Implicit Diffusion Models for Continuous Super-Resolution

Sicheng Gao[1*], Xuhui Liu[1*], Bohan Zeng[1*], Sheng Xu[1], Yanjing Li[1], Xiaoyan Luo[1]
Jianzhuang Liu[2], Xiantong Zhen[3], Baochang Zhang[1,4†]
[1]Beihang University [2]Shenzhen Institute of Advanced Technology, Shenzhen, China
[3]United Imaging [4]Zhongguancun Laboratory, Beijing, China

Abstract

Image super-resolution (SR) has attracted increasing attention due to its widespread applications. However, current SR methods generally suffer from over-smoothing and artifacts, and most work only with fixed magnifications. This paper introduces an Implicit Diffusion Model (IDM) for high-fidelity continuous image super-resolution. IDM integrates an implicit neural representation and a denoising diffusion model in a unified end-to-end framework, where the implicit neural representation is adopted in the decoding process to learn continuous-resolution representation. Furthermore, we design a scale-adaptive conditioning mechanism that consists of a low-resolution (LR) conditioning network and a scaling factor. The scaling factor regulates the resolution and accordingly modulates the proportion of the LR information and generated features in the final output, which enables the model to accommodate the continuous-resolution requirement. Extensive experiments validate the effectiveness of our IDM and demonstrate its superior performance over prior arts. The source code will be available at `https://github.com/Ree1s/IDM`.

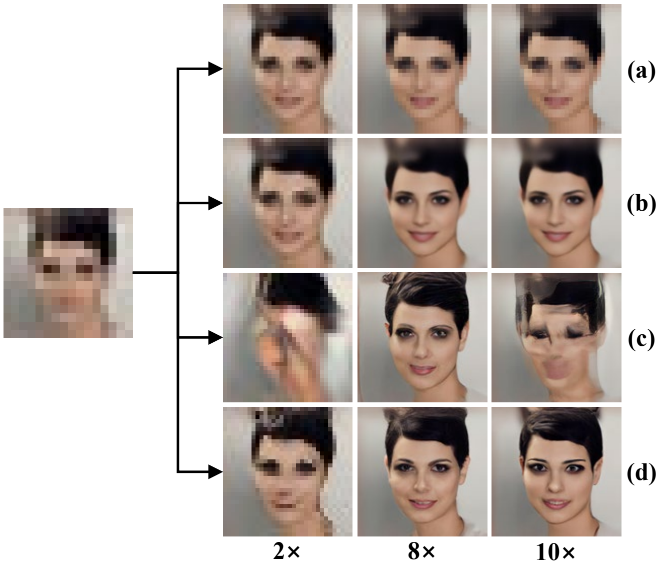


Figure 1. Visual comparison, where training is on 8× SR and testing on 2×, 8×, and 10×. (a) EDSR [23] and (b) LIIF [6] are regression-based models; (c) SR3 [35] and (d) IDM (ours) are generative models. Among them, LIIF and IDM employ the implicit neural representation.

1. Introduction

Image super-resolution (SR) refers to the task of generating high-resolution (HR) images from given low-resolution (LR) images. It has attracted increasing attention due to its far-reaching applications, such as video restoration, photography, and accelerating data transmission. While significant progress has been achieved recently, existing SR models predominantly suffer from suboptimal quality and the requirement for fixed-resolution outputs, leading to undesirable restrictions in practice.

Regression-based methods [21, 23] offer an intuitive way to establish a mapping from LR to HR images. LIIF [6] specifically achieves resolution-continuous outputs through implicit neural representation. However, these methods often fail to generate high-fidelity details needed for high magnifications (see Fig. 1(a) and (b)) since their regression losses tend to calculate the averaged results of possible SR predictions. Deep generative models, including autoregressive [30, 43], GAN-based [15, 16, 18, 26], flow-based [8, 24] and variational autoencoders (VAEs) [17, 42], have emerged as solutions that enrich detailed textures. Still, they often exhibit artifacts and only apply to pre-defined fixed magnifications. Despite the ability to generate realistic images with high perceptual quality with the help of extra priors, GAN-based models are subject to mode collapse and struggle to capture complex data distributions, yielding unnatural textures. Recently, Diffusion Probabilistic Models (DMs) [12, 39] have been used in image synthesis to improve the fidelity of SR images and have shown impressive performance. Nonetheless, DM-based methods are still limited to fixed magnifications, which would result in corrupted output once the magnification changes (see Fig. 1(c)). There-

[*] These authors contributed equally.

[†] Corresponding Author: bczhang@buaa.edu.cn.

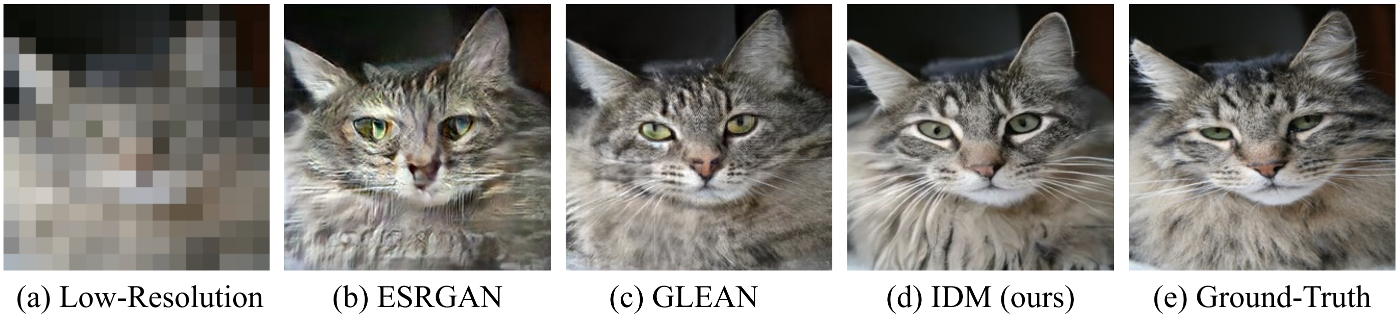


Figure 2. Examples of 16 × super-resolution. (a) LR input. (b) ESRGAN [45] which trains a simple end-to-end structure GAN, and loses the inherent information. (c) GLEAN [4] which achieves more realistic details through additional StyleGAN [16] priors, but still generates unnatural textures and GAN-specific artifacts. (d) With implicit continuous representation based on a scale-adaptive conditioning mechanism, IDM generates the output with high-fidelity details and retains the identity of the ground-truth. (e) The ground-truth.

fore, they turn to a complicated cascaded structure [13] or two-stage training strategies [10, 33, 34] to achieve multiple combined magnifications, or retrain the model for a specific resolution [35], which brings extra training cost.

To address these issues, this paper presents a novel Implicit Diffusion Model (IDM) for high-fidelity image SR across a continuous range of resolutions. We take the merit of diffusion models in synthesizing fine image details to improve the fidelity of SR results and introduce the implicit image function to handle the fixed-resolution limitation. In particular, we formulate continuous image super-resolution as a denoising diffusion process. We leverage the appealing property of implicit neural representations by encoding an image as a function into a continuous space. When incorporated into the diffusion model, it is parameterized by a coordinate-based Multi-Layer Perceptron (MLP) to capture the resolution-continuous representations of images better.

At a high level, IDM iteratively leverages the denoising diffusion model and the implicit image function, which is implemented in the upsampling layers of the U-Net architecture. Fig. 1(d) illustrates that IDM achieves continuously modulated results within a wide range of resolutions. Accordingly, we develop a scale-adaptive conditioning mechanism consisting of an LR conditioning network and a scaling factor. The LR conditioning network can encode LR images without priors and provide multi-resolution features for the iterative denoising steps. The scaling factor is introduced for controlling the output resolution continuously and works through the adaptive MLP to adjust how much the encoded LR and generated features are expressed. It is worth noting that, unlike previous methods with two-stage synthesis pipelines [9, 13, 33] or additional priors [4, 26, 44], IDM enjoys an elegant end-to-end training framework without extra priors. As shown in Fig. 2, we can observe that IDM outperforms other previous works in synthesizing photographic image details.

The main contributions of this paper are summarized as follows:

- We develop an Implicit Diffusion Model (IDM) for continuous image super-resolution to reconstruct photo-realistic images in an end-to-end manner. Iterative implicit denoising diffusion is performed to learn resolution-continuous representations that enhance the high-fidelity details of SR images.
- We design a scale-adaptive conditioning mechanism to dynamically adjust the ratio of the realistic information from LR features and the generated fine details in the diffusion process. This is achieved through an adaptive MLP when size-varied SR outputs are needed.
- We conduct extensive experiments on key benchmarks for natural and facial image SR tasks. IDM exhibits state-of-the-art qualitative and quantitative results compared to the previous works and yields high-fidelity resolution-continuous outputs.

2. Related Work

Implicit Neural Representation. In recent years, implicit neural representations have shown extraordinary capability in modeling 3D object shapes, synthesizing 3D surfaces of the scene, and capturing complicated 3D structures [3, 27–29, 36–38]. Particularly, methods based on Neural Radiance Fields (NeRF) [2, 28] utilize Multi-Layer Perceptrons (MLPs) to render 3D-consistent images with refined texture details. Because of its outstanding performance in 3D tasks, implicit neural representations have been extended to 2D images. Instead of parameterizing 2D shapes with an MLP with ReLU as in early works [31, 40], SIREN [37] employs periodic activation functions to model high-quality image representations with fast convergence. LIIF [6] significantly improves the performance of representing natural and complex images with local latent code, which can restore images in an arbitrary resolution. However, the high-resolution results generated by LIIF are constrained by prior LR information, resulting in over-smoothing with high-frequency information lost.

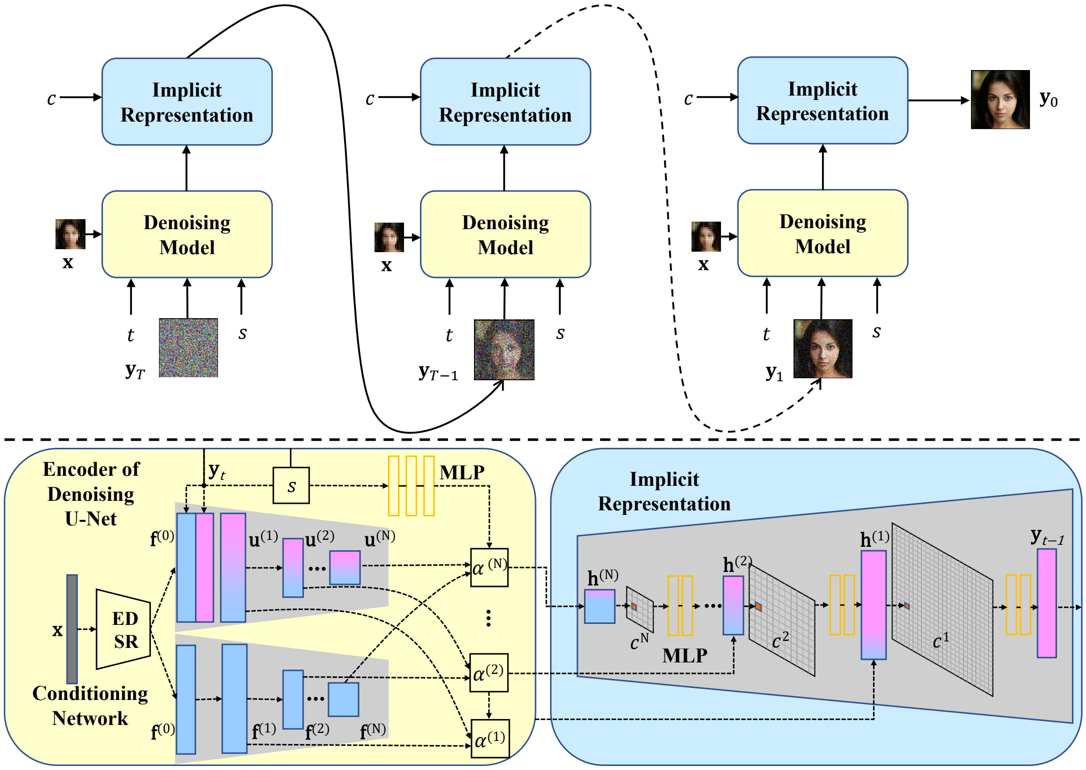


Figure 3. Overview of the IDM framework. **Upper Part:** Overall process of the inference. **Lower Part:** Detailed illustration of a denoising step, where the U-Net decoder is omitted for conciseness.

In our method, we introduce the denoising diffusion model to yield realistic details missed by LIIF while retaining the superiority of the implicit continuous image function. Based on the controllable scaling factor, IDM can dynamically maintain a balance between the LR information and generated fine details while meeting the size-varied requirement of output images.

Generative Image Super-Resolution Models. In image super-resolution, regression-based methods, such as EDSR [23], RRDB [45], and SWinIR [21], directly learn a mapping from LR to HR images with an MSE loss. Based on these algorithms, [6, 14, 19] further achieve continuous image super-resolution with meta-learning or implicit neural representation. While impressive PSNR results have been shown, they often suffer from duller edges and over-smoothing details in perceptual outputs. On the other hand, GAN-based and flow-based models, variational autoencoders (VAEs), and autoregressive models (ARMs) have been proposed to improve the fine details of SR images. SRGAN [18] uses an adversarial loss and the perceptual loss [48], rather than a pixel-wise loss (*e.g.*, L2 loss), to optimize the output. SFTGAN [51] and GLEAN [4] design new structures to fuse semantics and StyleGAN [16] priors to generate rich and realistic texture features. Moreover, flow-based models [22, 24] and VAEs [17, 42] introduce normalization flow and stochastic variational inference into image generation, respectively, but their sample quality underperforms GAN-based methods. Despite the strong performance in learning complex distributions, ARMs [30, 43] are limited to low-resolution images because of the high training cost and sophisticated sequential sampling process.

Recently, Diffusion Probabilistic Models (DMs) [12] have shown state-of-the-art results in image and speech synthesis [5], and time series forecasting [32], for example. Likewise, some diffusion frameworks have been applied to low-level vision tasks. For example, SR3 exhibits impressive performance on image SR after repeated refinement, LDM [34] employs the cross-attention conditioning mechanism for generating high-resolution images, and CDM [13] introduces the class condition and a cascaded structure to achieve realistic multi-resolution results. However, some drawbacks of existing models remain to be solved, including but not limited to unnatural artifacts, fixed magnification ratios, and complicated two-stage pipelines. In this paper, IDM combines the merits of diffusion models and implicit neural representations in a practical end-to-end framework, thereby obtaining photo-realistic SR images with continuous resolutions.

3. Method

This section presents the IDM approach, a simple end-to-end framework with an effective scale-adaptive conditioning mechanism and an implicit diffusion process, to generate high-fidelity resolution-continuous outputs. The architecture of IDM is shown in Fig 3.

3.1. Problem Statement

Given an LR-HR image pair denoted as $(\mathbf{x}_i, \mathbf{y}_i)$ and a scaling factor s, where $\mathbf{x}_i$ is degraded from $\mathbf{y}_i$ and s controls the resolution of the output in a continuous manner, IDM aims to learn a parametric approximation to the data distribution $p(\mathbf{y} \mid \mathbf{x})$ through a fixed Markov chain of length T. Following [35], we define the forward Markovian diffusion process q by adding Gaussian noise as:

$$\begin{aligned} q\left(\mathbf{y}_{1:T} \mid \mathbf{y}_0\right) &= \prod_{t=1}^{T} q\left(\mathbf{y}_t \mid \mathbf{y}_{t-1}\right), \\ q\left(\mathbf{y}_t \mid \mathbf{y}_{t-1}\right) &= \mathcal{N}\left(\mathbf{y}_t \mid \sqrt{1-\beta_t}\mathbf{y}_{t-1}, \beta_t \mathbf{I}\right), \end{aligned} \tag{1}$$

where $\beta_t \in (0,1)$ are the variances of the Gaussian noise in T iterations. Given $\mathbf{y}_0$, the distribution of $\mathbf{y}_t$ can be represented by:

$$q\left(\mathbf{y}_t \mid \mathbf{y}_0\right) = \mathcal{N}\left(\mathbf{y}_t \mid \sqrt{\gamma_t}\mathbf{y}_0, (1-\gamma_t)\mathbf{I}\right), \tag{2}$$

where $\gamma_t = \prod_{i=1}^{t}(1-\beta_i)$. In the inverse diffusion process, IDM learns the conditional distributions $p_\theta\left(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}\right)$ to denoise the latent features sequentially during training. Formally, the inference process can be conducted as a reverse Markovian process from Gaussian noise $\mathbf{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a target image $\mathbf{y}_0$ as:

$$\begin{aligned} p_\theta\left(\mathbf{y}_{0:T} \mid \mathbf{x}\right) &= p\left(\mathbf{y}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}\right), \\ p\left(\mathbf{y}_T\right) &= \mathcal{N}\left(\mathbf{y}_T \mid \mathbf{0}, \mathbf{I}\right), \\ p_\theta\left(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}\right) &= \mathcal{N}\left(\mathbf{y}_{t-1} \mid \mu_\theta\left(\mathbf{x}, \mathbf{y}_t, t\right), \sigma_t^2 \mathbf{I}\right). \end{aligned} \tag{3}$$

As shown in Fig. 3, we adopt a U-Net architecture as the denoising model similar to the vanilla DDPM [12] that encodes the noisy image $\mathbf{y}_t$ into multi-resolution feature maps $\mathbf{u}^{(i)}$, where $i \in \{1, \cdots, N\}$, and N is the number of depths in the U-Net backbone. Meanwhile, we introduce the implicit image function in the decoding part of the U-Net to generate realistic resolution-continuous images. IDM unifies the iterative diffusion refinement process and the implicit image function in an end-to-end framework.

3.2. Scale-Adaptive Conditioning Mechanism

LR Conditioning Network. Inspired by GLEAN [4] and GCFSR [11], we utilize a CNN, which is stacked by convolutional layers with a bilinear filtering downsampling operation and a leaky ReLU [25] activation, as the conditioning network to extract conditioning features in multiple resolutions from the LR image. To accomplish this, we first employ EDSR [23] to establish the initial LR feature $\mathbf{f}^{(0)}$ and make its resolution the same as $\mathbf{y}_t$'s by bilinear interpolation. Then, we concatenate $\mathbf{f}^{(0)}$ and $\mathbf{y}_t$ and feed the result into the U-Net for preliminary conditional guidance. Meanwhile, $\mathbf{f}^{(0)}$ is also sent to the CNN, where the feature is progressively downsampled as:

$$\mathbf{f}^{(i)} = \operatorname{Conv}\left(\mathbf{f}^{(i-1)}\right), \tag{4}$$

where Conv denotes the convolution layer with the bilinear filtering downsampling operation and a leaky ReLU activation. Unlike GAN-based methods that rely on additional priors [4, 44], our conditioning network only provides encoded multi-resolution features. It sends them into the U-Net without extra priors to model latent representations.

Scaling Factor Modulation. To lift the limitation of fixed magnification ratios, we introduce a scaling factor s as a condition for the diffusion process to enable magnification with continuous resolution. We first define an interval $(1, M]$, where M is the maximum magnification ratio, and randomly select s from the interval during training. We then reshape $\mathbf{y}_t$ according to s to control the resolution of generated images, as shown in the yellow part of Fig. 3. The scaling factor s is used to adjust the ratio of the original input information $\mathbf{f}^i$ from the conditioning network and the output $\mathbf{u}^i$ from the denoising network. As shown at the bottom of Fig. 3, unlike the cross-attention mechanism [34] and the concatenating operation [20], we map s to a set of scaling vectors $\alpha = \left\{\alpha_1^{(1)}, \alpha_2^{(1)}, \ldots, \alpha_1^{(i)}, \alpha_2^{(i)}, \ldots, \alpha_1^{(N)}, \alpha_2^{(N)}\right\}$ with an adaptive MLP, where i represents the depth index with different resolution outputs from the conditioning network and the denoising network. Next $\alpha_1^{(i)}$ and $\alpha_2^{(i)}$ are normalized by the L2 norm, and then used to modulate $\mathbf{f}^{(i)}$ and $\mathbf{u}^{(i)}$ channel-wisely and fuse them adaptively. In general, we conduct the modulation process with the scaling factor s as follows:

$$\alpha = \mathit{Reshape}(\mathrm{MLP}(s)), \tag{5}$$

$$\bar{\alpha}_1^{(i)} = \frac{\left|\alpha_1^{(i)}\right|}{\sqrt{{\alpha_1^{(i)}}^2 + {\alpha_2^{(i)}}^2 + \delta}}, \tag{6}$$

$$\bar{\alpha}_2^{(i)} = \frac{\left|\alpha_2^{(i)}\right|}{\sqrt{{\alpha_1^{(i)}}^2 + {\alpha_2^{(i)}}^2 + \delta}}, \tag{7}$$

$$\mathbf{h}^{(i)} = \bar{\alpha}_1^{(i)} \cdot \mathbf{f}^{(i)} + \bar{\alpha}_2^{(i)} \cdot \operatorname{Concat}\left({\mathbf{u}_{\text{up}}}^{(i)}, {\mathbf{u}_{\text{down}}}^{(i)}\right), \tag{8}$$

where $\delta = 1e-8$ to avoid zero denominators, and ${\mathbf{u}_{\text{up}}}^{(i)}$ and ${\mathbf{u}_{\text{down}}}^{(i)}$ are the feature maps from the decoder and encoder of the U-Net, respectively. The modulation result $\mathbf{h}^{(i)}$ is shown in Fig. 3.

Table 1. Datasets used in our experiments.

	Training	Testing
Human faces	FFHQ [16]	CelebA-HQ [15]
General scenes	DIV2K [1]	DIV2K-validate [1]
Cats	LSUN-train [47]	CAT [50]
Bedrooms	LSUN-train [47]	LSUN-validate [47]
Towers	LSUN-train [47]	LSUN-validate [47]

3.3. Implicit Neural Representation

Considering that prevailing SR methods are often burdened by a complicated cascaded pipeline [13] or two-stage training strategies [33,34] to produce outputs with multiple resolutions, we innovate the implicit neural representation to learn continuous image representations, simplifying IDM. As shown in the blue box in Fig. 3, we insert several coordinate-based MLPs into the upsampling of the U-Net architecture to parameterize the implicit neural representations, which can restore LR images with high-fidelity quality in a continuous scale range. Like LIIF [6], with the assumed continuous coordinates of multi-resolution features $c = \left\{c^{(1)}, \ldots, c^{(i)}, \ldots, c^{(N)}\right\}$ as a reference, which is obtained from the denoising network using the scaling factor s, we input the current features around the coordinates and then calculate the target features. Given the features $\mathbf{h}^{(i+1)}$ and the corresponding coordinates $c^{(i+1)}$, we formulate the implicit representation process as follows:

$$\mathbf{u}_{\text{up}}^{(i)} = D_i\left(\hat{\mathbf{h}}^{(i+1)}, c^{(i)} - \hat{c}^{(i+1)}\right), \qquad (9)$$

where D_i is a 2-layer MLP with hidden dimensionality 256, and $\hat{\mathbf{h}}^{(i+1)}$ and $\hat{c}^{(i+1)}$ are interpolated by calculating the nearest Euclidean distance from $\mathbf{h}^{(i+1)}$ and $c^{(i+1)}$ in the $(i+1)$-th depth, respectively.

3.4. Optimization

IDM aims to infer the target image $\mathbf{y}_0$ with a sequence of denoising steps. To this end, we optimize the denoising model ϵ_θ which is equivalent to restoring the target image $\mathbf{y}_0$ from a noisy target image $\tilde{\mathbf{y}}_t = \sqrt{\gamma_t}\mathbf{y}_0 + \sqrt{1-\gamma_t}\epsilon$. Meanwhile, to achieve resolution-continuous outputs, the denoising model $\epsilon_\theta\left(\mathbf{x}, t, s, \tilde{\mathbf{y}}_t, \gamma_t\right)$ should apply to arbitrary scales through training while ensuring the validity of the predicted noise ϵ. To conclude, we optimize the denoising network with

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\epsilon,\gamma_t,t,s}\left\|\epsilon - \epsilon_\theta(\mathbf{x}, t, s, \tilde{\mathbf{y}}_t, \gamma_t)\right\|_1^1, \qquad (10)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \{1, \cdots, T\}$, $s \sim \mathcal{U}(1, M]$, and $(\mathbf{x}, \mathbf{y})$ is sampled from the training set of LR-HR image pairs.

4. Experiments

In addition to the extensive experiments described in this section, we also provide more results with more magnifications and resolutions in the supplementary materials.

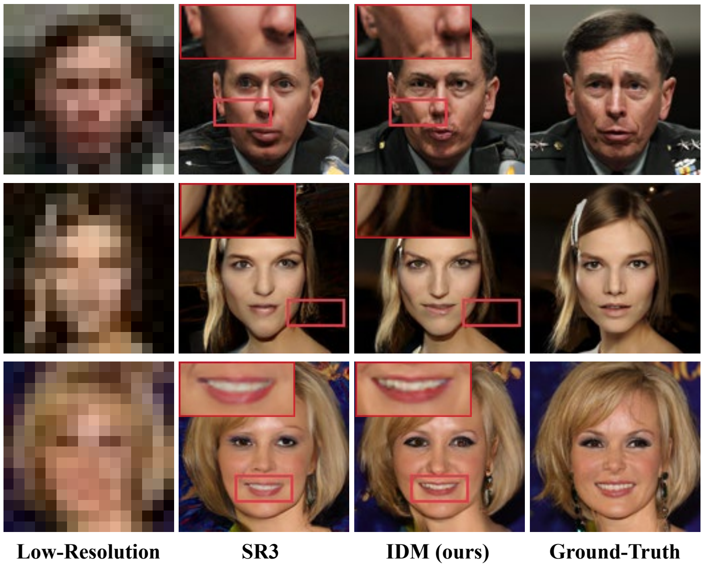


Figure 4. Qualitative comparison on 8× SR on CelebA-HQ [15]. The results of IDM maintain higher fidelity and more credible identities close to the ground-truth, generating more realistic facial components (*e.g.,* eyes, teeth, and hair).

4.1. Implementation Details

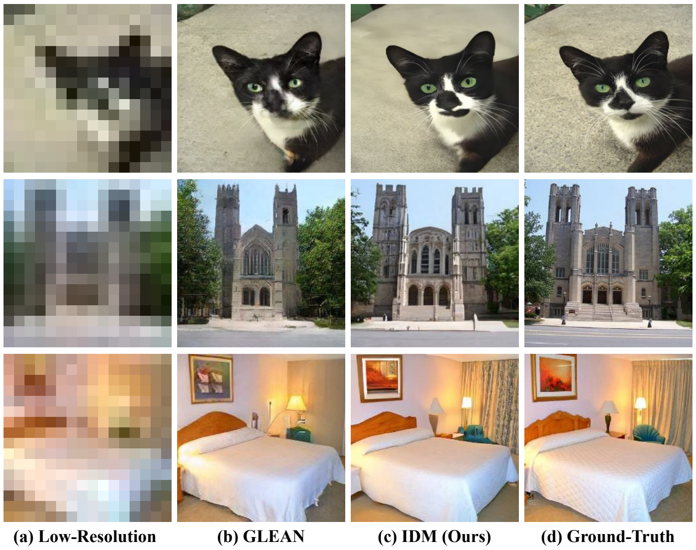
Datasets. We conduct our experiments on face datasets, natural image datasets, and a general scene dataset (DIV2K [1]), which are listed in Table 1. For face datasets, we train and evaluate IDM on FFHQ [16] and CelebA-HQ [15], respectively, which is the same as with SR3 [35]. We use DIV2K for general scenes to compare with various state-of-the-art (SOTA) methods based on other generative models. Finally, similar to GLEAN [4], we train and test our model on the LSUN [47] dataset.

Training Details. We train our IDM in an end-to-end manner. We set a milestone to 1M iterations, where the training is with a fixed downsampling scale to $M\times$, and after the milestone, the training is conducted for 0.5 M iterations with HR images randomly resized according to the uniform distribution $\mathcal{U}(1, M)$. Following the vanilla DDPM [12], we use the Adam optimizer with a fixed learning rate of 1e-4 for the former and 2e-5 for the latter. We utilize a dropout rate of 0.2 and two 24GB NVIDIA RTX A5000 GPUs for all experiments.

4.2. Qualitative Comparisons

We conduct qualitative comparisons with SOTA methods on both face and natural image SR.

Face Super-Resolution. Fig. 4 shows the qualitative comparison with the SOTA DM-based method (SR3) on the 16×16 → 128×128 face SR task. Although both SR3 and IDM can improve the diversity of generated outputs, SR3 loses many face attributes, so it is quite different between the identities of



(a) Low-Resolution (b) GLEAN (c) IDM (Ours) (d) Ground-Truth

Figure 5. Results of 16× SR on the LSUN dataset. IDM achieves more consistent textures with the ground-truth.

Table 2. Quantitative comparison (PSNR and SSIM) with several baselines on $16\times16 \rightarrow 128\times128$ face super-resolution. Consistency measures the MSE $(\times10^{-5})$ between LR and downsampled SR images.

Method	PSNR↑	SSIM↑	Consistency↓
PULSE [26]	16.88	0.44	161.1
FSRGAN [7]	23.01	0.62	33.8
Regression [35]	23.96	0.69	2.71
SR3 [35]	23.04	0.65	2.68
IDM	**24.01**	**0.71**	**2.14**

Table 3. Quantitative comparison (PSNR and LPIPS) on LSUN [47] with 16× SR.

Method	Cats	Bedrooms	Towers
PULSE [26]	19.78/0.5241	12.97/0.7131	13.62/0.7066
ESRGAN+ [45]	19.99/0.3482	19.47/0.3291	17.86/0.3132
GLEAN [4]	20.92/0.3215	19.44/0.3310	18.41/0.2850
IDM	**21.52/0.3131**	**20.33/0.3290**	**19.44/0.2549**

Table 4. Quantitative comparison of 4× SR on the DIV2K [1] validation set. D+F means the training datasets include both DIV2K and Flicker2K [41], and D means that IDM is only trained on DIV2K. Red and blue colors indicate the best and the second-best performance among generative models, respectively.

Method		Datasets	PSNR↑	SSIM↑
Bicubic		D+F	26.7	0.77
Reg.-based	EDSR [23]	D+F	28.98	0.83
	LIIF [6]	D+F	29.00	0.89
GAN-based	ESRGAN [45]	D+F	26.22	0.75
	RankSRGAN [49]	D+F	26.55	0.75
Flow-based	SRFlow [24]	D+F	27.09	0.76
	HCFlow [22]	D+F	27.02	0.76
Flow+GAN	HCFlow++ [22]	D+F	26.61	0.74
VAE+AR	LAR-SR [10]	D+F	27.03	0.77
Diffusion	IDM	D	27.10	0.77
Diffusion	IDM	D+F	27.59	0.78

the SR3 outputs and the ground-truth. For instance, the teeth and eyes are discrepant, and the wrinkles are not retained. In contrast, IDM maintains the identities and high-fidelity face details.

Natural Image Super-Resolution. Fig. 5 shows the qualitative comparison with the SOTA GAN-based method (GLEAN) on the 16× natural image SR task, including Cats, Towers, and Bedrooms. We directly take the examples provided in GLEAN for comparison. Although GLEAN can produce realistic SR results, it does not perform well on some detailed textures, such as the nose and eyes in the first row, the windows and doors in the send row, and the curtain, wall picture, and two lamps in the last row. IDM is more effective in reconstructing them and exhibits excellent fine details.

4.3. Quantitative Comparisons

Face Super-Resolution. Following SR3, we evaluate IDM on 100 face images extracted from CelebA-HQ [15] and compute the PSNR, SSIM [46], and Consistency metrics. Table 2 shows the PSNR, SSIM, and Consistency results on the 8× face super-resolution task. GAN-based models are up to par with human perception when the super-resolution magnification is large [4]. Nevertheless, their poor Consistency values show that their SR results deviate from the LR images. Compared with the diffusion model-based SR3, IDM obtains better results in all metrics (0.97 dB higher in PSNR, 0.06 higher in SSIM, and 0.53 lower in Consistency).

Natural Image Super-resolution. To demonstrate the performance of IDM on natural image SR, we provide the quantitative comparison in Table 3. We select 100 images in the validation dataset and compute the average PSNR and LPIPS [48]. Because PULSE generates SR objects incorrectly, its PSNR and LPIPS are significantly lower than other methods. Although GLEAN achieves better results with the pretrained latent banks, IDM outperforms it in all categories, with 0.60 dB, 0.89 dB, and 1.03 dB improvements in PSNR, respectively. Likewise, IDM decreases LPIPS in all categories whose SR results align more with human perception.

4.4. Comparison on a General Scene Dataset

We conduct comprehensive comparisons with various prior arts, including regression-based and generative methods, on the general scene dataset DIV2K. EDSR and LIIF are trained with the pixel-wise loss. Fig. 7 shows the qualitative comparison with LIIF on the 4× general scene SR task. LIIF generates

Table 5. Quantitative comparison (PSNR/LPIPS) of continuous SR results on CelebA-HQ [15] when training on 8× LR-HR pairs. Each method is trained on 8× face SR. " – " indicates the model is completely invalid with the magnification.

Method	in-distribution		out-of-distribution		
	5.3×	7×	10×	10.7×	12×
LIIF [6]	**27.52**/0.1207	**25.09**/0.1678	22.97/0.2246	22.39/0.2276	21.81/0.2332
SR3 [35]	–	21.15/0.1680	20.26/0.2856	–	19.48/0.3947
IDM	23.34/**0.0526**	23.55/**0.0736**	**23.46/0.1171**	**23.30/0.1238**	**23.06 /0.1800**

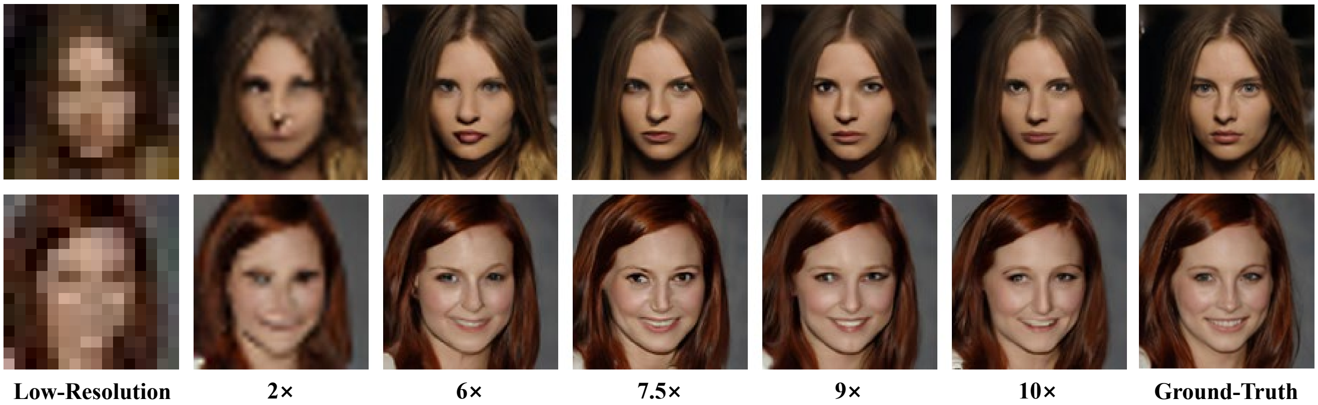


Figure 6. Visualization of continuous SR results on CelebA-HQ when training on 8× LR-HR training pairs, where the ground-truth has a resolution of 128×128. We specially select three arbitrary magnifications within the training range (1, 8] and another two out of the range (*i.e.*, 9× and 10×).

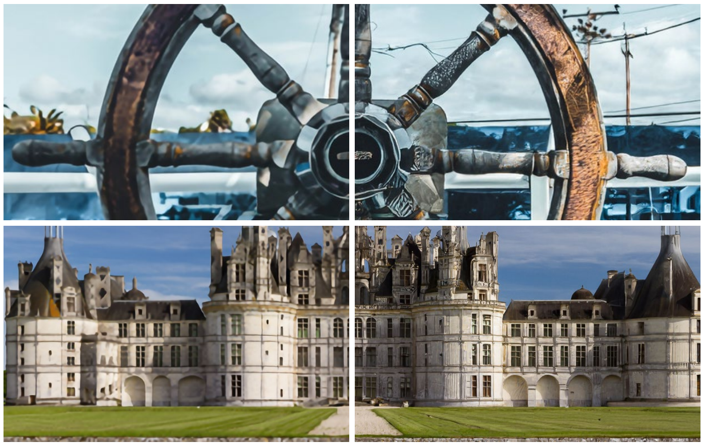


Figure 7. Two pairs of visual results by the regression-based method LIIF [6] (left part of each pair) and our IDM (right part of each pair) for 4× general scene SR.

clear, high-resolution outputs. However, its simple pixel interpolation leads to an obvious loss of realistic textures. In Table 4, the results of the GAN-based models (ESRGAN [45] and RankSRGAN [49]), flow-based models (SRFlow [24] and HCFlow [22]), and mixed generative models (HCFlow++ [22] and LAR-SR [10]) are from [10]. Table 4 demonstrates that our IDM outperforms other generative methods with a significant improvement (0.50dB on PSNR and 0.03 on SSIM). Even with less training data (800 images in ours vs. 2800 images in others), IDM still outperforms prior arts on both metrics.

4.5. Comparison of Continuous SR

Quantitative Results. Table 5 shows the quantitative comparisons on CelebA-HQ dataset with LIIF and SR3. LIIF and IDM are trained within the magnification range (1, 8] and tested on in-distribution and out-of-distribution scales, respectively. For in-distribution scales, although LIIF reports higher PSNR, our IDM exhibits much better performance in terms of LPIPS, demonstrating that the generated images of IDM are much more consistent with human perception. For out-of-distribution scales, IDM outperforms other methods in terms of both PSNR and LPIPS despite the variation of scales.

Visualization. To demonstrate the continuous SR achieved by IDM, we visualize some results with arbitrary testing magnifications when training on 8× face SR in Fig. 1 and Fig. 6. Fig. 1 shows that the regression-based models can achieve resolution-continuous results via implicit neural representation, but they suffer from the typical over-smoothing issue (second row). The generative model (SR3) performs well on the 8× magnification, consistent with that in training, but it encounters extreme distortions once the magnification changes (third row). In contrast, as shown in the fourth row of Fig. 1, IDM successfully synthesizes realistic results with the continuous resolution. In Fig. 6, even if the magnification is out of the training range (1, 8], *i.e.*, 9× and 10×, IDM still demonstrates outstanding effectiveness in representing continuous SR images.

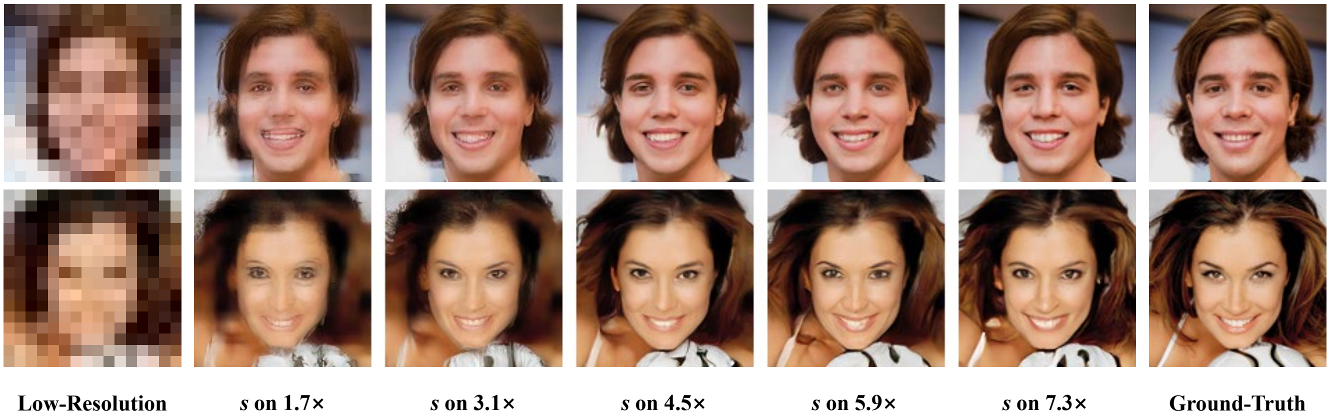
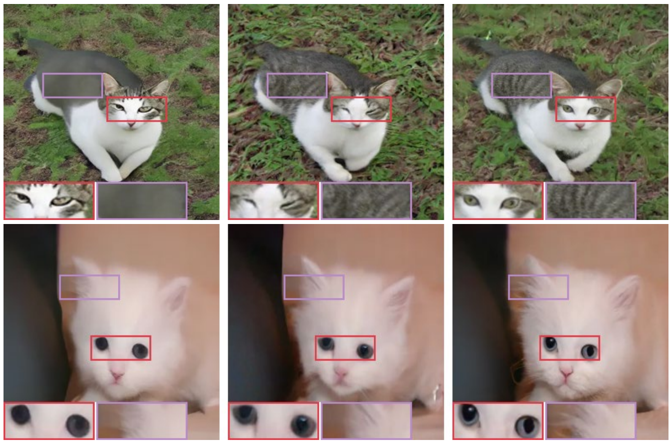


Figure 8. Visualization with different values of the scaling factor s when training on 8× face SR, where the ground-truth has a resolution of 128×128, and s takes the values of other magnifications.



(a) Concatenating **(b) Concatenating w/ encoder** **(c) LR conditioning network**

Figure 9. Effect of the LR conditioning network. (a) Conditioning our model via the concatenating operation in SR3 [35]. (b) Conditioning our model by adding an encoder [20]. (c) Our LR conditioning network.

4.6. Ablation Studies

Importance of the Scaling Factor. To demonstrate the significance of the scaling factor s, we provide qualitative visual results with different values of s when training on 8× face SR in Fig. 8. Specifically, we assign s with the value from other specific magnifications. For example, the third column in Fig. 8 is obtained using s on 3.1× face SR. Evidently, for 8× face SR, using the scaling factor assigned by smaller magnification leads to blurred textures. As the corresponding magnification increases, IDM synthesizes more fine details. It illustrates that the scaling factor is inclined to allocate more weights to generated features on large-magnification SR. Overall, the scaling factor effectively dynamically adjusts the proportion between LR condition and generated details.

Effect of the LR Conditioning Network. We conduct qualitative experiments on Cats with 16× SR to validate the effect of our LR conditioning network. Specifically, we construct two comparison models by replacing the scale-adaptive conditioning network in IDM with two types of conditioning mechanisms that concatenate (1) the upsampled LR image or (2) the LR features encoded by the EDSR encoder with the ground-truth, and feed them to the denoising model, where (1) is adopted by SR3 [35]. As shown in Fig. 9(a), directly using the upsampled LR image as the condition often leads to blurred textures. While introducing an encoder to extract features in advance can slightly alleviate this issue (Fig. 9(b)), it still performs poorly in generating high-fidelity details, such as eyes and hair. In contrast, the proposed scale-adaptive conditioning network develops a parallel architecture providing multi-resolution LR features for the denoising model, to enrich the texture information. Fig. 9(c) shows the superior performance over the others.

5. Conclusion

This paper presents an Implicit Diffusion Model (IDM) for achieving high-fidelity image super-resolution with continuous resolution. Specifically, we introduce the implicit image function in the decoding part of the diffusion denoising model. This practical end-to-end framework adopts an iterative process of diffusion denoising and implicit neural representation. We further design a scale-adaptive conditioning mechanism, which takes a low-resolution image as a condition to adjust the proportion between LR information and generated details dynamically. Extensive experiments illustrate that our IDM exhibits state-of-the-art performance.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 5, 6

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[3] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *ECCV*, 2020. 2

[4] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2, 3, 4, 5, 6

[5] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *ICLR*, 2021. 3

[6] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7

[7] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 6

[8] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv:1410.8516*, 2014. 1

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[10] Baisong Guo, Xiaoyun Zhang, Haoning Wu, Yu Wang, Ya Zhang, and Yan-Feng Wang. Lar-sr: A local autoregressive model for image super-resolution. In *CVPR*, 2022. 2, 6, 7

[11] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *CVPR*, 2022. 4

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 3, 4, 5

[13] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 2, 3, 5

[14] 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. 3

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017. 1, 5, 6, 7

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3, 5

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. 1, 3

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 3

[19] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *CVPR*, 2022. 3

[20] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022. 4, 8

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR*, 2021. 1, 3

[22] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *ICCV*, 2021. 3, 6, 7

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1, 3, 4, 6

[24] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, 2020. 1, 3, 6, 7

[25] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 4

[26] 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. 1, 2, 6

[27] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 2021. 2

[29] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 2

[30] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv:1609.03499*, 2016. 1, 3

[31] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *ICML*, 2019. 2

[32] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *ICML*, 2021. 3

[33] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 2, 5

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5

[35] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 1, 2, 4, 5, 6, 7, 8

[36] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *CVPR*, 2019. 2

[37] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 2

[38] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1

[40] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 2

[41] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 6

[42] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020. 1, 3

[43] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 1, 3

[44] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 2, 4

[45] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 2, 3, 6, 7

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[47] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015. 5, 6

[48] 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. 3, 6

[49] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, 2019. 6, 7

[50] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, 2008. 5

[51] Yutian Zhang, Xiaohua Li, and Jiliu Zhou. Sftgan: a generative adversarial network for pan-sharpening equipped with spatial feature transform layers. *JARS*, 2019. 3