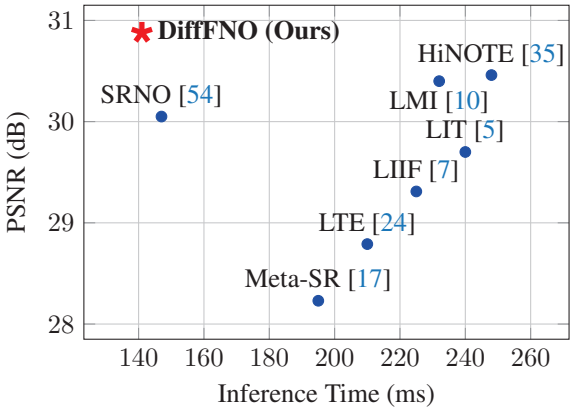
DiffFNO: Diffusion Fourier Neural Operator

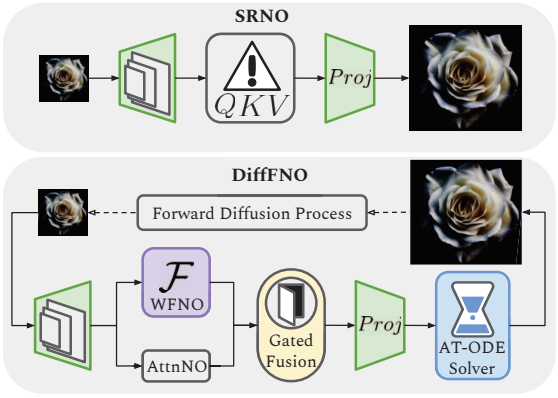
Xiaoyi Liu[1,*] Hao Tang[2,†]

[1]Washington University in St. Louis [2]Peking University

jasonl@wustl.edu haotang@pku.edu.cn



(a) **PSNR and inference time for ×4 super-resolution**



(b) **DiffFNO vs SRNO [54]**

Figure 1. (a) All models use the EDSR-baseline [31] encoder, except HiNOTE [35] which has its own. (b) Compared to SRNO [54], DiffFNO is strengthened by the fusion of spectral and spatial features and efficient refinement by a diffusion process.

Abstract

*We introduce DiffFNO, a novel diffusion framework for arbitrary-scale super-resolution strengthened by a Weighted Fourier Neural Operator (WFNO). Mode Rebalancing in WFNO effectively captures critical frequency components, significantly improving the reconstruction of high-frequency image details that are crucial for super-resolution tasks. Gated Fusion Mechanism (GFM) adaptively complements WFNO's spectral features with spatial features from an Attention-based Neural Operator (AttnNO). This enhances the network's capability to capture both global structures and local details. Adaptive Time-Step (ATS) ODE solver, a deterministic sampling strategy, accelerates inference without sacrificing output quality by dynamically adjusting integration step sizes ATS. Extensive experiments demonstrate that DiffFNO achieves state-of-the-art (SOTA) results, outperforming existing methods across various scaling factors by a margin of **2–4 dB in PSNR**, including those beyond the training distribution. It also achieves this at lower inference time (Fig. 1 (a)). Our approach sets a new standard in super-resolution, delivering both superior accuracy and computational efficiency.*

1. Introduction

Image super-resolution (SR) reconstructs high-resolution (HR) images from low-resolution (LR) inputs, recovering lost fine details to enhance visual quality. SR is crucial for applications like medical imaging [12], satellite imagery [23, 52], and video games [38]. However, multiple HR images can correspond to the same LR input due to information loss during downsampling. This ambiguity requires algorithms capable of inferring plausible and perceptually accurate high-frequency content from limited data.

Deep learning, particularly Convolutional Neural Networks (CNNs) [58], has significantly advanced SR. Dong et al. introduced SRCNN [9], demonstrating the effectiveness of end-to-end learning for SR. Subsequent models achieved remarkable performance using deeper architectures and at-

†Corresponding author.

*Work Done during the visit at Peking University.

tention mechanisms [6, 30, 31, 34, 57].

Diffusion models have emerged as powerful generative frameworks modeling complex data distributions via iterative denoising processes [11, 14, 45]. Their ability to generate high-fidelity images is well-suited for inferring missing fine details. In SR, diffusion models progressively refine an LR image by modeling the conditional distribution of HR images given the LR input [15, 25, 41, 51]. This iterative process reconstructs intricate textures and high-frequency components, producing realistic outputs.

However, diffusion models are computationally intensive due to the iterative reverse diffusion process [46]. To address this, recent research explores efficient sampling strategies to accelerate reverse diffusion. One approach is approximating the diffusion process through deterministic Ordinary Differential Equation (ODE), which can be solved in fewer steps [33]. This accelerates inference and provides consistent, reproducible results.

Arbitrary-scale SR models [7, 17, 24], which can upsample images at user-defined scales beyond those seen in training, have gained attention in recent years. Methods involving attention mechanisms [5] and representing images as continuous functions [10] have been explored. Operator-learning methods such as Super-Resolution Neural Operators (SRNO) [54] and HiNOTE [35] have further advanced this field. However, the inherent differences between physics simulations and real-world images introduce challenges from computational demands to the difficulty in restoring high-frequency details.

To address these limitations, our contributions are:
(1) We propose Weighted Fourier Neural Operator (WFNO) strengthened by iterative refinement from a diffusion framework for high-frequency reconstruction, detailed in Fig. 2. Through Mode Rebalancing (MR), WFNO learns to emphasize the most critical frequency components. This greatly enhances high-frequency image detail reconstruction, overcoming the limitations of standard FNOs and MLPs, which underrepresent such details due to mode truncation and spectral bias, respectively. (2) We develop Gated Fusion Mechanism (GFM) to dynamically adjust the influence of Fourier space features from WFNO and complementary spatial domain features from an Attention-based Neural Operator (AttnNO). AttnNO is lightweight, sharing an encoder with and running in parallel to WFNO. (3) Additionally, we present Adaptive Time-Step (ATS) ODE solver, which flexibly adjusts integration step sizes based on data characteristics by assessing the complexity of image regions, thereby reducing computational overhead without compromising quality. (4) DiffFNO achieves state-of-the-art results on multiple SR benchmarks, outperforming existing methods by **2–4 dB in PSNR** in reconstruction quality. It also offers competitive inference time as Fig. 1 (a) shows. DiffFNO remains robust across various upscaling factors—even those unseen during training.

2. Related Work

Neural Operators and Fourier Methods. *Neural Operators (NO)* [22] have emerged as a powerful framework for learning mappings between infinite-dimensional function spaces, providing resolution-invariant models that generalize across different input resolutions. Unlike traditional neural networks that map finite-dimensional vectors to other vectors, neural operators learn mappings from functions to functions [27], making them well-suited for tasks involving continuous data or data at varying resolutions.

Multi-Layer Perceptrons (MLPs) often exhibit a spectral bias, favoring low-frequency functions [39]. This limits their ability to capture fine textures and sharp edges. To overcome these limitations, techniques like positional encodings and Fourier feature mappings capture high-frequency details by embedding input coordinates into a higher-dimensional sinusoidal space, allowing the network to represent complex patterns [44, 47].

Fourier Neural Operator (FNO) [26] is a variant of NO that uses spectral convolution to efficiently capture global data patterns, modeling long-range dependencies with lower computational complexity than traditional CNNs. In physics and climate settings [19, 29, 55], FNOs can handle arbitrary input resolutions without retraining. Although successful, FNOs may still lose high-frequency information due to mode truncation (discarding higher-frequency Fourier modes). This loss impairs tasks such as SR that rely on detailed reconstruction [13, 48, 49].

The Mode Rebalancing mechanism in WFNO overcomes these limitations. Instead of being truncated, all Fourier modes are preserved, with additional learnable weights to modulate their impact on reconstruction. Fine-grained feature representation is further enhanced by AttnNO, which captures local details by processing data directly in the spatial domain

Diffusion-Based SR and Efficient Sampling. Diffusion models have gained prominence as powerful generative models capable of producing high-quality images through iterative denoising techniques [14, 42]. In the context of SR, diffusion models have been employed to model the conditional distribution of HR images given LR inputs, achieving higher resolutions after progressive enhancement [40, 41].

Despite their effectiveness, diffusion models are computationally intensive due to the large number of time steps required in the reverse diffusion process. Such computational demands pose significant challenges for practical applications [20], especially in real-time or resource-constrained settings. Current solutions include: (i) *Deterministic Sampling via ODE Solvers*: By reformulating the stochastic reverse diffusion as a deterministic ODE, advanced ODE solvers can be employed to reduce the number of sam-

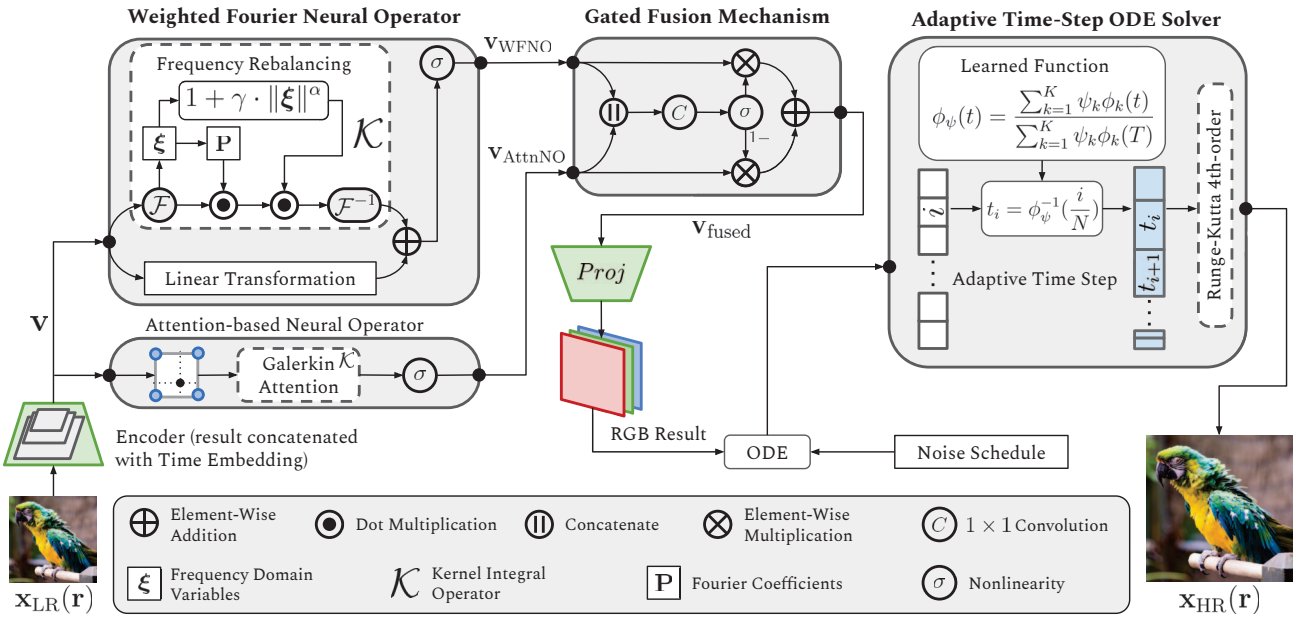


Figure 2. The proposed Diffusion Fourier Neural Opeartor (DiffFNO) architecture for arbitrary-scale super-resolution begins by lifting a low-resolution input image $\mathbf{x}_{\text{LR}}(\mathbf{r})$ into a feature space using a convolutional encoder. Features extracted by the Weighted Fourier Neural Operator (WFNO) and an Attention-based Neural Operator (AttnNO) are combined using a Gated Fusion Mechanism (GFM). The fused features are then projected into RGB space, where Adaptive Time-Step (ATS) ODE solver efficiently completes the reverse diffusion process with both accuracy and speed. This pipeline generates $\mathbf{x}_{\text{HR}}(\mathbf{r})$, a high-resolution version of the input image.

pling steps [21]. Methods like Denoising Diffusion Implicit Models (DDIM) [45] and DPM-Solver [2, 33] have demonstrated the ability to generate high-quality images with significantly fewer steps. (ii) *Operator Learning for Fast Sampling*: Neural operators accelerates sampling by learning the solution operator of the reverse diffusion process [8, 59]. (iii) *Progressive Distillation*: Training a distilled model to approximate the behavior of the full diffusion model allows faster sampling with fewer steps [37, 43]. Although effective, this method may require extensive retraining and potentially compromise image quality for increased speed.

Applying these acceleration methods to diffusion-based SR enables faster inference while maintaining high image quality. With efficient sampling methods, diffusion models become more practical for SR tasks, balancing performance and computational efficiency [32].

DiffFNO adopts (i) for its simplicity and the robustness of established numerical methods. We also strength it with the ATS strategy, which adjusts integration step sizes adaptively to balance speed and quality.

3. The Proposed DiffFNO

3.1. Network Architecture and Novel Components

An overview of the proposed DiffFNO is shown in Fig. 1 (b). It has three parts: (i) A CNN encoder extracts features from LR images. Unlike the simple linear transformations in standard FNO setups for physics simulations, our encoder is tailored for SR, extracting complex patterns and textures needed for high-quality reconstructions. We use the EDSR-baseline [31] and RDN models [58] in our experiments. (ii) WFNO and GFM: WFNO captures both global and local details alongside the AttnNO. GFM combines these into a unified HR feature map, which is then projected into RGB. (iii) ATS ODE solver accelerates inference speed by taking fewer, larger, and dynamically adjusted steps toward the reconstructed HR image. Fig. 2 illustrates these components in detail.

The network minimizes the difference between the predicted image and the true image. The loss function is:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_0}\left[\|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right], \quad (1)$$

where $\mathbf{x}_t$ (i.e. $\mathbf{x}_{\text{LR}}$) is obtained by adding noise to $\mathbf{x}_0$ (i.e. $\mathbf{x}_{\text{HR}}$). $s_\theta(\mathbf{x}_t, t)$ is the neural network approximating the true score function. $\nabla \mathbf{x}_t \log p_t(\mathbf{x}_t|\mathbf{x}_0)$ is the true score function.

3.2. Weighted Fourier Neural Operator

The Fourier Neural Operator (FNO) [26] is an efficient NO variant designed to learn mappings between function spaces. It operates directly on inputs of arbitrary resolutions, performing upscaling by mapping low-resolution inputs to high-resolution outputs. It first transforms the input data into the frequency domain, applies the learned filters, and then transforms the data back to the spatial domain.

Spectral convolution and mode truncation greatly enhance computational efficiency.

Let $\mathbf{x}_{\text{LR}}(\mathbf{r})$ denote the low-resolution input function (e.g., an image), where $\mathbf{r} \in \mathbb{R}^2$ represents spatial coordinates. The goal is to learn an operator $\mathcal{G}$ such that:

$$\mathbf{x}_{\text{HR}}(\mathbf{r}) = \mathcal{G}[\mathbf{x}_{\text{LR}}(\mathbf{r})], \tag{2}$$

where $\mathbf{x}_{\text{HR}}(\mathbf{r})$ is the output function (e.g., the super-resolved image). The FNO models $\mathcal{G}$ by stacking:

$$\mathbf{v}_{l+1}(\mathbf{r}) = \sigma\left(\mathcal{W}_l \mathbf{v}_l(\mathbf{r}) + \mathcal{K}_l \mathbf{v}_l(\mathbf{r})\right), \tag{3}$$

where $\mathbf{v}_l(\mathbf{r})$ is the feature representation at layer l evaluated at spatial location $\mathbf{r}$, and $\mathbf{v}_{l+1}(\mathbf{r})$ is its updated representation in the following layer; σ is a nonlinear activation function; $\mathcal{W}_l$ is a linear transformation. $\mathcal{K}_l$, the integral operator at layer l, transforms the features into the Fourier domain. Fourier modes are then truncated for computational efficiency. Global convolution is performed with a pointwise multiplication between the transformed features and the learned Fourier coefficients.

$$\mathcal{K}_l \mathbf{v}_l(\mathbf{r}) = \mathcal{F}^{-1}\left(\mathbf{P}_l(\boldsymbol{\xi}) \cdot \mathcal{F}[\mathbf{v}_l](\boldsymbol{\xi})\right), \tag{4}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, respectively. $\boldsymbol{\xi}$ is the frequency domain variables. $\mathcal{F}[\mathbf{v}_l](\boldsymbol{\xi})$ is the Fourier transform of $\mathbf{v}_l$, evaluated at frequency $\boldsymbol{\xi}$. $\mathbf{P}_l(\boldsymbol{\xi})$ is a complex-valued tensor of learnable parameters representing the Fourier domain filters.

However, mode truncation underrepresents high-frequency components that are critical to SR of real-world images. To address this limitation, we introduce Mode Rebalancing. A learned weighting function $\mathbf{w}_l(\boldsymbol{\xi})$ is applied to the Fourier modes to amplify or attenuate specific frequency components. It is defined at layer l as:

$$\mathbf{w}_l(\boldsymbol{\xi}) = 1 + \gamma_l \cdot \|\boldsymbol{\xi}\|^{\alpha}, \tag{5}$$

where γ_l is a learnable scalar parameter at layer l that controls the strength of the weighting; α is a hyperparameter (0.7 in our experiments or optionally a learnable parameter) that determines how the weight scales with the frequency magnitude $\|\boldsymbol{\xi}\|$. $\mathbf{w}(\boldsymbol{\xi})$ assigns higher weights to higher frequencies when $\alpha > 0$, thus emphasizing high-frequency components. This yields an updated $\mathcal{K}_l$:

$$\mathcal{K}_l \mathbf{v}_l(\mathbf{r}) = \mathcal{F}^{-1}\left(\mathbf{w}_l(\boldsymbol{\xi}) \cdot \mathbf{P}_l(\boldsymbol{\xi}) \cdot \mathcal{F}[\mathbf{v}_l](\boldsymbol{\xi})\right). \tag{6}$$

3.3. Gated Fusion Mechanism

While WFNO excels at capturing global dependencies through spectral convolutions, it may not fully exploit local interactions critical for detailed image reconstruction. We incorporate AttnNO to complement WFNO by capturing local dependencies. Working in tandem, they learn mappings from the low-resolution input function to the high-resolution output function. Gated Fusion Mechanism optimally combines the complementary features from both operators, adaptively balancing the contributions of each to a fused feature map, which is then fed to a projection layer.

Efficient implementation of the kernel integral using the Galerkin-type attention mechanism [4] has been explored in the neural operator applied to SR tasks [35, 54]. Our AttnNO is composed of bicubic interpolation, Galerkin attention, and nonlinearity, sharing an encoder with WFNO. AttnNO models local interactions in the spatial domain, focusing on the most relevant spatial regions during the convolution process. Given the complementary role of AttnNO to WFNO, we simplify its structure to improve runtime.

While previous works have applied gating mechanisms in different contexts, our approach differs significantly. Zheng et al. [60] use gating within recurrent CRF networks primarily for semantic segmentation, controlling the information flow for boundary refinement rather than fusing feature maps with distinct representations. Hu et al. [16] introduced channel-wise gating in Squeeze-and-Excitation (SE) blocks, focusing on adaptively recalibrating feature channels within a single network stream. In contrast, our Gated Fusion Mechanism applies spatial gating to integrate global dependencies captured by WFNO and local information from AttnNO. This mechanism adaptively combines both operators' feature maps, enhancing high-resolution image reconstruction by balancing global and local contributions at each spatial location.

Let $\mathbf{v}_{\text{WFNO}} \in \mathbb{R}^{B\times H\times W\times C}$ and $\mathbf{v}_{\text{AttnNO}} \in \mathbb{R}^{B\times H\times W\times C}$ denote the feature maps obtained from WFNO and AttnNO, respectively, where B is the batch size, H and W are the height and width of the feature maps, and C is the number of channels. We first concatenate the feature maps along the channel dimension and pass them through a convolutional layer followed by a sigmoid activation to produce a gating map $\mathbf{G} \in \mathbb{R}^{B\times H\times W\times 1}$:

$$\mathbf{G} = \sigma\left(\text{Conv}_{1\times 1}\left([\mathbf{v}_{\text{WFNO}}, \mathbf{v}_{\text{AttnNO}}]\right)\right), \tag{7}$$

where $[\cdot,\cdot]$ denotes concatenation along the channel dimension; $\text{Conv}_{1\times 1}$ is a 1×1 convolutional layer that reduces the concatenated features to a single-channel gating map; $\sigma(\cdot)$ is the sigmoid activation function applied element-wise.

The fused feature map $\mathbf{v}_{\text{fused}} \in \mathbb{R}^{B\times H\times W\times C}$ is the element-wise weighted sum of the two feature maps:

$$\mathbf{v}_{\text{fused}} = \mathbf{G} \odot \mathbf{v}_{\text{WFNO}} + (1 - \mathbf{G}) \odot \mathbf{v}_{\text{AttnNO}}, \tag{8}$$

where $\odot$ denotes element-wise multiplication, and subtraction is performed element-wise. The gating map $\mathbf{G}$ is broadcast across the channel dimension to match the dimensions of the feature maps.

Gated Fusion Mechanism brings two advantages compared to a naive concatenation strategy: (i) Captures complementary Information: WFNO models global dependencies through spectral convolutions, effectively modeling long-range interactions and overall structure. In contrast, AttnNO excels at capturing local dependencies and fine-grained details via attention mechanisms. (ii) Balances contributions dynamically: Gated Fusion Mechanism elicits the importance of each feature map at each spatial location, dynamically balancing global and local information.

3.4. Forward Diffusion Process and Noise Schedule

Motivation. NOs are well-suited for SR tasks due to their inherent resolution invariance and their ability to model global dependencies efficiently. Diffusion models can iteratively refine a low-resolution image to a high-resolution one, capturing the complex conditional distribution of high-resolution images given low-resolution inputs.

DiffFNO leverages the strengths of both frameworks. WFNO is a powerful mechanism for handling arbitrary resolutions and capturing high-frequency details, while the diffusion process iteratively improves reconstruction output.

In our framework, the forward diffusion process models the degradation of HR images to LR images, which in our case is primarily due to downscaling. To incorporate this degradation into the diffusion model framework, we define a forward process that simulates the downscaling effect over continuous time $t \in [0, T]$. At, the image $\mathbf{x}_T$ closely resemble the observed LR image $\mathbf{x}_{\text{LR}}$ after significant degradation. We adopt a modified variance-preserving (VP) stochastic differential equation (SDE):

$$d\mathbf{x}_t = -\frac{1}{2}\beta(t)\left(\mathbf{x}_t - \mathbf{D}\mathbf{x}_t\right)dt + \sqrt{\beta(t)}d\mathbf{w}, \qquad (9)$$

where $\beta(t)$ is the noise schedule; $\mathbf{D}$ is the downsampling operator that reduces the resolution of the image; $d\mathbf{w}$ is the standard Wiener process.

In this formulation, the term $\mathbf{x} - \mathbf{D}\mathbf{x}$ quantifies the high-frequency details lost during downscaling. The drift term $-\frac{1}{2}\beta(t)\left(\mathbf{x} - \mathbf{D}\mathbf{x}\right)dt$ models the gradual removal of these details, while the diffusion term $\sqrt{\beta(t)}d\mathbf{w}$ adds Gaussian noise to simulate further degradation.

Noise Schedule $\beta(t)$. We define the noise schedule $\beta(t)$ as a simple and effective linear function increasing over the time interval $[0, T]$:

$$\beta(t) = \beta_{\text{min}} + (\beta_{\text{max}} - \beta_{\text{min}}) \cdot \frac{t}{T}, \qquad (10)$$

where β_{min}=0.1 and β_{max}=20. This linear schedule ensures a gradual increase in the degradation strength from minimal degradation at t=0 to maximum degradation at t=T.

Relation to Image Degradation. At each time t, the image $\mathbf{x}_t$ progressively loses high-frequency details due to the drift toward $\mathbf{D}\mathbf{x}_t$, the downscaled version of the image. The added Gaussian noise further simulates the information loss inherent in downscaling. At $t = T$, the image $\mathbf{x}_T$ approximates the observed low-resolution image $\mathbf{x}_{\text{LR}}$. The reverse diffusion process then aims to recover the high-resolution image $\mathbf{x}_0$ (i.e.$\mathbf{x}_{\text{HR}}$) from $\mathbf{x}_T$ by reversing the degradation.

Choice of $\beta(t)$. The linear noise schedule is chosen for its simplicity and effectiveness. It provides a straightforward way to control the rate of degradation over time. Parameters β_{min} and β_{max} are selected to balance the trade-off between sufficient degradation (to simulate downscaling) and numerical stability of the diffusion process.

Downsampling Operator $\mathbf{D}$. The operator $\mathbf{D}$ is defined to reduce the spatial dimensions of the image by the desired scaling factor. We use bicubic downsampling.

3.5. Adaptive Time-Step

The standard reverse diffusion process is stochastic and requires a large number of sampling steps, making it computationally expensive. To accelerate inference, we reformulate the reverse diffusion as a deterministic Ordinary Differential Equation (ODE), allowing us to use advanced ODE solvers for faster sampling. The ODE solver integrates the reverse diffusion process, and its output is the super-resolved image. The reverse diffusion process can be described by a Stochastic Differential Equation (SDE) [46]:

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] dt + g(t) d\bar{\mathbf{w}} \qquad (11)$$

where $\mathbf{x}$ is the data; t is the time variable; $f(\mathbf{x}, t)$ and $g(t)$ are drift and diffusion coefficients; $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the score function; $\bar{\mathbf{w}}$ is the reverse-time Wiener process. By removing the stochastic term, we obtain the probability flow ODE, which deterministically transports the data from the noise distribution to the data distribution.

Our ATS ODE solver comprises three key components:

1. Adaptive Time Step Selection Using a Learned Function. Optimizing the allocation of time steps based on data characteristics has been explored in previous works [28, 53]. We discretize the time interval $[0, T]$ into N non-uniform time steps $\{t_i\}_{i=0}^{N}$, where $t_0 = 0$ and $t_N = T$. We introduce a learned function $\phi_\psi(t)$ as a weighted sum of polynomial basis functions, where the weight is parameterized by a set of learnable coefficients $\psi = \{\psi_1, \psi_2, \ldots, \psi_K\}$, which adaptively determines the distribution of time steps based on the data characteristics.

Parameterization of $\phi_\psi(t)$. We define $\phi_\psi(t)$ as a normalized weighted sum of K predefined monotonically increasing basis functions $\{\phi_k(t)\}_{k=1}^{K}$:

$$\phi_\psi(t) = \frac{\sum_{k=1}^{K} \psi_k \phi_k(t)}{\sum_{k=1}^{K} \psi_k \phi_k(T)}, \quad \psi_k = \exp(\omega_k), \qquad (12)$$

where each basis function $\phi_k(t) = t^k$ for $k = 1, 2, \ldots, K$ is polynomial. ω_k are unconstrained learnable parameters that ensure $\psi_k \geq 0$ through the exponential mapping. We set $K = 3$ to balance model flexibility with computational efficiency. This setup allows $\phi_\psi(t)$ to capture nonlinear timestep distributions without excessive complexity.

Selection of Time Steps. Using the learned function $\phi_\psi(t)$, we map uniformly spaced normalized values $s_i = \frac{i}{N}$ to non-uniform time steps t_i:

$$t_i = \phi_\psi^{-1}(s_i) = \phi_\psi^{-1}\left(\frac{i}{N}\right), \quad i = 0, 1, \ldots, N \tag{13}$$

Since $\phi_\psi(t)$ is monotonically increasing, its inverse function $\phi_\psi^{-1}(s)$ exists and can be efficiently computed.

2. Neural Operator Score Network. The score function, representing the gradient of the log probability density $\log p_t(\mathbf{x})$, is approximated using a neural network $s_\theta(\mathbf{x}, t)$ parameterized by θ:

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \approx s_\theta(\mathbf{x}, t). \tag{14}$$

In our architecture, s_θ consists of: (i) An encoder that extracts features from $\mathbf{x}_{\text{LR}}$); (ii) WFNO for capturing global dependencies and high-frequency details; (iii) AttnNO for modeling local dependencies and fine-grained structures; (iv) Gated Fusion Mechanism to dynamically combine features; (v) Time embedding $\mathbf{e}(t)$ incorporating the time variable t into our neural network $s_\theta(\mathbf{x}, t)$ using sinusoidal positional embeddings [14, 50], concatenating it and encoded features along the channel dimension.

3. Efficient Solver. We solve the reverse-time stochastic differential equation (SDE) of the diffusion process, transformed into an ODE:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}), \tag{15}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the image estimate at time t in the reverse diffusion process; $f(\mathbf{x}, t)$ and $g(t)$ are coefficients derived from the forward diffusion process.

For the Variance Preserving (VP) SDE commonly used in diffusion models, the coefficients are defined as:

$$f(\mathbf{x}, t) = -\frac{1}{2}\beta(t)\mathbf{x}, \quad g(t) = \sqrt{\beta(t)}, \tag{16}$$

where $\beta(t)$ is a predefined noise schedule specific to the diffusion process and is consistent with our DiffFNO. By substituting the score function approximation from Eq. (14), we define the approximate drift function:

$$f_\theta(\mathbf{x}, t) = f(\mathbf{x}, t) - \frac{1}{2} g(t)^2 s_\theta(\mathbf{x}, t). \tag{17}$$

The adaptive time steps $\{t_i\}_{i=0}^N$ discretize the ODE. We apply the Runge-Kutta 4th-order (RK4) method, as it balances computational cost and accuracy, requiring fewer steps than lower-order methods while retaining precision.

The benefits of ATS are threefold: (i) Deterministic Sampling: It consistently produces the same results for identical inputs, improving reproducibility. (ii) Reduced Computation: Fewer sampling steps significantly decrease inference time. (iii) High-Quality Reconstruction: It maintains high-quality reconstruction by efficiently allocating computational resources.

4. Experiments

Datasets and Evaluation Metrics. We use the DIV2K [1] dataset for training. For evaluation, we use the DIV2K validation set and four standard datasets: Set5 [3], Set14 [56], BSD100 [36], and Urban100 [18].

We evaluate our model on upscaling factors of $\times 2$, $\times 3$, $\times 4$, $\times 6$, $\times 8$, and $\times 12$. Notably, scales $\times 6$, $\times 8$, and $\times 12$ are outside the training distribution, as the training scales are uniformly sampled from $\times 1$ to $\times 4$. This setup assesses our model's ability to generalize to arbitrary scales. We use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) as our evaluation metrics.

Quantitative Results. Building on the quantitative gains of our model, we also present qualitative results to illustrate the visual improvements achieved. We compare our proposed DiffFNO model with several SOTA arbitrary-scale SR methods, including Meta-SR [17], LIIF [7], LTE [24], SRNO [54], LIT [5], LMI [10], and HiNOTE [35]. All models are trained on the DIV2K dataset with identical settings to ensure a fair comparison in Tables 1 and 2.

Among the compared methods, Meta-SR performs adequately at lower scales but struggles at higher scaling factors due to its generalized approach that lacks specialized mechanisms for fine detail capture. LIIF and LTE improve upon Meta-SR by using local implicit functions and frequency-based estimations, respectively, which enhance high-frequency texture representation. However, they still face limitations in capturing non-periodic textures and high-frequency details, resulting in blurred textures at larger scales. LIT and LMI further advance performance by integrating attention mechanisms and MLP-mixer architectures, effectively preserving high-frequency textures and handling diverse scales, but they may not generalize well across datasets with varying distributions. SRNO and HiNOTE employ neural operator frameworks with attention mechanisms and frequency-aware loss priors to better capture global spatial properties and enhance high-frequency detail reconstruction. We observed mixed results between SRNO and HiNOTE: at certain scaling factors, one outperforms the other, indicating their varying strengths at different resolutions. Overall, their neural operator foundation improves the handling of arbitrary scaling but may increase computational demands.

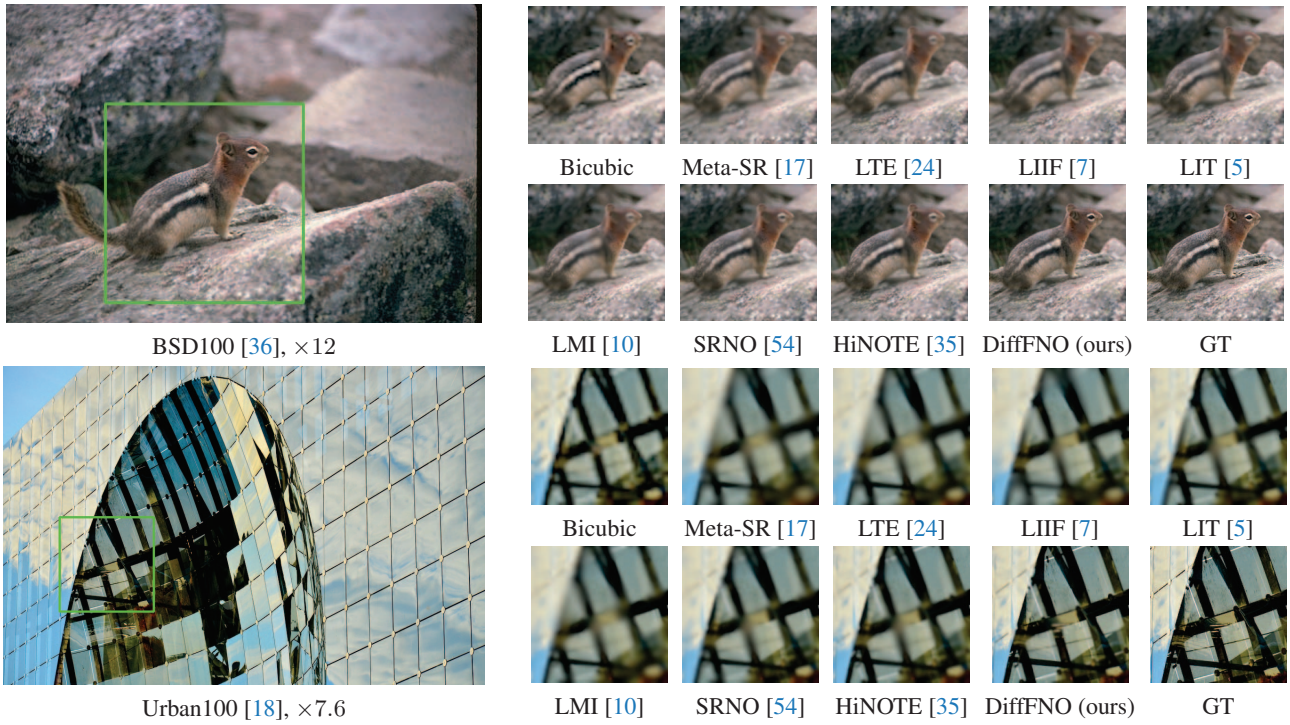


Figure 3. Qualitative comparison on integer and continuous super-resolution scales. The models use RDN [58] as their encoder (except HiNOTE [35], has its own). In the HR image, the cropped patch is outlined in green.

Model	×2		×3		×4		×6		×8		×12	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
EDSR-MetaSR [17]	33.32	0.913	30.10	0.800	28.23	0.830	26.10	0.792	24.77	0.742	23.95	0.720
EDSR-LTE [24]	33.83	0.921	30.50	0.880	28.79	0.852	26.55	0.800	25.05	0.760	24.20	0.736
EDSR-LIIF [7]	34.36	0.925	30.94	0.885	29.31	0.855	27.02	0.814	25.44	0.771	24.32	0.743
EDSR-LIT [5]	34.81	0.928	31.39	0.890	29.70	0.860	27.44	0.815	25.78	0.775	24.69	0.745
EDSR-LMI [10]	35.40	0.930	31.88	0.895	30.40	0.865	27.95	0.820	26.16	0.780	25.56	0.750
EDSR-SRNO [54]	34.85	0.928	31.45	0.890	30.05	0.863	27.36	0.810	26.00	0.772	25.91	0.760
EDSR-DiffFNO (Ours)	**35.72**	**0.932**	**32.50**	**0.905**	**30.88**	**0.870**	**28.29**	**0.830**	**26.78**	**0.790**	**26.48**	**0.775**
HiNOTE† [35]	35.29	0.931	31.90	0.895	30.46	0.842	27.83	0.799	26.41	0.772	26.23	0.732
RDN-MetaSR [17]	33.50	0.920	30.32	0.893	28.41	0.861	26.29	0.810	24.90	0.780	24.01	0.790
RDN-LTE [24]	33.98	0.922	30.65	0.882	28.94	0.852	26.70	0.802	25.20	0.762	24.35	0.732
RDN-LIIF [7]	34.51	0.927	31.09	0.887	29.46	0.857	27.17	0.812	25.59	0.772	24.47	0.742
RDN-LIT [5]	34.96	0.930	31.54	0.892	29.85	0.862	27.59	0.817	25.93	0.777	24.84	0.747
RDN-LMI [10]	35.55	0.932	32.03	0.897	30.55	0.867	28.10	0.822	26.31	0.782	25.71	0.752
RDN-SRNO [54]	35.00	0.930	31.60	0.892	30.20	0.862	27.51	0.812	26.15	0.772	26.06	0.762
RDN-DiffFNO (Ours)	**35.87**	**0.934**	**32.65**	**0.902**	**31.03**	**0.872**	**28.44**	**0.832**	**26.93**	**0.792**	**26.63**	**0.777**

Table 1. PSNR/SSIM comparison on the DIV2K [1] validation set using EDSR [31] and RDN [58] encoders. HiNOTE [35] uses its own.

Our DiffFNO model consistently achieves the highest PSNR and SSIM scores across all scaling factors and datasets. The performance gap widens at larger scaling factors (×8 and ×12), demonstrating a superior generalization to the out-of-distribution scales. The improvements are more pronounced on complex datasets like Urban100, which contain intricate textures and structures. By combining WFNO and AttnNO features through the Gated Fusion Mechanism, which adaptively balances global and local features, DiffFNO synthesizes global and local dependencies. The ATS ODE solver efficiently refines high-resolution images, further enhancing quality. This combination addresses the limitations of prior models, such as spectral bias and insufficient high-frequency detail capture.

Qualitative Results. Fig. 3 compares arbitrary-scale SR methods on a BSD100 image (scaling factor of ×12) with

Model	Set5					Set14					BSD100					Urban100				
	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8
MetaSR [17]	37.50	34.05	31.52	28.23	26.02	33.51	30.03	28.02	25.53	24.02	31.02	28.05	26.52	24.82	23.52	32.02	28.03	25.82	23.52	22.03
LIIF [7]	38.02	34.42	32.04	28.57	26.25	34.03	30.43	28.43	25.84	24.33	31.52	28.55	27.03	25.03	23.83	32.52	28.53	26.03	23.83	22.33
LTE [24]	38.21	34.63	32.25	28.76	26.44	34.22	30.65	28.64	26.05	24.52	31.71	28.73	27.23	25.23	24.03	32.72	28.75	26.23	24.03	22.53
SRNO [54]	38.32	34.84	32.69	29.38	27.28	34.27	30.71	28.97	26.76	25.26	32.43	29.37	27.83	26.04	24.99	33.33	29.14	26.98	24.43	23.02
LIT [5]	38.53	35.02	32.82	29.51	27.42	34.44	30.83	29.03	26.82	25.33	32.52	29.51	27.92	26.12	25.01	33.42	29.22	27.02	24.52	23.12
LMI [10]	38.72	35.14	32.95	29.63	27.55	34.63	31.02	29.24	27.05	25.55	32.72	29.74	28.04	26.25	25.14	33.62	29.44	27.24	24.63	23.23
HiNOTE [35]	39.01	35.22	33.08	29.85	27.74	35.02	31.25	29.55	27.35	25.85	33.02	30.05	28.15	26.35	25.25	34.03	29.83	27.55	24.73	23.34
DiffFNO (Ours)	**39.72**	**35.30**	**33.16**	**30.23**	**27.93**	**36.01**	**31.54**	**30.22**	**27.58**	**26.02**	**33.56**	**30.24**	**28.21**	**26.45**	**25.30**	**34.19**	**29.99**	**27.74**	**24.80**	**23.35**

Table 2. PSNR comparison on four benchmark datasets: Set5 [3], Set14 [56], BSD100 [36], and Urban100 [18]. All models use RDN [58] as their encoder, besides HiNOTE [35] which has its own.

Model	×2		×3		×4		×6		×8		×12		Inference	Steps
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
SRNO [54]	33.81	0.920	30.53	0.880	28.74	0.850	26.59	0.800	25.10	0.760	24.18	0.730	147	-
FNO [26]	34.36	0.925	30.94	0.885	29.31	0.855	27.02	0.810	25.44	0.770	24.32	0.740	**85**	-
WFNO	34.81	0.928	31.39	0.888	29.70	0.858	27.44	0.815	25.78	0.775	24.69	0.745	97	-
WFNO-AttnNO	35.40	0.930	31.88	0.892	30.40	0.862	27.95	0.820	26.16	0.780	25.56	0.750	139	1000
DiffFNO(-w, -a, -s)	34.85	0.928	31.45	0.890	30.05	0.860	27.36	0.815	26.00	0.775	25.91	0.760	204	1000
DiffFNO(-a, -s)	35.29	0.930	31.90	0.893	30.46	0.863	27.83	0.820	26.41	0.780	26.23	0.765	231	1000
DiffFNO(-s)	35.70	**0.932**	32.48	0.896	30.85	0.866	28.26	0.825	26.75	0.785	26.45	0.770	266	1000
DiffFNO	**35.72**	**0.932**	**32.50**	**0.900**	**30.88**	**0.870**	**28.29**	**0.830**	**26.78**	**0.790**	**26.48**	**0.775**	141	**30**

Table 3. Ablation study of variants of DiffFNO on the DIV2K [1] validation set. All use EDSR-baseline [31] backbone as their encoder. Inference times are measured in milliseconds (ms). WFNO-AttnNO has Gated Fusion Mechanism.

fine-grained details like animal fur and rock textures and an Urban100 (continuous scaling factor of $\times 7.6$) image featuring large structures and fine local details such as reflections on the grass. SRNO and HiNOTE capture multiscale details effectively, from the animal's body to tiny gaps between glass panels. However, DiffFNO reconstructs crisper edges with fewer artifacts, enhancing texture in animal fur pattern and reflections. WFNO captures large-scale patterns, while AttnNO and Gated Fusion Mechanism preserve intricate textures. This multiscale approach followed by a diffusion process enhanced by the ATS 's ODE solver further reduces visual artifacts.

Ablation Studies. Extensive ablation studies validate the effectiveness and complementary nature of new components in DiffFNO. Table 3 reports the PSNR results on the DIV2K validation set for different model variants with scaling factors from $\times 2$ to $\times 12$. **-w** denotes leaving out the Mode Rebalancing (yielding the default FNO [26]). **-a** denotes omitting AttnNO. **-s** denotes the removal of ATS ODE solver. We also measure inference time by averaging over 100 runs, and report inference steps. We establish a baseline with SRNO, whose architecture is the most similar to our DiffFNO aamong the methods covered in our study. Overall, we observe notable improvements with the addition of model components. The complete DiffFNO achieves the highest PSNR and SSIM values across all upscaling factors.

Effect of Mode Rebalancing. Incorporating Mode Rebalancing into WFNO boosted performance compared to the default FNO [26], at the cost of a slightly increased number of parameters and inference time.

Effect of Gated Fusion Mechanism and AttnNO. The Gated Fusion Mechanism introduces minimal computational overhead. In addition, extra computational cost incurred by Attention-based Neural Operator is effectively mitigated by running it in parallel with WFNO while employing a shared encoder.

Effect of ATS ODE Solver: ATS dramatically reduces the number of inference steps from $1,000$ to just 30, which substantially improves the inference time while delivering competitive performance (Tab. 3). DiffFNO outperforms the SOTA in both PSNR and inference time (Fig. 1 (a)).

5. Conclusion

We propose Diffusion Fourier Neural Opeartor (DiffFNO) for arbitrary-scale image super-resolution. DiffFNO is made of Weighted Fourier Neural Operator with a Mode Rebalancing mechanism to emphasize high-frequency details. It is complemented by a Attention-based Neural Operator through a Gated Fusion Mechanism that effectively adjusts the influence of global and local features. Image reconstruction is further refined by a diffusion process augmented with an Adaptive Time-Step ODE solver that dynamically allocates time steps, drastically cutting down inference time without compromising output quality. Experiments demonstrate DiffFNO's competitiveness in both reconstruction quality and inference time across various benchmarks, establishing a new state-of-the-art.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 6, 7, 8

[2] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022. 3

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 135.1–135.10, 2012. 6, 8

[4] Shuhao Cao. Choose a transformer: Fourier or galerkin. *Advances in neural information processing systems*, 34:24924–24940, 2021. 4

[5] Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18257–18267, 2023. 1, 2, 6, 7, 8

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023. 2

[7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1, 2, 6, 7, 8

[8] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations*, 2022. 3

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2014. 1

[10] Huiyuan Fu, Fei Peng, Xianwei Li, Yejun Li, Xin Wang, and Huadong Ma. Continuous optical zooming: A benchmark for arbitrary-scale image super-resolution in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3035–3044, 2024. 1, 2, 6, 7, 8

[11] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023. 2

[12] Hayit Greenspan. Super-Resolution in Medical Imaging. *The Computer Journal*, 52(1):43–63, 2008. 1

[13] Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In *Advances in Neural Information Processing Systems*, 2021. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 6

[15] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[17] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1575–1584, 2019. 1, 2, 6, 7, 8

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 6, 7, 8

[19] Peishi Jiang, Zhao Yang, Jiali Wang, Chenfu Huang, Pengfei Xue, T. C. Chakraborty, Xingyuan Chen, and Yun Qian. Efficient super-resolution of near-surface climate modeling using the fourier neural operator. *Journal of Advances in Modeling Earth Systems*, 15(7), 2023. 2

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 2

[21] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021. 3

[22] Nikola B Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(146):1–63, 2023. 2

[23] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. 1

[24] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1929–1938, 2022. 1, 2, 6, 7, 8

[25] 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*, 479:47–59, 2022. 2

[26] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric

partial differential equations. In *International Conference on Learning Representations*, 2021. 2, 3, 8

[27] Zongyi Li, Nikola Kovachki, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. 2

[28] Zhengyu Li, Kyungmin Kim, Jungwoo Lee, and Thomas Huang. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion sampling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[29] Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/JMS Journal of Data Science*, 1(3):1–27, 2024. 2

[30] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[31] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1, 2, 3, 7, 8

[32] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 3

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 2, 3

[34] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022. 2

[35] Xihaier Luo, Xiaoning Qian, and Byung-Jun Yoon. Hierarchical neural operator transformer with learnable frequency-aware loss prior for arbitrary-scale super-resolution. *arXiv preprint arXiv:2405.12202*, 2024. 1, 2, 4, 6, 7, 8

[36] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 416–423, 2001. 6, 7, 8

[37] Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, and Juho Lee. Early exiting for accelerated inference in diffusion models. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. 3

[38] NVIDIA Corporation. Deep learning super sampling (dlss), 2024. 1

[39] 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 *International conference on machine learning*, pages 5301–5310. PMLR, 2019. 2

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 2

[42] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Sr3: Image super-resolution via repeated refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2736–2745, 2022. 2

[43] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 3

[44] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, 2020. 2

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 5

[47] 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. *Advances in neural information processing systems*, 33:7537–7547, 2020. 2

[48] Huy Tran, Levon Nurbekyan, and Houman Owhadi. Factorized fourier neural operators. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. 2

[49] Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023. 2

[50] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 6

[51] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 2

[52] Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232: 104110, 2022. 1

[53] Rachel Watson, Anirudh Mehta, Hyojin Choi, and Ajay Singh. Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024. 5

[54] Min Wei and Xuesong Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18247–18256, 2023. 1, 2, 4, 6, 7, 8

[55] Qidong Yang, Paula Harder, Venkatesh Ramesh, Alex Hernandez-Garcia, Daniela Szwarcman, Prasanna Sattigeri, Campbell D Watson, and David Rolnick. Fourier neural operators for arbitrary resolution climate data downscaling. In *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023. 2

[56] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730, 2010. 6, 8

[57] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2

[58] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 1, 3, 7, 8

[59] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 42390–42402. PMLR, 2023. 3

[60] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhile Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015. 4