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2022)

17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
19. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
20. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
21. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. arXiv preprint arXiv:2301.11093 (2023)
22. Jiménez, Á.B.: Mixture of diffusers for scene composition and high resolution image generation. arXiv preprint arXiv:2302.02412 (2023)
23. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free diffusion model adaptation for variable-sized text-to-image synthesis. arXiv preprint arXiv:2306.08645 (2023)
24. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. NeurIPS **36** (2024)
25. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503 (2023)
26. Lu, Z., Wang, Z., Huang, D., Wu, C., Liu, X., Ouyang, W., Bai, L.: Fit: Flexible vision transformer for diffusion model. arXiv preprint arXiv:2402.12376 (2024)
27. Midjourney: (2024), `https://www.midjourney.com`, accessed: 17, 01, 2024
28. Pattichis, M.S., Bovik, A.C.: Analyzing image structure by multidimensional frequency modulation. IEEE TPAMI **29**(5), 753–766 (2007)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
30. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
31. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831. PMLR (2021)
32. Riad, R., Teboul, O., Grangier, D., Zeghidour, N.: Learning strides in convolutional neural networks. In: ICLR (2021)
33. Rippel, O., Snoek, J., Adams, R.P.: Spectral representations for convolutional neural networks. NeurIPS **28** (2015)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS **35**, 36479–36494 (2022)
36. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022)
37. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497 (2023)

38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
39. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint arXiv:2309.03350 (2023)
40. Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., Xu, D.: Diffusion model is secretly a training-free open vocabulary semantic segmenter. arXiv preprint arXiv:2309.02773 (2023)
41. Xiao, C., Yang, Q., Zhou, F., Zhang, C.: From text to mask: Localizing entities using the attention of text-to-image diffusion models. arXiv preprint arXiv:2309.04109 (2023)
42. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. arXiv preprint arXiv:2210.06978 (2022)
43. Zhang, R.: Making convolutional networks shift-invariant again. In: ICML. pp. 7324–7334. PMLR (2019)
44. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV. pp. 286–301 (2018)
45. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. ICCV (2023)
46. Zheng, Q., Guo, Y., Deng, J., Han, J., Li, Y., Xu, S., Xu, H.: Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. arXiv preprint arXiv:2308.16582 (2023)
47. Zhu, Q., Zhou, M., Huang, J., Zheng, N., Gao, H., Li, C., Xu, Y., Zhao, F.: Fouridown: Factoring down-sampling into shuffling and superposing. In: NeurIPS (2023)