Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder

Jinseok Kim[1,3] Tae-Kyun Kim[1,2]
[1]KAIST [2]Imperial College London
[3]AI Lab, LG Electronics

(a) Arbitrary-Scale Image Generation (b) Arbitrary-Scale Super-Resolution

Figure 1. The proposed method generates novel images and super-resolves low-resolution images at arbitrary-scales with high fidelity, diversity, and fast inference speed.

Abstract

Super-resolution (SR) and image generation are important tasks in computer vision and are widely adopted in real-world applications. Most existing methods, however, generate images only at fixed-scale magnification and suffer from over-smoothing and artifacts. Additionally, they do not offer enough diversity of output images nor image consistency at different scales. Most relevant work applied Implicit Neural Representation (INR) to the denoising diffusion model to obtain continuous-resolution yet diverse and high-quality SR results. Since this model operates in the image space, the larger the resolution of image is produced, the more memory and inference time is required, and it also does not maintain scale-specific consistency. We propose a novel pipeline that can super-resolve an input image or generate from a random noise a novel image at arbitrary scales. The method consists of a pre-trained auto-encoder, a latent diffusion model, and an implicit neural decoder, and their learning strategies. The proposed method adopts diffusion processes in a latent space, thus efficient, yet aligned with output image space decoded by MLPs at arbitrary scales. More specifically, our arbitrary-scale decoder is designed by the symmetric decoder w/o up-scaling from the pre-trained auto-encoder, and Local Implicit Image Function (LIIF) in series. The latent diffusion process is learnt by the denoising and the alignment losses jointly. Errors in output images are backpropagated via the fixed decoder, improving the quality of output images. In the extensive experiments using multiple public benchmarks on the two tasks i.e. image super-resolution and novel image generation at arbitrary scales, the proposed method outperforms relevant methods in metrics of image quality, diversity and scale consistency. It is significantly better than the relevant prior-art in the inference speed and memory usage.

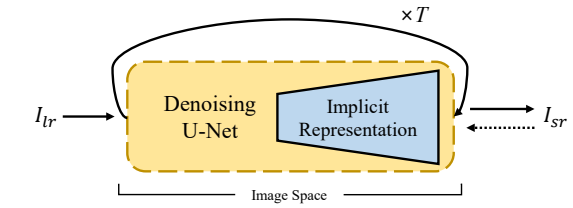
1. Introduction

Super-resolution (SR) is the task of restoring a high-resolution (HR) image by estimating the high-frequency details of an input low-resolution (LR) image. SR has applications in various fields, including medical imaging, satellite imaging, surveillance, and digital photography. It helps to enhance the visual quality of images, making them more suitable for analysis, interpretation, or presentation. SR is a

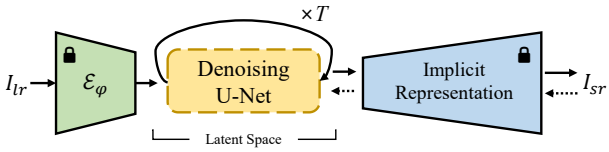
long-standing study in the field of computer vision; it is still a challenging problem. Since multiple HR images can be converted to the same LR image, obtaining the original HR image given a LR image is an 'ill-posed problem' in which there is no single correct answer. Therefore, the SR models should be able to generate diverse HR images with high perceptual quality while maintaining the coarse feature of the LR image well.

Image generation is the task of generating new images from an existing dataset. It has a wide range of applications, including data augmentation for machine learning, computer graphics, art creation, content creation for virtual environments, and more. However, high-dimensional data that look more realistic and contain fine details are required, and diverse images must be generated while maintaining high quality.

For both super-resolution and image generation tasks, several methodologies using deep neural networks have been proposed. The emergence of GAN and Diffusion Models (DMs) has brought a new paradigm to these areas. Various methods [11, 13, 14] based on them have enabled high-quality image synthesis at various scales. However, these models only work at a fixed integer scale factor($2\times$, $4\times$, $8\times$) and do not address consistency of results across different scales. As SR methods, regression-based models, such as EDSR [21], RRDB [32], RCAN [38] and SwinIR [20], learn a mapping from LR images to HR images in an end-to-end manner. Unfortunately, these models also work solely on specific trained scales. Recent methods have been developed for upscaling on continuous scales, using an implicit image function. Their MLP-based decoder takes an arbitrary query point in 2D pixel space and yields predicted pixel colors, such as Meta-SR [12], LTE [18], and LIIF [7]. However, since they often suffer from duller edges and over-smoothing details, the perceptual quality is not as high as those of fixed-scale methods. The most relevant work to this study is IDM [9]. It applies Implicit Neural Representation (INR) to the denoising diffusion model to obtain continuous-resolution yet diverse and high-quality SR results. This model, however, applies implicit representation to the denoising U-Net at each diffusion step, raising the complexity of training/inference and the need of a large memory. See Fig. 2a. The method also yields poor scale consistency. The aforementioned methods are limited to SR than image generation. Continuous scale image generation has also received an attention. Ntavelis *et al*. [23] proposed a scale-consistent positional encoding to generate images of arbitrary-scales with high fidelity based on GANs. In contrast, we adopt a diffusion probabilistic model to enable diverse yet improved output qualities. Different from IDM [9], our neural decoder is taken out of the diffusion process and the diffusion is done in a latent space(Fig. 2b). See also Section 2 for the related works.



(a) IDM



(b) Ours

Figure 2. Model structure comparison with IDM. The solid arrow represents the inference process, and the dotted arrow represents the error backpropagation process.

To address the aforementioned issues, this paper proposes a diffusion-based model for arbitrary-scale image upsampling and image synthesis simultaneously. The main contributions of this paper are as follows:

- We design a decoder that combines an auto-decoder and MLPs to generate images of arbitrary scales from the latent space.
- We introduce a two-stage model alignment process to reduce errors and misalignment between a Latent Diffusion Model (LDM) and image decoder that may occur during training.
- The proposed method enables faster and more efficient image generation compared to the other diffusion-based super-resolution model, as well as offers high fidelity and diverse output images.

2. Related Work

Diffusion Models. The diffusion model is one of the most powerful and widely used generative models. It eliminates noise from a noisy input (which usually follows a Gaussian distribution) and transforms it into sample data that follow the desired data distribution. These operations are often performed directly in the pixel space. However, recent developments have been made called Latent Diffusion Models (LDMs) [24], where LDMs perform diffusion operations in a lower-dimensional latent space, being computationally efficient and capturing complex patterns more effectively. Diffusion models enable the production of results with high quality and diversity due to its stochastic nature. Recently, diffusion-based models have been extended to various domains such as 3D data, demonstrations, and voices, and have achieved state-of-the-art results on several benchmarks [4, 5, 19, 31].

Arbitrary-Scale Image Generation. Anokhin *et al*. [1] and Skorokhodov *et al*. [29] introduced CIPS and INR-GAN respectively, in which MLPs were applied instead of using spatial convolutions commonly used in existing GAN-based models. They allow models learned on single scale to produce images on several different scales. Ntavelis *et al*. [23] proposed a scale-consistent positional encoding to generate images of arbitrary-scales with high fidelity and maintaining consistency according to the scale.

A standard diffusion model uses U-Net [25] structure consisting of a convolutional network. Therefore, as long as certain conditions are met (*e.g*. resolution in multiples of 8), there are no restrictions on the input size. Despite supporting flexible input sizes, they are trained with a fixed image size, which can lead to out-of-domain problems when generating images of different sizes. For example, a diffusion model trained with 256×256 face datasets can only generate the faces images of that size. Recently, Wang *et al*. [33] and Bar-Tal *et al*. [2] introduced the model that can generate high-quality arbitrary-resolution images by generating multiple patch images and blending them. However, they require multiple inferences to generate a single image and do not address the problem of generating the same image only at different scales, i.e. arbitrary-scale super-resolution.

Arbitrary-Scale Super-Resolution. Since the inception of MetaSR [12], there has been a surge in the exploration of various approaches for single-model arbitrary-scale super-resolution. LIIF [7] employed an implicit decoding function that takes a 2D coordinate and neighboring feature vectors as input and produces the RGB value of the corresponding pixel. However, the perceptual quality was not as good as that of high PSNR due to the duller edges and over-smoothing details. Additionally, these approaches do not provide enough diversity of output images or image consistency at different scales.

By conditioning the LR images on the diffusion models, they can be extended to super-resolution models. CDM [11] and SR3 [27] gradually upsampled LR images into HR images. However, they require training multiple networks separately at each scale. PDDPM [26] enabled upsampling at multiple scales with a single model by exploiting the positional embedding, but the aforementioned models still only worked at fixed integer scales. Recently, IDM [9] applied INR to the U-Net decoder to allow generating images on arbitrary scales. Although it has shown impressive performance, the larger the image is produced, the lower the inference speed, and the exponentially more memory is used, because it has to pass through MLPs repeatedly in each reverse process. In this paper, we applied diffusion processes on a latent space then a MLP-based decoder super-resolves the denoised latent vector to an arbitrary-scale image. The diffusion model and decoder can be separately learnt, offering a much simpler architecture.

3. Preliminary

3.1. Latent Diffusion Models

The core idea behind Latent Diffusion Models (LDMs) is to train a diffusion model in the latent space of a pre-trained auto-encoder, which allows for high-quality image synthesis with a flexible range of styles and resolutions. The diffusion process consists of a forward process and a reverse process. The forward process gradually adds random noise into the data, whereas the reverse process constructs desired data samples from the noise. The forward process is a fixed process and the noise latent z at time step t can be expressed as:

$$\mathrm{z}_t = \sqrt{\bar{\alpha}_t}\mathrm{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}), \tag{1}$$

where, $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_t)$ and $\{\beta_t\}_{t=0}^{T}$ is a variance schedule. In contrast, the reverse process is the inverse of the forward process. However, the posterior distribution $p(\mathrm{z}_{t-1}|\mathrm{z}_t)$ is intractable since it necessitates knowledge of the distribution encompassing all potential images to compute this conditional probability. Therefore, the distribution is approximated through a neural network, and can be expressed using the Bayesian theorem as follows:

$$p_\theta(\mathrm{z}_{t-1}|\mathrm{z}_t) = \mathcal{N}(\mathrm{z}_{t-1}; \mu_t(\mathrm{z}_t, \mathrm{z}_0), \sigma_t^2\mathrm{I}), \tag{2}$$

where

$$\mu_t(\mathrm{z}_t, \mathrm{z}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\mathrm{z}_t + \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right), \quad \sigma_t^2 = \beta_t. \tag{3}$$

The authors of DDPM [10] have shown that better results can be obtained by predicting noise at each time step t. Therefore, the objective function of the diffusion model $\epsilon_\theta(\cdot)$ can be expressed as:

$$L_\theta = \|\epsilon - \epsilon_\theta(\mathrm{z}_t, t)\|_2^2 \tag{4}$$

3.2. Local Implicit Image Function

Local Implicit Image Function (LIIF) [7] is a technique for representing images in a continuous way for arbitrary scale image upsampling. It uses a decoding function that takes a 2D coordinate and neighboring feature vectors as input and outputs the RGB value of the corresponding pixel. In LIIF representation, the RGB values of continuous images I at coordinates c are defined as

$$I(c) = f_\theta(\mathrm{z}^*, c^*), \tag{5}$$

where z^* is the interpolated feature vector obtained by calculating the nearest Euclidean distance from z and through the relative coordinate c^* to image domain. f_θ is a decoding function parameterized as a MLP, and shared by all images.

It also improved the quality of continuous representation through other techniques such as feature unfolding, local ensemble and cell decoding. In conclusion, LIIF leaned continuous representation by self-supervised learning and it can naturally exploit the information provided in ground-truths in different resolutions, even extrapolate to ×30 higher resolution.

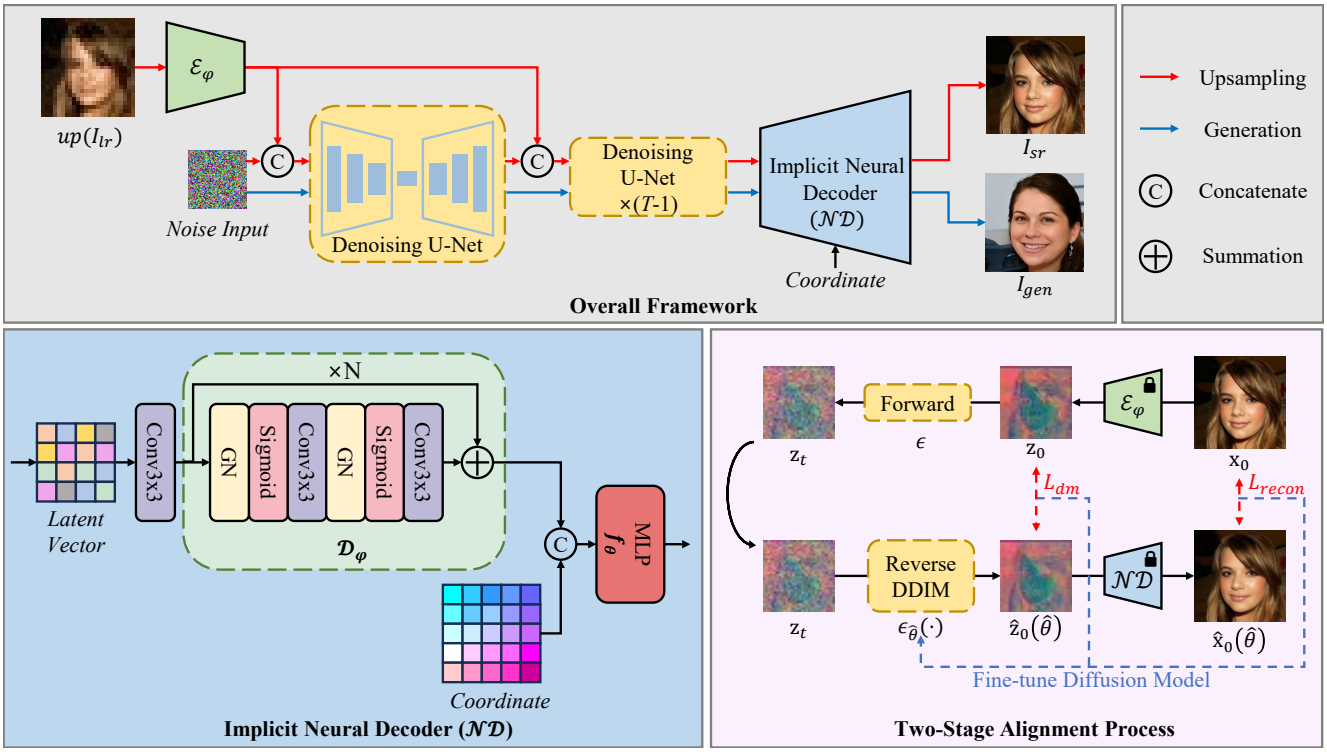


Figure 3. **Upper Part:** Overall process of proposed networks. Red line is a super-resolution process, and Blue line is a generation process. **Lower Left Part:** Detail architecture of Implicit Neural Decoder. It contains a series of auto-decoder $\mathcal{D}_\varphi$ and a neural decoding function f_θ. **Lower Right Part:** Pipeline of two-stage alignment process.

4. Method

4.1. Overview

We propose a simple architecture that combines the LDM and LIIF decoder for both arbitrary-scale SR and image generation (Fig. 3). As the encoder/decoder is fixed and the diffusion process is independently applied to the latent space in the LDM (note this significantly decreases learning complexity yet LDMs have shown outperforming other variants of DMs), we follow this stage learning strategy than end-to-end learning. An auto-encoder consisting of an encoder and a symmetric decoder w/o upsampling is pre-trained. Our implicit neural decoder combines the convolutional decoder from the auto-encoder and MLP-based decoder, that can map to arbitrary-scale output images. The diffusion process is done in a latent space, freezing the encoder and the neural decoder. The pre-trained decoder successfully super-resolves the denoised latent vector to any scale output images. In order to align the image space and latent space better, we backpropagate the image losses via the decoder to the 0-th diffusion step and any t-th step [16, 24, 28, 35]. Thus, the latent diffusion occurs with its original denoising objective plus the image loss. We experimented diverse variants of the proposed architecture, including the end-to-end learning, different decoder architectures, combinatorial loss functions (see the Supplementary material), and the proposed pipeline delivers the best results on all tasks and benchmarks. The simple architecture benefits in terms of efficient inference to arbitrary scales and learning complexity as well as image quality and diversity.

4.2. Encoder-Decoder

Our model is composed of the encoder part, denoising diffusion part, decoder part. The encoder-decoder structure follows a basic auto-encoder with convolutional and transposed convolutional neural networks, the encoder $\mathcal{E}_\varphi$ extracts the image $\mathrm{x} \in \mathbb{R}^{H\times W\times 3}$ into the latent vector $\mathrm{z} \in \mathbb{R}^{h\times w\times c}$, and the decoder part reconstructs the z back into the image space. To improve the decoding ability, the symmetric structure decoder in the auto-encoder was used (this exploits N ResBlocks where GN and Conv layers repeat), and only the upsamping layer was removed. Additionally, a MLP is combined behind it to generate arbitrary-scale images (Fig. 3).

In other words, the auto-decoder is learned to increase the amount of meaningful information by expanding the dimension of the latent feature vector, and the MLP is learned to map the latent to RGB values along with output coordinates. The feature vectors sequentially pass through the symmetric convolutional decoder and MLP and are decoded into an image. Finally, to represent the RGB values of continuous images I, we propose to modify LIIF (on Eq. (5)) as

$$I(c) = \mathcal{ND}(\mathrm{z}, c^*) = f_\theta(\mathcal{D}_\varphi(\mathrm{z}), c^*). \tag{6}$$

$\mathcal{D}_\varphi$ is the decoder network from the auto-encoder, z is a sampled latent vector by the diffusion model and c^* the relative pixel coordinate. The latent vector z is primarily decoded by the decoder $\mathcal{D}_\varphi$, and then interpolated through the Euclidean distance from c^*. f_θ is modeled as an MLP with four hidden layers consisting of 256 hidden units. Fig. 3 shows the overview of our model architecture.

4.3. Conditioning for Upsampling

In the upsampling task, the model should be able to generate the details of the HR image while maintaining the low-frequency information of the given LR image well. As conditioning information, we use the feature vector of LR image extracted by the auto-encoder E_φ. At this time, the LR image is linearly interpolated to match the size of the target latent vector. The conditioning information is strategically concatenated with intermediate feature maps at each denoising step. The LDM is learned to restore high-resolution image features from LR image features.

4.4. Two-Stage Alignment Process

The Latent Diffusion Model (LDM) is a two-stage model that has been designed to enhance the speed of both learning and inference, compared to diffusion models that operate in pixel space. The second stage model relies on the intermediate expression or features extracted by the first stage model, an auto-encoder. During the training process, the encoder can cause errors or inaccuracies, which are then passed on to the LDM. Since the LDM also produces additional errors and is trained separately from the decoder, we anticipate that these errors would reduce the effectiveness of the decoding. To address this issue, we presented a two-stage alignment process to improve the quality of the output images by reducing misalignment errors between the two-stage models.

Ultimately, what we want is to generate the output image as similar as the ground truth image as possible. Inspired by [16, 35], we want to induce an image similar to ground-truth image to be generated through the loss between the result and the ground-truth image. However, the larger the time step t, the more difficult it is to accurately predict $\hat{\mathrm{z}}_0$, so weights were given according to the time step t. The reconstruction loss with the ground truth image is defined as

$$L_{recon} = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \left\| \mathrm{x}_0 - \hat{\mathrm{x}}_0 \right\|_2^2 \tag{7}$$

where $\hat{\mathrm{x}}_0$ is generated image from the predicted latent $\hat{\mathrm{z}}_0$. By combining Eq. (1) and Eq. (4) in Sec. 3, the denosing objective of LDM can be expressed as

$$\begin{aligned} L_{dm} &= \left\| \epsilon - \epsilon_\theta \left(\mathrm{z}_t \right) \right\|_2^2 \\ &= \left\| \frac{1}{\sqrt{1-\bar{\alpha}_t}} \left(\mathrm{z}_t - \sqrt{\bar{\alpha}_t} \mathrm{z}_0 \right) - \frac{1}{\sqrt{1-\bar{\alpha}_t}} \left(\mathrm{z}_t - \sqrt{\bar{\alpha}_t} \hat{\mathrm{z}}_0 \right) \right\|_2^2 \\ &= \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \left\| \mathrm{z}_0 - \hat{\mathrm{z}}_0 \right\|_2^2 \end{aligned} \tag{8}$$

Specifically, to fine-tune the diffusion model, we modified the objective function as:

$$L_{align} = \lambda_1 L_{dm} + \lambda_2 L_{recon} \tag{9}$$

The denoising network is fine-tuned at each reverse step in a similar way to the training strategy proposed by Kim *et al*. [16]. We also adopted reverse DDIM [30] process for fast sampling. The overall flow of the proposed alignment process is shown in Fig. 3.

5. Experiment

5.1. Implementation

To train the implicit neural decoder and diffusion model, we use Adam optimizer with learning rates of 5e-5 and 1e-6, respectively. We set both λ_1 and λ_2 to 1.0. We utilized a single 24GB NVIDIA RTX 4090 GPU for all experiments.

5.2. Evaluation

Datasets. We evaluate using the datasets below:

- The Human Face Dataset contains two sub-datasets: Flick-Faces-HQ (FFHQ) [15] and CelebA-HQ [14]. These datasets consist of 70K and 30K different human face images, respectively.
- We used LSUN [36] for general scenes. The LSUN database is divided into various subcategory images, with the smaller size of image is 256x256 pixels.
- To demonstrate the upsampling potential of our model on ultra-high-resolution images, we used the wild datasets DIV2K and Flickr2K.

The image resolutions used for comparison methods are different. For more detailed setup of the training dataset, see Sec. S.2 of the supplementary material.

Metrics. We use evaluation metrics commonly used in Image Generation and Super-Resolution tasks. For image generation task, Fréchet Inception Distance(FID), Precision, and Recall are used to measure the perceptual quality and reliability of the generated image and to measure how well it covers the distribution range of real images, respectively. In addition, SelfSSIM [23] is also calculated to measure the consistency between images generated at different scales.

In SR task, PSNR is used to compare how close the upscale is to the ground truth. However, since PSNR does not capture high texture details well, it is known to not correlate well with human perception of image quality [3, 17, 37]. Therefore, we also use LPIPS [37] to compare perceptual quality with higher precision. Finally, FPS (Frames Per Second) is measured for inference speed comparison.

5.3. Comparisons on Image-Generation

Quantitative Comparisons. For the qualitative comparison of arbitrary-scale image generation performance with

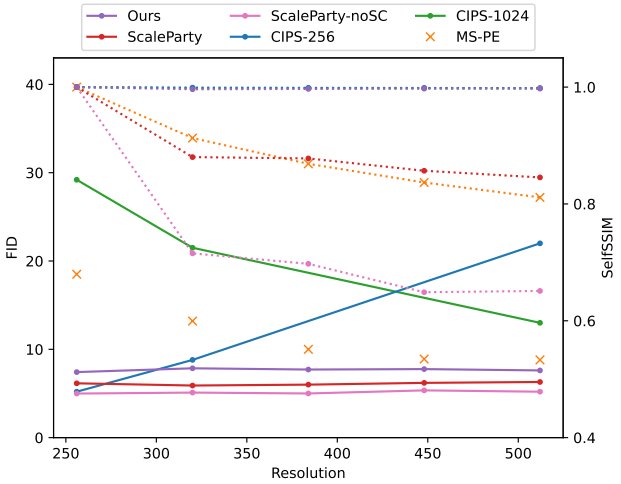


Figure 4. We compare our model quantitatively with FID and SelfSSIM scores on the FFHQ datasets. The solid lines represent the FID scores of the methods that generate images of arbitrary-scale, while the dotted lines indicate the SelfSSIM scores. The '×' symbol indicates the method that only generates images of a fixed scale. Our model demonstrates competitive performance in both evaluation metrics.

Dataset:	LSUN Bedroom						
Method	Res	FID↓	Prec↑	Rec↑	SelfSSIM (5k)↑		
MSPIE	128	11.39	**66.45**	26.97	1.00	0.10	0.10
	160	16.45	63.84	23.09	0.10	1.00	0.12
	192	12.65	58.10	25.93	0.10	0.12	1.00
Scaleparty	128	10.15	62.50	20.63	1.00	0.94	0.92
	160	9.85	**64.14**	22.02	0.92	1.00	0.95
	192	9.91	**64.77**	21.10	0.89	0.94	1.00
Ours	128	**7.20**	59.69	**38.26**	1.00	**0.98**	**0.99**
	160	**7.43**	58.52	**32.12**	**0.96**	1.00	**0.99**
	192	**7.73**	59.57	**27.98**	**0.95**	**0.97**	1.00

Table 1. Comparison of quantitative results on LSUN Bedroom datasets. For more results on other dataset, see Tab. S1 in the supplementary material.

other methods, we use arbitrary-scale generative models, CIPS [1] and ScaleParty [23], and predefined multi-scale generative modes, MSPIE [34] and MS-PE [8]. Fig. 4 and Tab. 1 show the qualitative comparison on FFHQ and LSUN Bedroom datasets, respectively. For evaluation, 50k images were sampled.

CIPS, an INR-based model, has very high consistency between images generated at each scale. However, the further away from the resolution is used for training, the lower the FID score is obtained. ScaleParty guarantees good FID scores regardless of scales. Although it also achieved a high score in terms of consistency, it clearly falls short compared to the INR-based model. Our model not only achieves good FID scores on all scales, but also shows high consistency. Moreover, our model shows much better diversity (recall) than other models, see Tab. 1.

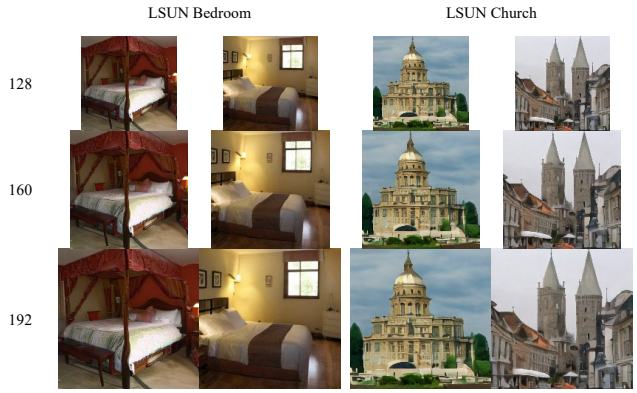


Figure 5. Qualitative results on LSUN Bedroom and Church datasets. For more results, see supplementary.

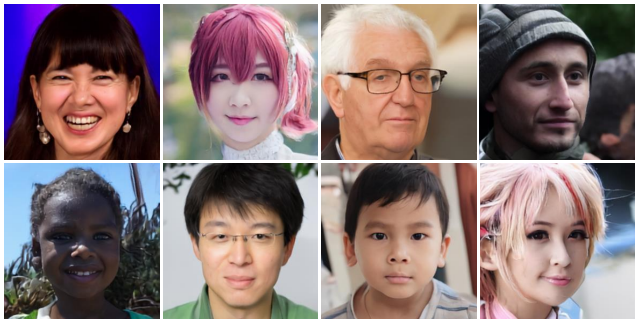


Figure 6. Various generated images on face dataset. See Fig. S3 in the supplementary to see larger and more diverse images.

Visualization. The qualitative results of our model is illustrated in Fig. 1 and Fig. 5. We selected some arbitrary scales and visualized the results. We can observe the model's effectiveness in generating diverse images, with high perceptual quality on various datasets. It also shows high scale-consistency on different scales. See Supplementary for more qualitative results.

5.4. Comparisons on Arbitrary SR

Quantitative Comparisons. Following IDM, we evaluated our model on CelebA-HQ face images. Tab. 2 shows the quantitative results with LIIF, SR3, and IDM. LIIF and IDM, as well as our model, are arbitrary scale super-resolution models, trained in the scale range (1, 8] using FFHQ datasets. Although SR3 is a model that operates only at a fixed integer magnification, it shows a lower PSNR and higher LPIPS compared to other methods. LIIF shows good PSNR scores for in-distribution scales, but still shows low perceptual quality. IDM achieved a low LPIPS score while maintaining a high PSNR. Except for the in-distribution PNSR score of LIIF, our model outperformed other methods in PSNR and LPIPS at all scales.

Similar results were obtained with the in-the-wild dataset (see Tab. 3 and Tab. 4). Our model shows superior performance compared to other generative models despite using less training data. Although regression-based methods like

Dataset:	CelebA-HQ					Dataset:	Lsun Bedroom	LSUN Tower
Method	5.3×	7×	10×	10.7×	12×	Method	16×	
LIIF [7]	**27.52** / 0.1207	**25.09** / 0.1678	22.97 / 0.2246	22.39 / 0.2276	21.81 / 0.2332	PULSE [22]	12.97 / 0.7131	13.62 / 0.7066
SR3 [27]	-	21.15 / 0.1680	20.25 / 0.2856	-	19.48 / 0.3947	GLEAN [6]	19.44 / 0.3310	18.41 / 0.2850
IDM [9]	23.34 / 0.0526	23.55 / 0.0736	23.46 / 0.1171	23.30 / 0.1238	23.06 / 0.1800	IDM [9]	**20.33** / 0.3290	19.44 / 0.2549
Ours	24.66 / **0.0455**	24.13 / **0.0690**	**23.64 / 0.1110**	**23.62 / 0.1183**	**23.52 / 0.1427**	Ours	20.08 / **0.3269**	**21.24 / 0.1897**

Table 2. Quantitative results of arbitrary-scale super-resolution on CelebA-HQ and LSUN Bedroom datasets. For each method, PSNR↑/LPIPS↓ scores are reported.

Method		Datasets	PSNR↑	SSIM↑
Reg.-based	EDSR	D+F	28.98	0.83
	LIIF	D+F	**29.00**	**0.89**
GAN-based	ESRGAN	D+F	26.22	0.75
	RankSRGAN	D+F	26.55	0.75
Flow-based	SRFlow	D+F	27.09	0.76
Flow+GAN	HCFlow++	D+F	26.61	0.74
Diffusion	IDM	D	27.10	0.77
	IDM	D+F	27.59	0.78
	Ours	D	**27.61**	**0.81**

Table 3. Quantitative comparison of 4× super-resolution using in-the-wild datasets. D and F refer to the DIV2k and Flickr2k.

Method	8×	12×	17×
LIIF	**23.97** / 0.4790	22.28 / 0.5900	21.23 / 0.6560
Ours	23.82 / **0.4265**	**22.73 / 0.5463**	**21.83 / 0.6225**

Table 4. Comparison (PSNR↑ / LPIPS↓) on the DIV2K at out-of-distribution scales.

EDSR and LIIF have shown high scores on 4×, Tab. 4 indicates that our approach surpasses LIIF at larger scale. Additionally, our model, based on diffusion models, offers the benefits of diversity and scale consistency.

Visualization. We visualize the upsampling results of several face datasets to demonstrate qualitative comparisons on arbitrary-scale upsampling in Fig. 7. A very low-resolution image does not contain enough features, making it very difficult to up-sample. Despite these difficulties, LIIF restores a face with an expression very similar to that of the ground-truth image. However, the output image is over-smoothed, and detailed textures such as hair are not sufficiently restored. IDM generates images very realistically and delicately. Although high consistency is maintained across scales, differences in facial expressions are still visible. In contrast, our model not only maintains high fidelity and perceptual quality, but also produces consistent images well without significant distortion even at arbitrary large scales, see Fig. 8.

Inference Speed and Memory Usage. Given that diffusion models typically suffer from slow inference speeds and high memory usage compared to other methods, we have intentionally designed our model to alleviate these shortcomings. We compared the inference speed of each model by measuring FPS (Frames Per Second) and Tab. 5 shows

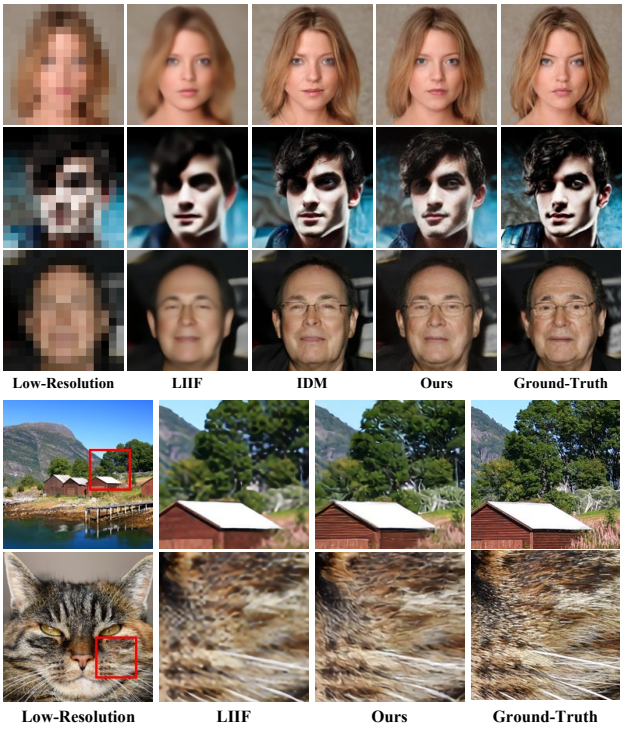


Figure 7. Qualitative comparisons of arbitrary-scale upsampling on face (**upper**) and in-the-wild (**lower**) dataset. Our model restored more detailed information and generated faces with similar expressions to ground-truth images compared to other models.

the results. MLP-based implicit networks require a long time compared to other network structures. In IDM, an implicit network must be passed at every denoising step, which leads to significant slowdown. Additionally, since a high-dimensional denoising process is required to generate a large image, the larger the scale is, the more severely the inference speed decreases. In contrast, our model compensates the speed by using the implicit network only once during the decoding process. As a result, our model is able to inference faster than IDM (about 12.7 times faster on 8× task), while showing better output quality. Furthermore, since IDM operates in the pixel space, it was not possible to generate very large-scale images such as 200× in our environment (Single RTX 4090), but our model can generate them by adjusting the input batch size of the MLP. Our model is still slower than other GAN-based SR models or regression models, but it shows a fast inference speed compared to other diffusion-based models.

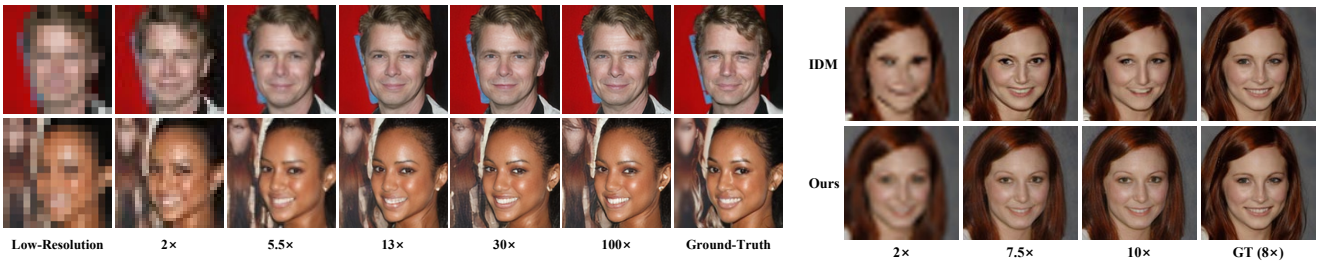


Figure 8. **Left:** Qualitative results of the proposed method for arbitrary-scale upsampling on face datasets. **Right:** Comparison of scale consistency on face dataset.

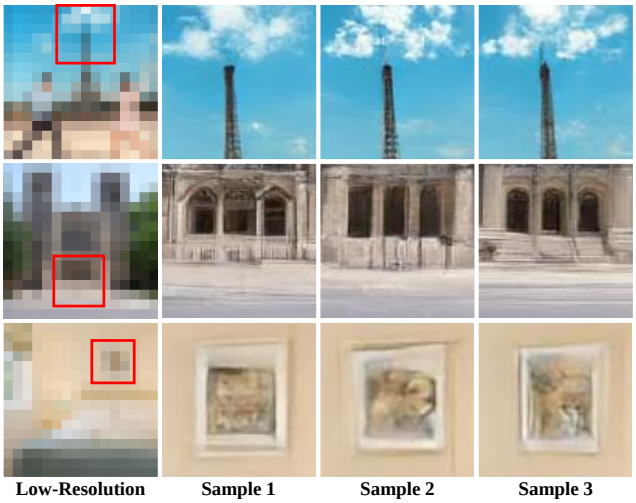


Figure 9. Visualization of result diversity in super-resolution tasks. This allows our model to better handle the 'ill-posed problem'. See Fig. S9 in the supplementary to see larger and more diverse results.

Method	FPS↑				
	8×	12×	30×	100×	200×
IDM	0.0202	0.0200	6.54e-3	3.98e-4	-
Ours	**0.2568**	**0.2510**	**0.2473**	**0.1833**	**0.0982**

Table 5. Comparison of inference Speed in terms of FPS (Frames Per Second). '-' indicates that the model did not work in our environment due to memory overflow.

	Encoder-Decoder (PSNR↑)	
	16 → 128 (8×)	64 → 256 (4×)
MLP	22.96	32.54
AE+MLP	**35.28**	**35.93**

	Aligment Process (PSNR↑ / LPIPS↓)			
	5.3×	7×	10×	12×
Before	22.79 / 0.0600	22.49 / 0.0932	23.23 / 0.1691	22.14 / 0.2077
After	**24.66 / 0.0455**	**24.13 / 0.0690**	**23.64 / 0.1110**	**23.52 / 0.1427**

Table 6. Ablation studies of decoder structures and alignment process.

Figure 10. Comparison of qualitative results before and after the two-stage alignment process.

5.5. Ablation Studies

Decoder Architecture. Tab. 6 shows the reconstruction capabilities according to the decoder structures. Rather than directly reconstructing the extracted features using MLPs, its reconstruction quality is better when using the symmetric decoder from the pre-trained auto-encoder.

Alignment Process. Tab. 6 and Fig. 10 present the quantitative and qualitative results of the two-stage alignment process, respectively. To demonstrate the effectiveness of the two-stage alignment process, we compare the results before and after the process. Overall, artifacts are reduced and textures are more realistic.

6. Conclusions

In this paper, we proposed an Implicit Neural Decoder with a latent diffusion model for arbitrary-scale image generation and upsampling. We show that our Implicit Neural Decoder can effectively reconstruct latent features into continuous-scale images. This has enabled our model to train and infer efficiently in the latent space. As a result, our method is an order-of-magnitude faster compared to IDM. Furthermore, we also proposed a two-stage alignment process to improve the quality of output images by reducing misalignment errors between the two-stage models. We achieved high-diversity image generation and high-fidelity super-resolution at arbitrary scales while maintaining perceptual quality and scale consistency.

Acknowledgements. This work was in part supported by NST grant (CRC 21011, MSIT), KOCCA grant (R2022020028, MCST), IITP grant (RS-2023-00228996, MSIT), and LG Electronics Co., Ltd.

References

[1] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *CVPR*, pages 14278–14287, 2021. 3, 6

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 3

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 5

[4] Rumeysa Bodur, Erhan Gundogdu, Binod Bhattarai, Tae-Kyun Kim, Michael Donoser, and Loris Bazzani. iedit: Localised text-guided image editing with weak supervision. *arXiv:2305.05947*, 2023. 2

[5] Rumeysa Bodur, Binod Bhattarai, and Tae-Kyun Kim. Prompt augmentation for self-supervised text-guided image manipulation. In *CVPR*, 2024. 2

[6] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021. 7

[7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 2, 3, 7

[8] Jooyoung Choi, Jungbeom Lee, Yonghyun Jeong, and Sungroh Yoon. Toward spatially unbiased generative models. In *ICCV*, pages 14253–14262, 2021. 6

[9] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, pages 10021–10030, 2023. 2, 3, 7

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[11] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021. 2, 3

[12] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *CVPR*, 2019. 2, 3

[13] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *CVPR*, 2020. 2

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 2, 5

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 5

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435, 2022. 4, 5

[17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 5

[18] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *CVPR*, pages 1929–1938, 2022. 2

[19] Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, and Tae-Kyun Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *CVPR*, 2024. 2

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 2

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2

[22] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 7

[23] Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, and Luc Van Gool. Arbitrary-scale image synthesis. In *CVPR*, pages 11533–11542, 2022. 2, 3, 5, 6

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 2015. 3

[26] Dohoon Ryu and Jong Chul Ye. Pyramidal denoising diffusion probabilistic models. *arXiv preprint arXiv:2208.01864*, 2022. 3

[27] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv:2104.07636*, 2021. 3, 7

[28] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 4

[29] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. Adversarial generation of continuous images. In *CVPR*, pages 10753–10764, 2021. 3

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 5

[31] Tao Wang, Kaihao Zhang, Ziqian Shao, Wenhan Luo, Bjorn Stenger, Tae-Kyun Kim, Wei Liu, and Hongdong Li. Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement. *arXiv:2307.14659*, 2023. 2

[32] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: En-

hanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 2

[33] Yinhuai Wang, Jiwen Yu, Runyi Yu, and Jian Zhang. Unlimited-size diffusion restoration. In *CVPRW*, pages 1160–1167, 2023. 3

[34] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. In *CVPR*, pages 13569–13578, 2021. 6

[35] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 4, 5

[36] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[38] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2

Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder

Supplementary Material

In this Supplementary Material, we describe the evaluation metrics in more details. In addition, we present more qualitative results for comparing other models and visualizing the quality of output images.

S.1. Metrics

FID [5]. Fréchet inception distance is a metric used for evaluating the quality of generated images produced by generative models. It measures the similarity between the distribution of real images and the distribution of generated images by computing the Fréchet distance in feature space.

Precision and Recall [8]. Precision and recall are proposed metrics to evaluate fidelity and diversity. Precision refers to the ratio of the generated image to the real image distribution and refers to the precision of how accurately the generated image depicts the real image sample. Recall refers to the ratio of the actual image to the distribution of the generated image sample and refers to the diversity of the generated image.

SelfSSIM [11]. It is a metric for evaluating the scale consistency of the generated images. We downsample two images of different resolutions to a lower resolution and then measure the SSIM [12] between them.

PSNR. Peak Signal-to-Noise Ratio (PSNR) is a widely used metric to quantify the quality of a reconstructed image compared to the original image. A higher PSNR score indicates a lower distortion. It suggests that the processed image is closer to the original in terms of pixel-wise similarity. However, recent research shows that this metric has limitations in indicating actual perceptual quality [1, 9, 14].

LPIPS [14]. Learned Perceptual Image Patch Similarity (LPIPS) is a metric created to measure the similarity between image patches from a perceptual standpoint. It has been demonstrated to accurately reflect human perception. A low LPIPS score indicates that the patches are perceptually similar.

S.2. Experiment Details

In this section, we describe the target scale at which our model and the comparison models were trained on each task of the experiment.

S.2.1. Image-Generation

For the results in the comparative experiments, we referred to the results of Ntavelis *et al*. [11]. Note ScaleParty, MSPIE, MS-PE were trained at larger resolutions, e.g. over 256×256 and 128×128 for FFHQ and LSUN respectively, while our method was trained at less than those resolutions.

FFHQ [7]. For the human face generation task, each model generated images at five different scales, 256, 320, 384, 448 and 512. The training policy for each model is as follows.

- MS-PE is trained for every scale in comparison (*i.e*. 256, 320, 384, 448 and 512), since it is a multi-scale generation model.
- CIPS is trained on single scale, 256.
- ScaleParty is trained with two different resolutions, 256 and 384, for its scale consistency approach.
- Our model is trained to generate an image of arbitrary resolution between (32, 256] from a latent vector of 32.

LSUN [13]. For generic scene (bedroom and church) generation tasks, each model generated images at three different scales, 128 160 and 192. The training policy for each model is as follows.

- MSPIE and ScaleParty are trained for 128 and 192.
- Our model is trained to generate an image of arbitrary resolution between (64, 128] from a latent vector of 64.

S.2.2. Super-Resolution

In the super-resolution operation, 16×16 low-resolution images are upsampled to arbitrary scales. All models were trained within a scale range of 8× for human faces and 16× for generic scenes.

S.3. More Results

S.3.1. Quantitative Results

Tab. S1 show additional quantitative results of image generation for LSUN Church. In the generation task, the maximum resolution of the images used by our model for training is lower than that of other models, as mentioned in Sec. S.2.1. Nevertheless, as shown in Fig. 4 and Tab. 1 of the main text and Tab. S1, our model shows competitive results. In particular, our model shows great strengths in terms of diversity and scale consistency. And in the super-resolution task, our model achieves significantly better performance not only in terms of fidelity but also in terms of perceptual quality. All methods were trained at the same scales for the super-resolution task.

S.3.2. Qualitative Results

To demonstrate the performance of our model, we provide more provide more generated images and comparison

Dataset:	LSUN Church						
Method	Res	FID↓	Prec↑	Rec↑	SelfSSIM (5k)↑		
MSPIE	128	**6.67**	**71.95**	44.59	1.00	0.32	0.43
	160	10.76	66.21	36.95	0.31	1.00	0.40
	192	**6.02**	66.70	46.13	0.39	0.38	1.00
Scaleparty	128	9.08	70.52	39.93	1.00	0.95	0.93
	160	**7.96**	**70.87**	32.07	0.94	1.00	0.95
	192	7.52	**68.14**	33.33	0.90	0.94	1.00
Ours	128	8.25	65.27	**47.02**	1.00	**0.98**	**0.98**
	160	8.58	64.02	**43.04**	**0.97**	1.00	**0.99**
	192	8.81	62.36	**42.80**	**0.96**	**0.97**	1.00

Table S1. Quantitative comparison of image generation on LSUN Church datasets.

results. In Figs. S1 to S5, we visualize various randomly sampled results for the FFHQ, LSUN-Bedroom and LSUN-Church datasets, respectively. Our model shows remarkable performance in synthesizing high-quality details with a variety of styles and scale-consistency.

Figs. S6 to S8 show the qualitative comparison of *SR* for CelebA-HQ [6], LSUN-Bedroom and LSUN-Tower, respectively. LIIF has over-smoothing issues in contrast to high PSNR scores. Both IDM and our model are good at capturing high-resolution details, and furthermore, our model has achieved relatively few distortions. In addition, Fig. S9 shows various *SR* results for the LSUN datasets. The top image is an *LR* image, and the images below are different *SR* results in the red area. As the scale increases, the number of high-resolution solutions that can be recovered from low-resolution becomes more diverse. However, INR-based models such as LIIF [3] always achieve only the same results. In contrast, our stochastic model can generate a variety of patterns and textures for blankets, clouds, and buildings, etc. while maintaining *LR* information. This allows our model to better handle the 'ill-posed problem'.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 1

[2] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021.

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 2

[4] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, pages 10021–10030, 2023.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 1

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 2

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[8] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019. 1

[9] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1

[10] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020.

[11] Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, and Luc Van Gool. Arbitrary-scale image synthesis. In *CVPR*, pages 11533–11542, 2022. 1

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[13] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[14] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

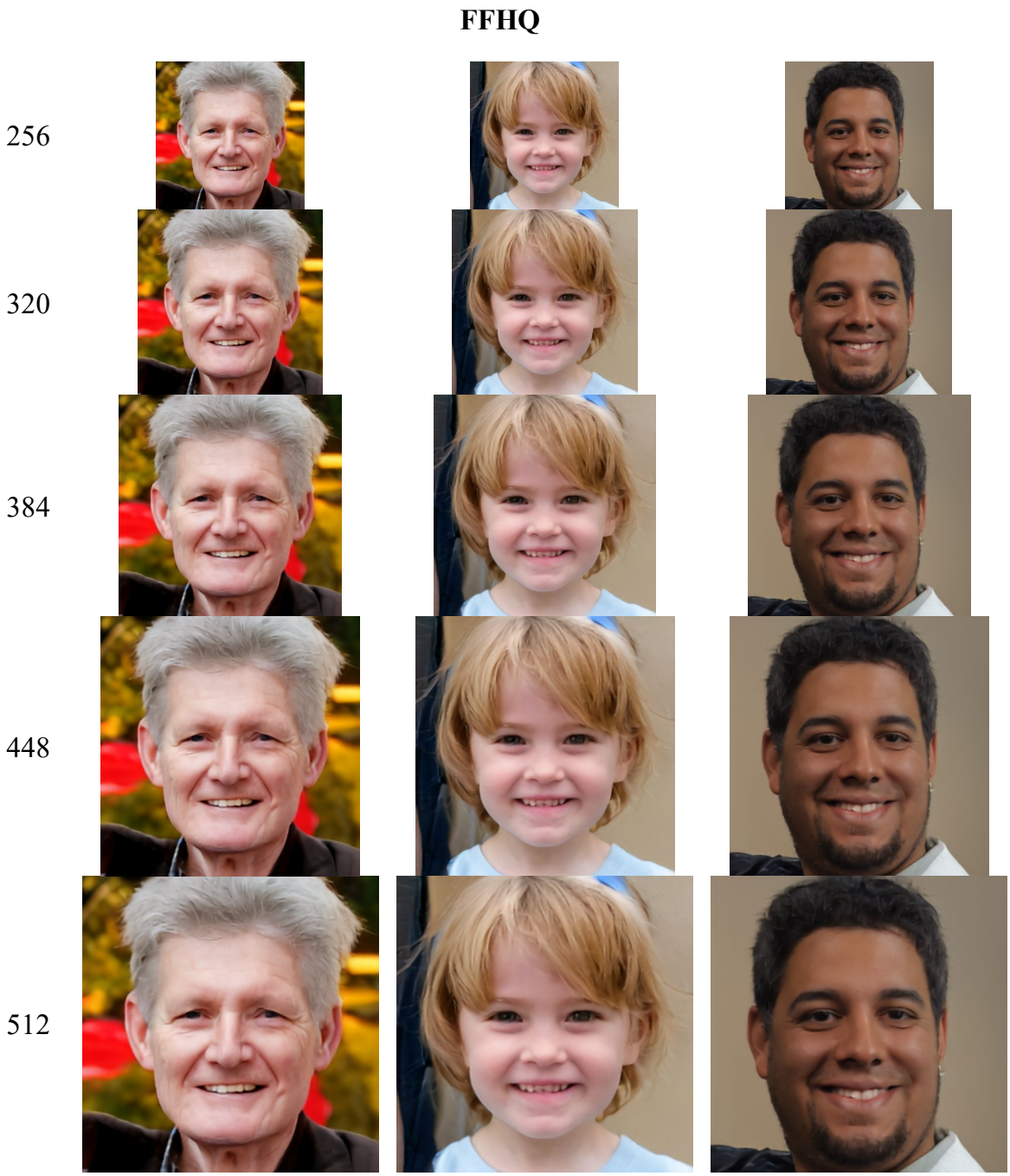


Figure S1. Scale consistency results of image generation on the FFHQ datasets.

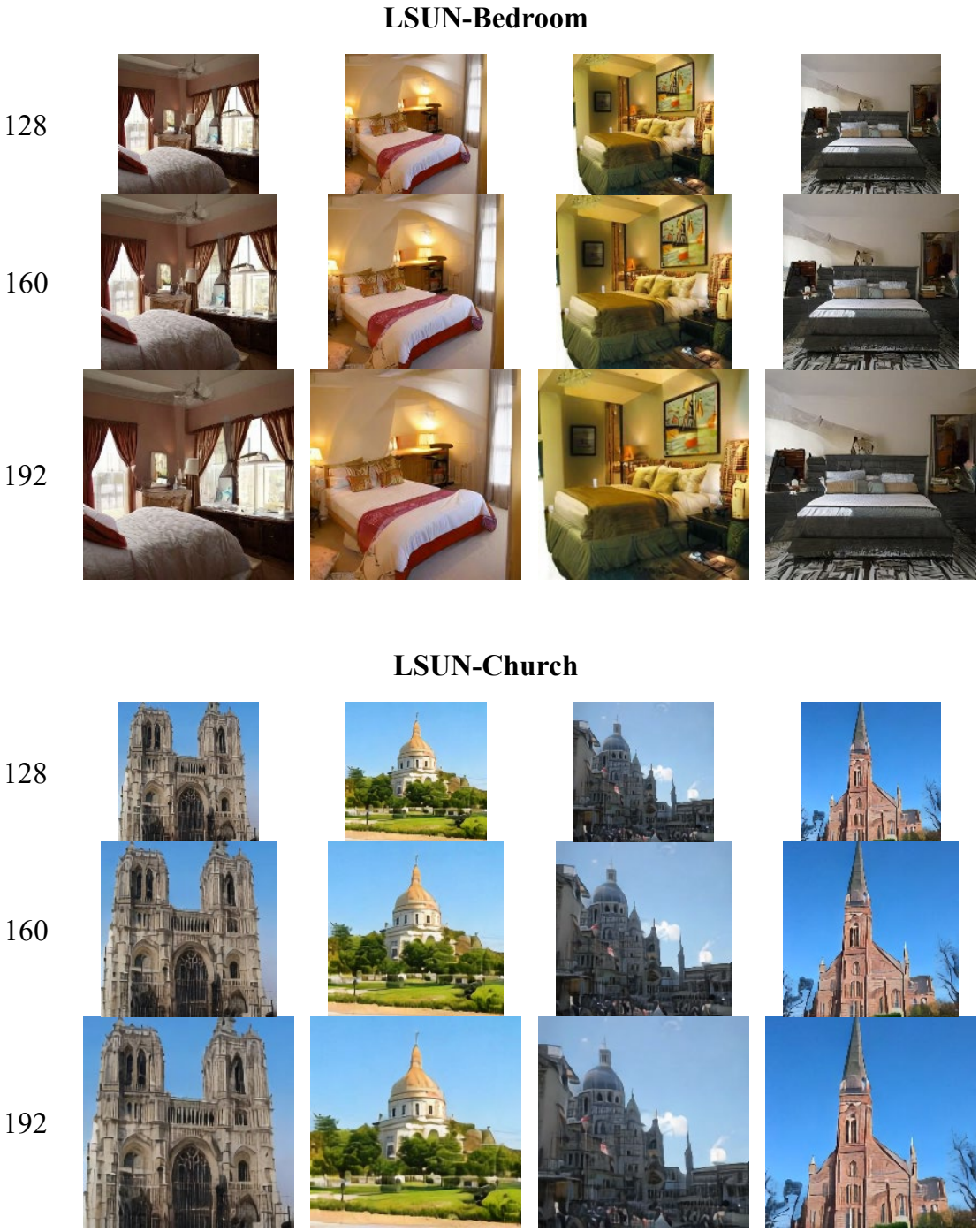


Figure S2. Scale consistency results of image generation on the LSUN Bedroom, Church datasets.

FFHQ

Figure S3. Visual results of image generation on the FFHQ datasets.

LSUN-Bedroom

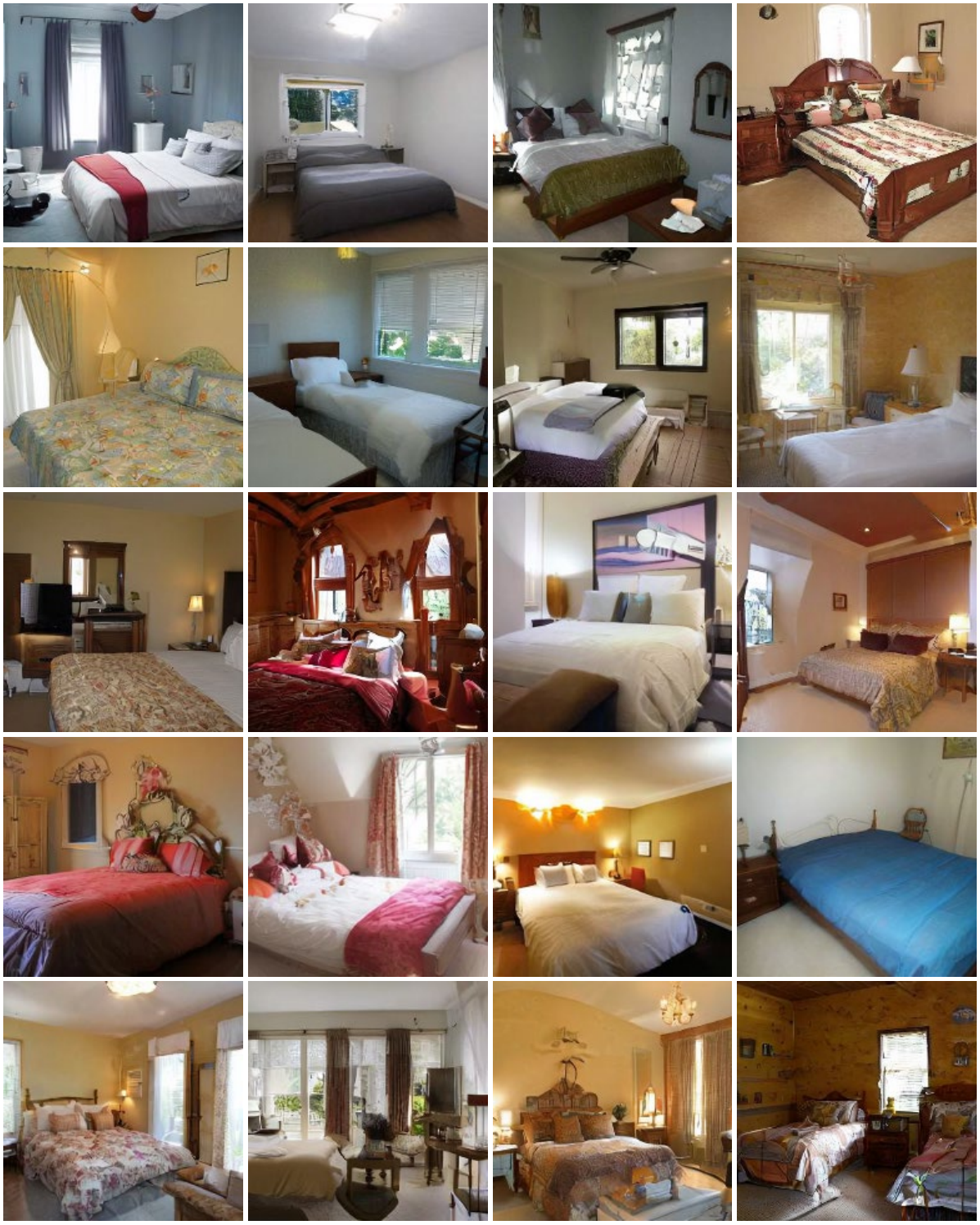


Figure S4. Visual results of image generation on the LSUN Bedroom datasets.

LSUN-Church

Figure S5. Visual results of image generation on the LSUN Church datasets.

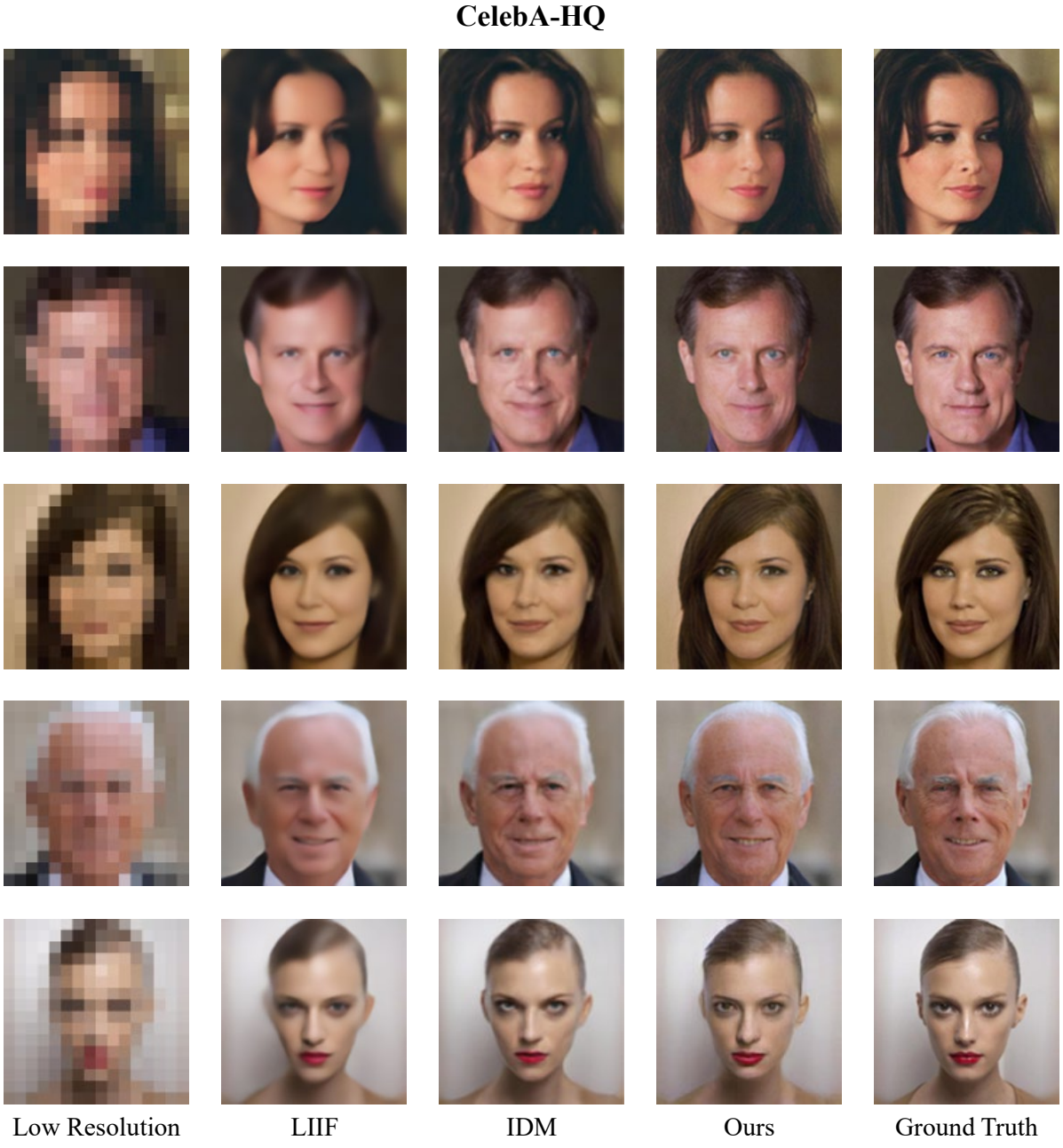


Figure S6. Comparison of $16 \times 16 \rightarrow 128 \times 128$ super-resolution on the CelebA-HQ datasets.

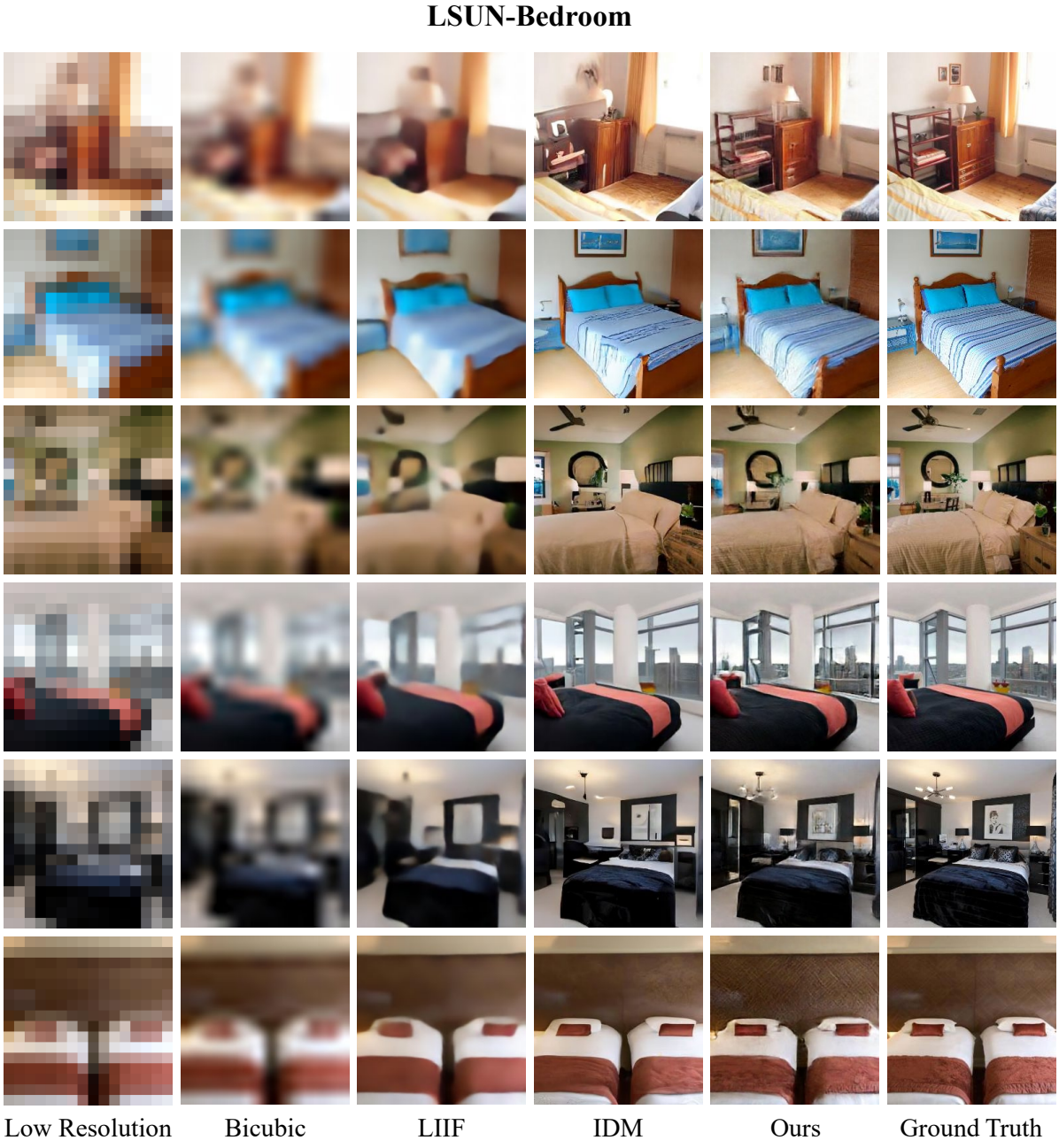


Figure S7. Comparison of $16 \times 16 \rightarrow 256 \times 256$ super-resolution on the LSUN Bedroom datasets.

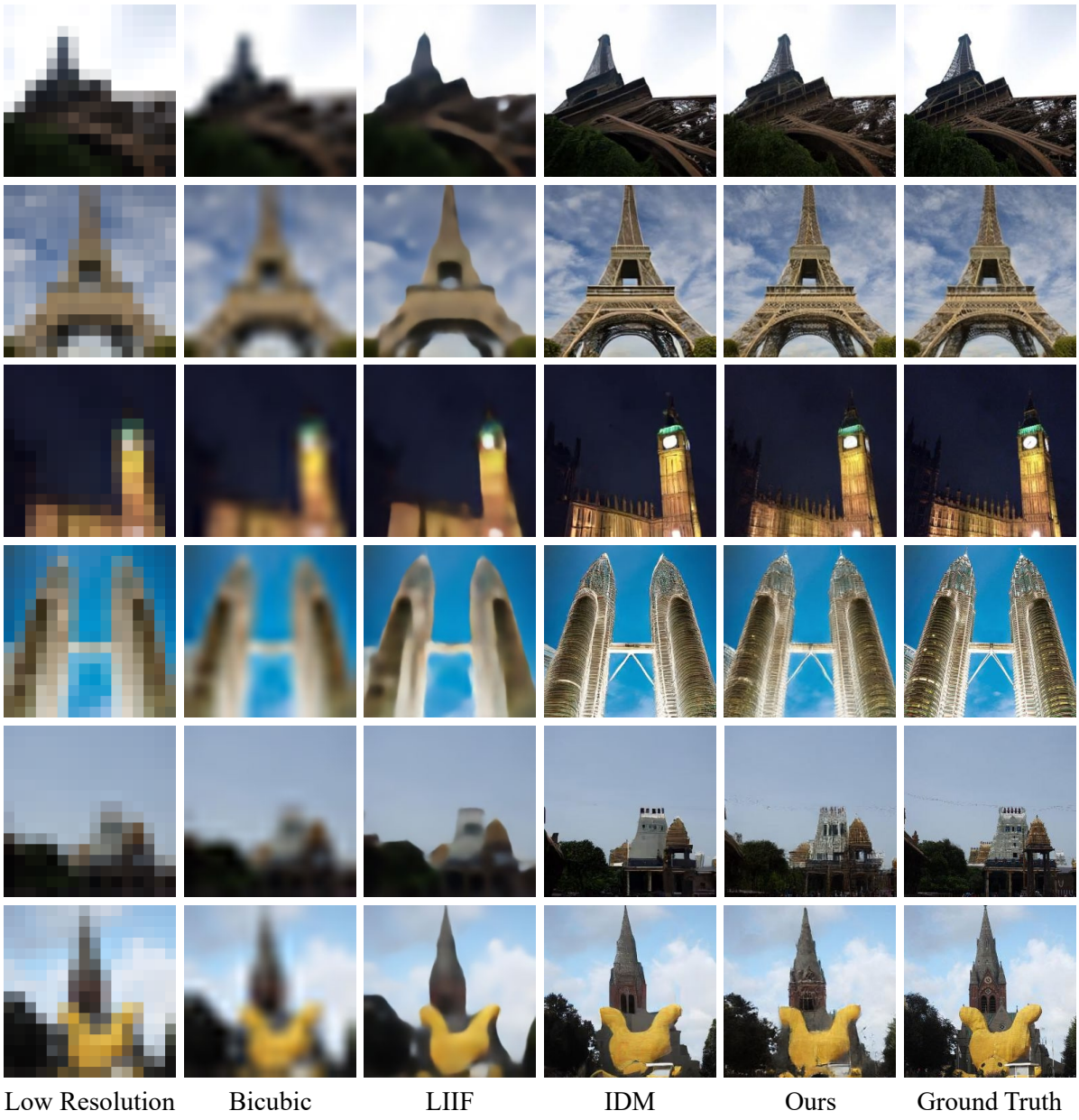


Figure S8. Comparison of $16 \times 16 \rightarrow 256 \times 256$ super-resolution on the LSUN Tower datasets.

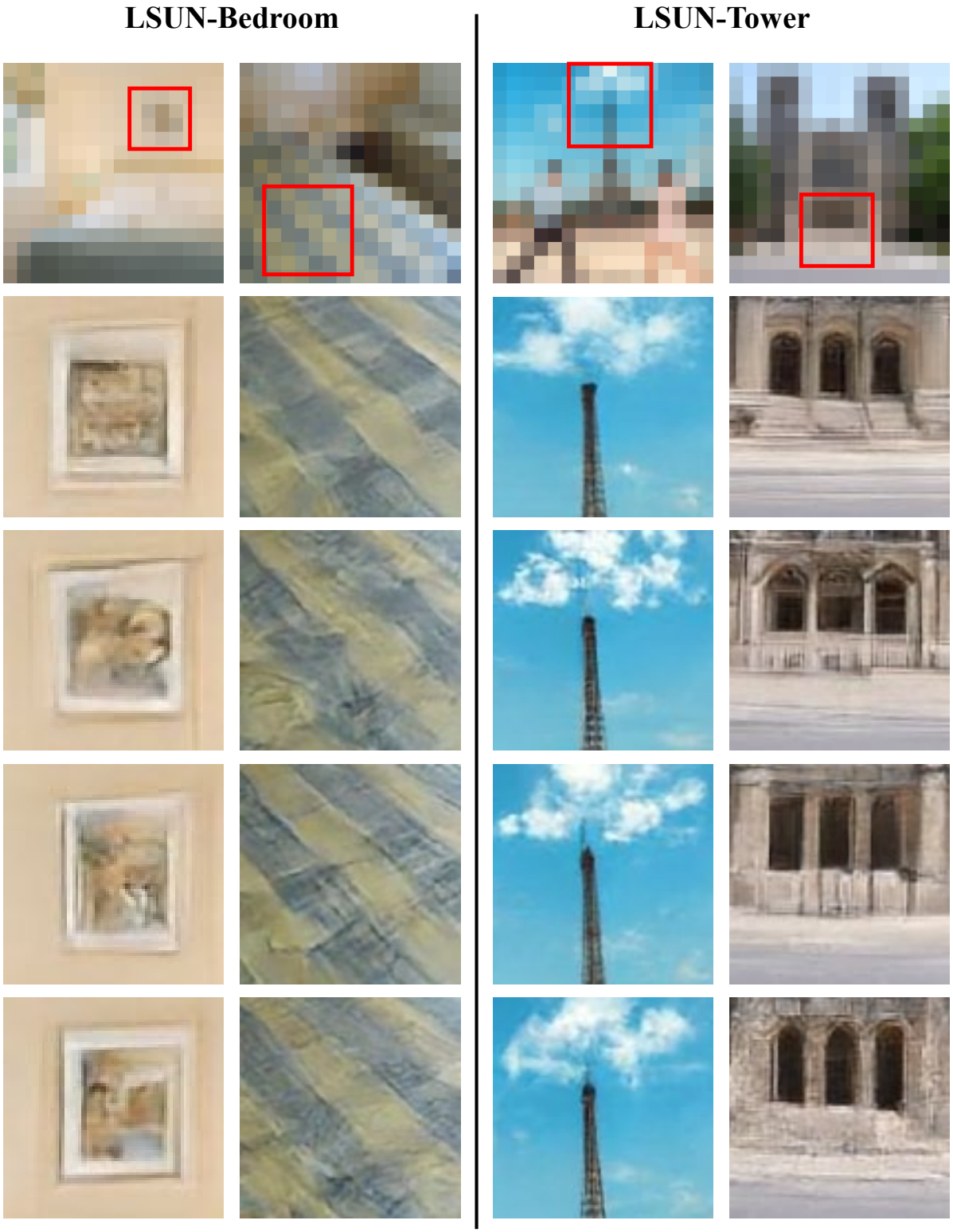


Figure S9. Visualization results of diversity in super-resolution tasks. The top image is an *LR* image, and the images below are different *SR* results in the red area.