

---

# DFU: scale-robust diffusion model for zero-shot super-resolution image generation

---

**Alex Havrilla**

Department of Mathematical Sciences  
Georgia Institute of Technology  
Atlanta, GA 30332  
ahavrilla3@gatech.edu

**Kevin Rojas**

Department of Mathematical Sciences  
Georgia Institute of Technology  
Atlanta, GA 30332  
kevin.rojas@gatech.edu

**Wenjing Liao**

Department of Mathematical Sciences  
Georgia Institute of Technology  
Atlanta, GA 30332  
wliao60@gatech.edu

**Molei Tao**

Department of Mathematical Sciences  
Georgia Institute of Technology  
Atlanta, GA 30332  
mtao@gatech.edu

## Abstract

Diffusion generative models have achieved remarkable success in generating images with a fixed resolution. However, existing models have limited ability to generalize to different resolutions when training data at those resolutions are not available. Leveraging techniques from operator learning, we present a novel deep-learning architecture, Dual-FNO UNet (DFU), which approximates the score operator by combining both spatial and spectral information at multiple resolutions. Comparisons of DFU to baselines demonstrate its scalability: 1) simultaneously training on multiple resolutions improves FID over training at any single fixed resolution; 2) DFU generalizes beyond its training resolutions, allowing for coherent, high-fidelity generation at higher-resolutions with the same model, i.e. zero-shot super-resolution image-generation; 3) we propose a fine-tuning strategy to further enhance the zero-shot super-resolution image-generation capability of our model, leading to a FID of 11.3 at 1.66 times the maximum training resolution on FFHQ, which no other method can come close to achieving.

## 1 Introduction

Diffusion models have been shown to be a powerful method for image generation [1–7]. Of particular interest is the diffusion model’s ability to generate high-resolution images. One approach is to generate lower-resolution samples and then upscale using existing learning-based methods [e.g., 4, 8, 9]. Also notable is image outpainting [e.g., 10, 11], which increases the resolution of an image by extending its content outside of the original palette of the diffusion model, again in a supervised fashion. However, both approaches require access to *high-resolution training data* and do not truly generalize the learned score beyond training resolution. In contrast, we seek to learn a model which can directly **sample at high resolutions** to generate the entire image after being **trained at low resolutions**, without access to high resolution training data.

We call this problem of generating resolution higher than that of training data **zero-shot super-resolution generation**. To tackle it, we start off by first considering a standard diffusion process as a discretization of an infinite dimensional generative process[12]. Under appropriate conditions we can recover the backward process with an associated score operator [13–15]. Then we propose Dual-Fourier Neural Operator UNet (DFU) as a scale-robust deep learning architecture that

can approximate, and in fact further improve the score operator across multiple resolutions. Evaluations on complex real-world image datasets demonstrate DFU’s superior zero-shot super-resolution generation capabilities. We observe that it is able to generate samples up to twice the resolution of the training data with good *coherence* (the overall global image structure) and *fidelity* (the clarity of fine-details). It is also especially suitable for mixed-resolution training, which can significantly improve image quality at all resolutions when compared to single-resolution training. In particular, multi-resolution DFU achieves superior FID to even single resolution UNet models evaluated at their training resolution. Notably, this benefit is only enjoyed by DFU, as the same mixed-resolution training procedure seems to slightly harm UNet image quality.

**Related Work** Operator learning via deep neural networks has exploded in popularity in recent years [16–19]. This work uses Fourier Neural Operator (FNO) [20], which learns solution operators to PDEs by composing blocks which simultaneously learn linear transformations in the spatial and Fourier domains. This architecture has led to many follow-up works [21–23]. Notably, recent independent work by Rahman et al. [24] extends the FNO architecture to UNet for solving PDEs.

Diffusion process in infinite dim. and its time reversal have been extensively studied [12, 13, 25]. Recent works have also adopted this perspective with the purpose of learning diffusion generative models in infinite dimensions. However these papers focus more on constructing a rigorous formulation of such processes with empirical experiments learning Gaussian process distributions [26], well-behaved PDEs [15], or simple image datasets such as MNIST [27, 28]. In contrast, this paper focuses on the design and training of neural networks to approximate the score operator for significantly more complex distributions with less regularity.

## 2 Designing and training Dual-FNO UNet

**Learning the score operator** An infinite dimensional denoising diffusion process can be constructed if one has the corresponding score operator (see A for details or [14, 15]). We wish to design an architecture that can approximate the score operator across *multiple resolutions*. It was believed UNet [29], a popular architecture used in the majority of diffusion models [2, 3, 7], is multiscale, and in fact, UNet does have shown its ability to learn high-frequency components of images via spatial convolutions at multiple scales. In addition, the field of operator learning offers explicit designs of neural networks capable of learning mappings between function spaces, such as for solving PDEs [17, 20, 21, 24].

However, as will be demonstrated in our experiments, both of these choices come with drawbacks for our purpose of zero-shot super-resolution generation. The fact that UNet relies on spatial convolutions makes it difficult to maintain global coherence (i.e. lower frequency information) when sampling beyond training resolution (see e.g., Fig. 1). On the other hand, FNO was designed to learn mappings between regular functions, e.g., those commonly appearing as strong solutions to PDEs. Meanwhile images are less smooth due to sharp edges and maintaining only lower-frequency information results in blurred images at higher resolutions.

**Designing Dual-FNO UNet** These observations lead us to introduce the Dual-Convolution (*DC*) of a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  as:

$$DC(f) = K_{\text{spatial}} * f + \mathcal{F}^{-1}(K_{\text{spectral}} \odot \mathcal{F}(f)) \in \mathcal{H} \quad (1)$$

We call the first term the *spatial convolution* and the second term the *spectral convolution*. We use this operation as the base for DFU by introducing it into the Unet architecture. (For precise definitions refer to B). We implement DFU with  $L = 4$  “UNet layers” and 4 Dual-FNO blocks per layer. Each Dual-FNO block consists of 2 Dual-Convolutions with an affinely transformed time-embedding added in between. A spatial kernel size of  $k = 3$  is used at all levels. The  $M = 16$  lowest Fourier modes are used in the top level spectral convolution. The number of modes used is then halved at each subsequent level. A full diagram of the architecture is attached in the appendix 4.

Note we tweaked the spatial convolution part of the implementation: the size of the discrete kernel remains fixed even if resolution increases, and thus it is not directly a discretization but a rescaled version. This implies that our architecture is not an operator. Note this is by design, as it allows the learning of fine scale details at lower resolutions, which then can be transferred to higher scale and become finer scales. For instance, when trained on  $96 \times 96$  data, the spectral convolution would



Figure 1: Visual comparison of DFU to various baselines at zero-shot super-resolution image generation. **Panel 1:** Single-resolution UNet trained on data with resolution up to 96x96, sampled at 128x128. **Panel 2:** Multi-resolution UNet sampled at 128x128. **Panel 3:** Multi-resolution DFU sampled at 128x128. Note both UNets struggle to generalize global coherency past the max training resolution  $r = 96$ .

enlarge a  $2 \times 2$  patch of the image to  $4 \times 4$  when sampling a  $192 \times 192$  image. However, the spatial convolution would remain  $2 \times 2$  ensuring the details are refined at the smallest scale possible. This contributes to DFU’s ability of zero-shot generation of higher resolution images with high fidelity, setting it apart from pre-existing architectures or techniques such as FNO [20], UNO [24], or bilinear interpolation.

### 3 Experiments

Code: <https://github.com/Dahoas/edm>

**Setup** We evaluate on Flicker Faces HQ (FFHQ) [30] and Lsun-Church [31]. For each dataset, we start with a ground truth resolution  $r = 256$ . To construct our mixed-resolution training data, we downsample to resolutions  $r \in \{32, 48, 64, 80, 96\}$ . Training is run for 400k steps with a batch size of 256. Sampling is done in 18 steps using the EDM scheme [5]. Note, while the majority of discussion here analyzes FFHQ, a comparison of results for LSUN can be found in Appendix E.

	32x32	64x64	96x96	128x128	160x160
Single-res UNet + Bilinear	2.9	5.3	6.4	14.5	31.9
Multi-res UNet	2.5	4.3	7.5	25.2	45.5
FNO UNet	30.5	40.1	31.4	53.5	64.1
DFU	<b>1.47</b>	<b>2.92</b>	<b>4.96</b>	<b>7.85</b>	<b>14.8</b>

Table 1: FIDs on **FFHQ** across resolutions. Note all models are trained on a maximum resolution of  $r = 96$ . Evaluation done on  $r = 128, 160$  is **zero-shot super-resolution** image generation.

**Baselines** We compare DFU against several baselines. *Single-res UNet + Bilinear* is the standard *single-resolution UNet* trained on a fixed resolution  $r = 96$ . Bilinear interpolation is then used to resize an image to the target resolution. The *Multi-resolution UNet* architecture adapts single-resolution UNet to training on multiple resolutions. We achieve this by removing the resolution dependent operations from a vanilla UNet. *FNO UNet* replaces the spatial convolutions in DFU with kernels of size 1, mimicking the FNO architecture per block. This architecture is similar to the UNO architecture in [24]. Notice that by comparing against these baselines we demonstrate the need for both spatial and spectral convolutions. Table 1 reports a subset of the results for FFHQ. Other than for single-resolution Unet, which is trained on a fix resolution, training is run on the same mixture of resolutions as used to train DFU. Figure 2 plots  $\text{FID}_r$  on both datasets for DFU and the best performing baseline. Full tables for all FIDs from both datasets are provided in the appendix.

**DFU generalizes to higher resolutions** DFU achieves a lower FID score than all baselines across all training resolutions. Furthermore, DFU significantly outperforms all baselines in the zero-shot super-resolution image-generation setting. The closest contender is the single-res UNet + bilinear interpolation baseline which can maintain coherent facial structures at higher resolutions via but

quickly loses fidelity. For a direct comparison between the two approaches see Figure 5. In contrast both multi-res UNet and FNO UNet struggle, with multi-res UNet unable to learn structural coherency and FNO UNet unable to learn high-frequency details. For a direct comparison between super-resolved images from the multi-res UNet and our model, see Figure 1. Further we remark DFU is able to generalize, not just the low-frequency structure of images to higher resolutions, but also higher-frequency details than presented in the training data. See for example Figure 3 which displays the generation ability of our model across multiple resolutions. These observations suggest spatial convolution learn detail while low-frequency spectral convolutions learn structure.

**Resolution training mixture impacts zero-shot super-resolution image-generation** By default we train all models on a uniform mixture of training resolutions. We also experiment with adjusting this mixture to place higher probability on higher resolution samples. However, we found concentrating too much mass on high resolutions leads to overfitting. This damages zero-shot super-resolution performance. Our best performing mixture sets resolution weight  $w_{96} = 0.4$  and  $w_{80} = 0.3$  with the rest of the weight decaying exponentially. The resulting model achieves slightly better FID on higher training resolutions and worse FID on lower resolutions than uniform mixing. However, we see a significant improvement in zero-shot super-resolution image-generation FID, e.g., improving FID<sub>160</sub> on FFHQ to 12.1 from the 14.8 of our uniformly mixed model.

Model	32x32	64x64	96x96	128x128	160x160	196x196
Dual FNO Unet	<b>1.31</b>	2.40	3.43	<b>4.86</b>	<b>8.78</b>	<b>20.00</b>
Multi-res Unet	1.52	<b>2.29</b>	3.62	<b>6.67</b>	<b>11.95</b>	20.69
Single res Unet	127.28	27.94	<b>3.31</b>	<b>15.44</b>	<b>32.06</b>	NA
Single res Unet + Bilinear Interpolation	2.96	3.26	3.31	<b>7.15</b>	<b>15.08</b>	<b>23.71</b>

Table 2: FIDs on **LSUN Church** across resolutions for different models. The single res Unet model is trained on  $96 \times 96$  data.

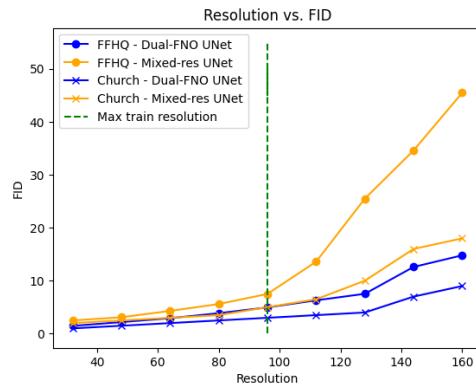


Figure 2: Sampled resolution versus FID of multi-res UNet and DFU. DFU has lower FID before losing local coherence at 2x training resolution.

**Mixed-resolution training improves single-resolution generation** Surprisingly, DFU significantly outperforms the baselines, not just in the zero-shot super-resolution generation regime, but also within training resolutions. For example, on FFHQ our model attains FID<sub>96</sub> = 4.96 in comparison to a vanilla UNet architecture with FID<sub>96</sub> = 6.4. This suggests mixed-resolution training with DFU can

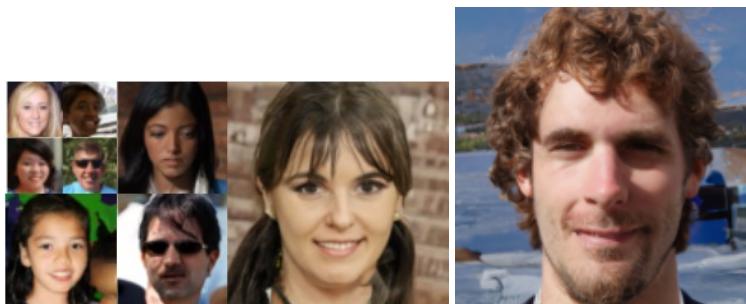


Figure 3: Samples from DFU across resolutions. **Left:** DFU sampled at  $r = 32, 64, 128$ . **Right:** DFU sampled at  $r = 160$ . DFU is trained on a mixture of resolutions from  $r = 32$  to  $96$ .

improve  $\text{FID}_r$  over training only at the single resolution  $r$ . To test this hypothesis, we train a series of single-resolution UNet and DFU models from  $r = 32$  to  $r = 96$ . We find DFU, when trained at a single resolution, slightly under-performs UNet. However, when trained at multiple resolutions, DFU outperforms single resolution UNet models. This even holds for the lowest resolutions, where UNet trained on  $r = 32$  achieves  $\text{FID}_{32} = 1.75$  whereas our  $\text{FID}_{32} = 1.47$ . Full tables can be found in the appendix.

## References

- [1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [2] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=PxTIG12RRHS>.
- [3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [4] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.
- [5] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=k7FuTOWM0c7>.
- [6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [7] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.
- [8] 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*, 2022.
- [9] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. *arXiv preprint arXiv:2303.05511*, 2023.
- [10] 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(2):295–307, 2015.
- [11] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.
- [12] Giuseppe Da Prato and Jerzy Zabczyk. Frontmatter, pages i–vi. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1992.
- [13] H. Föllmer and A. Wakolbinger. Time reversal of infinite-dimensional diffusions. *Stochastic Processes and their Applications*, 22(1):59–77, 1986. ISSN 0304-4149. doi: [https://doi.org/10.1016/0304-4149\(86\)90114-6](https://doi.org/10.1016/0304-4149(86)90114-6). URL <https://www.sciencedirect.com/science/article/pii/0304414986901146>.
- [14] Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models for function spaces. *arXiv preprint arXiv:2302.10130*, 2023.

- [15] Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. [arXiv preprint arXiv:2302.07400](#), 2023.
- [16] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218 – 229, 2019.
- [17] Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. [CoRR](#), abs/2108.08481, 2021. URL <https://arxiv.org/abs/2108.08481>.
- [18] Somdatta Goswami, Aniruddha Bora, Yue Yu, and George Em Karniadakis. Physics-informed deep neural operator networks. [ArXiv](#), abs/2207.05748, 2022.
- [19] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model Reduction And Neural Networks For Parametric PDEs. *The SMAI Journal of computational mathematics*, 7:121–157, 2021. doi: 10.5802/smai-jcm.74. URL <https://smai-jcm.centre-mersenne.org/articles/10.5802/smai-jcm.74/>.
- [20] Zongyi Li, Nikola Borislavov 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. URL <https://openreview.net/forum?id=c8P9NQVtmn0>.
- [21] Md Ashiqur Rahman, Manuel A Florez, Anima Anandkumar, Zachary E Ross, and Kamyar Azizzadenesheli. Generative adversarial neural operators. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=X1VzbBU6xZ>.
- [22] Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. U-fno—an enhanced fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022.
- [23] Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. *Advances in neural information processing systems*, 34:24048–24062, 2021.
- [24] Md Ashiqur Rahman, Zachary E. Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. [ArXiv](#), abs/2204.11127, 2022.
- [25] Annie Millet, David Nualart, and Marta Sanz. Time reversal for infinite-dimensional diffusions. *Probability theory and related fields*, 82(3):315–347, 1989.
- [26] Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. [arXiv preprint arXiv:2212.00886](#), 2022.
- [27] Paul Hagemann, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. [arXiv preprint arXiv:2303.04772](#), 2023.
- [28] Giulio Franzese, Simone Rossi, Dario Rossi, Markus Heinonen, Maurizio Filippone, and Pietro Michiardi. Continuous-time functional diffusion processes. [arXiv preprint arXiv:2303.00800](#), 2023.
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018.

- [31] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *ArXiv*, abs/1506.03365, 2015.
- [32] Werner Heisenberg. *Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik*. Springer, 1985.

## Acknowledgment

We sincerely thank Tianrong Chen, Guan-Horng Liu and Qinsheng Zhang for precious tips, such as for training existing models. AH is grateful to Stability.AI for partial support from a Graduate Research Fellowship and providing compute to train the models. KR and MT are grateful for partial support from NSF DMS-1847802, GT Cullen-Peck Scholarship, and GT-Emory Humanity.AI Award.

## A Diffusion and Diffusion Generative Modeling in Infinite Dimensions

We first review the approach in [14] for defining an infinite dim. forward SDE, its time-reversal, and the score operator: Let  $\mathcal{D}$  be a data distribution supported on a separable Hilbert space with inner product  $\langle \cdot, \cdot \rangle$ . For a fixed positive-definite, symmetric covariance operator  $C : \mathcal{H} \rightarrow \mathcal{H}$  with  $\text{trace}(C) < \infty$ , the forward SDE is:

$$dX_t = -\frac{1}{2}X_t dt + \sqrt{C}dW_t, \quad X_0 \sim \mathcal{D}$$

where  $CW_t$  is called the C-Wiener process [12]. As  $t \rightarrow +\infty$ , the distribution of  $X_t$  will converge to the stationary distribution  $\mathcal{N}(0, C)$ .

The score, generalized to infinite dim as an operator, can be defined as [14]

$$s(t, x) = \frac{1}{1 - e^{-t}} \left( x - e^{\frac{-t}{2}} \mathbb{E}[X_0 | X_t = x] \right), \quad \text{for } x \in \mathcal{H}, \quad (2)$$

and it will correct the drift in the infinite dim. backward SDE

$$dY_t = \frac{1}{2}Y_t dt + s(t, Y_t)dt + \sqrt{C}dW_t, \quad Y_0 \sim p_T$$

so that again  $Y_t$  and  $X_{T-t}$  agree in distribution.

*Remark 1.* An alternative approach to infinite dim diffusion generative model is adopted in [27], where one can define the process using a countable orthonormal basis (thanks to separable Hilbert space)  $\{e_i\}_{i=1}^\infty$  where each  $e_i$  is an eigenvector of  $C$  and their associated eigenvalues are  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq 0$ . One can then define the process of interest coordinate-wise [13] and discretize as

$$dX_t^i = f(X_t^i, t)dt + g(t)dW_t^{C_n}, \quad X_0 \sim D \quad i = 1, \dots, n \quad (3)$$

where  $W_t^C$  denotes  $\sqrt{C}W_t$  and  $C_n$  is a  $n$ -th truncated Karhunen-Loève decomposition of  $W_t^C$ , i.e:

$$W_t^{C_n} = \sum_{k=1}^n \sqrt{\lambda_k} \beta_k(t) e_k$$

for mutually independent standard Wiener processes  $\beta_k$ . These processes can then be treated as standard finite dimensional processes. However, we move away from this approach, since we would like to parametrize the score as an operator.

## B Dual FNO Precise Definitions

**Images as discretizations of functions** We model 2D images as functions on  $[0, 1]^2$ . In particular, we consider 2D image with three channels to correspond to a spatial discretization of three functions of continuous 2D space in the separable Hilbert space  $\mathcal{H} = L^2([0, 1]^2)$ . An image at resolution  $r$

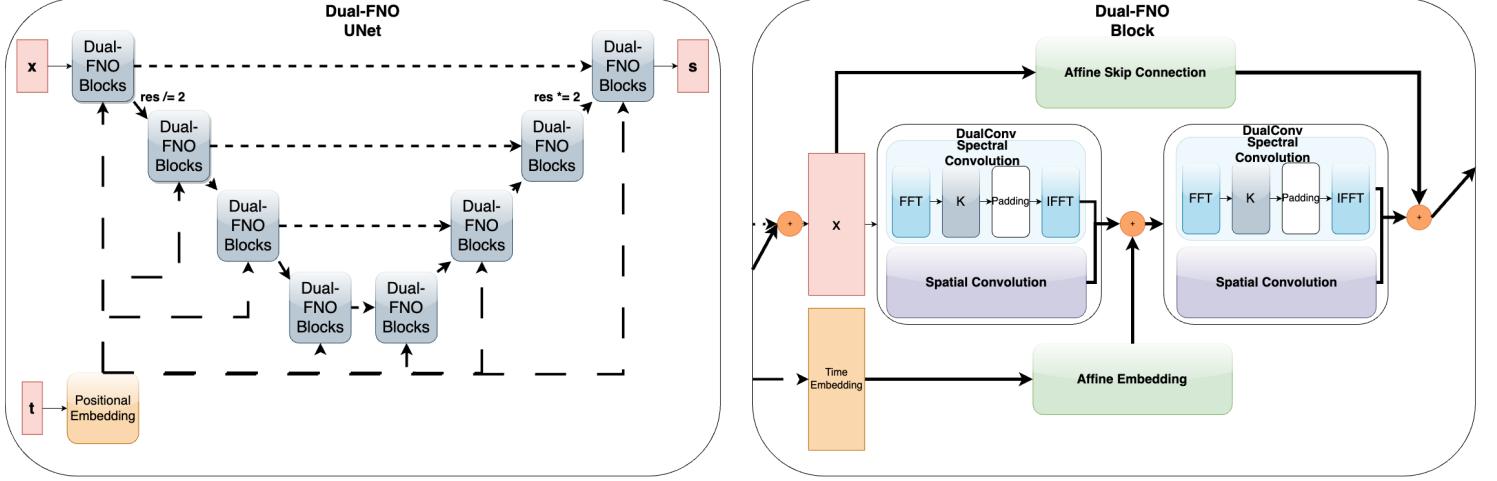


Figure 4: **Left:** DFU architecture. **Right:** Dual-FNO Block. Right is integrated in left by connecting arrows of corresponding line styles (solid for passing matrices, short dashed for skip connection, and long dashed for time embedding).

discretizes an image function  $f$  by sampling  $f$  from a uniform grid of  $r^2$  points  $G_r \subseteq [0, 1]^2$ . The key to achieving zero-shot super-resolution image generation is then to design a model which can learn features of the underlying image function distribution from a set of training image resolutions. Note for ease of presentation our further description will often be based on operations on functions from 1D space  $[0, 1]$  to  $\mathbb{R}$ . However, generalizing to 2D space (and higher-dim. if needed) and multiple channels is natural and simply done coordinate-wise.

**FNO Blocks** FNO is composed of a series of *FNO blocks* of the form

$$B_{\text{FNO}}(f) = W \odot f + \mathcal{F}^{-1}(K \odot \mathcal{F}(f))$$

where  $\mathcal{F}(f)$  is the Fourier transform of  $f$ , and  $\odot$  is Hadamard (i.e. pointwise) product.  $W, K \in \mathcal{H}$  are two learnable functions, and  $K$  induces a linear operator  $f \mapsto \mathcal{F}^{-1}(K \odot \mathcal{F}(f))$  that is a pointwise multiplication in frequency domain but a convolution in spatial domain.

In practice,  $f$  and  $K$  are both discretized. In order for the learned  $K$  to extend to different resolutions, FNO adopts the approach of fixing its bandwidth, i.e. keep the number of nonzero elements in  $K$  constant based on a fixed cutoff frequency.

**Dual Convolution** We introduce the Dual-Convolution (*DC*) of a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  as:

$$DC(f) = K_{\text{spatial}} * f + \mathcal{F}^{-1}(K_{\text{spectral}} \odot \mathcal{F}(f)) \in \mathcal{H} \quad (4)$$

where  $*$  is convolution and  $K_{\text{spatial}} : \mathbb{R}^d \rightarrow \mathbb{R}$  is a function supported on  $[-\varepsilon_s, \varepsilon_s]^d$ , corresponding to a learnable spatial kernel, and  $K_{\text{spectral}} : \mathbb{R}^d \rightarrow \mathbb{C}$  is a function supported on  $[-M, M]^d$  (i.e. with cut-off frequency  $M > 0$ ), corresponding to a learnable spectral kernel. We call the first term the *spatial convolution* and the second term the *spectral convolution*.

Note that these two components are not equivalent since  $K_{\text{spatial}}$  is localized in space and  $K_{\text{spectral}}$  is localized in frequency. Due to the Heisenberg Uncertainty Principle [32], a kernel cannot be localized in space and frequency simultaneously. Thus the spatial kernel  $K_{\text{spatial}}$  cannot be replaced by  $\mathcal{F}^{-1}(K_{\text{spectral}})$  or vice versa. The DC operator augments the analogous spectral convolution in FNO such that the global low-frequency features can be learned by the spectral convolution and the high-frequency features can be learned by the spatial convolution.

## C Fine-tuning via bilinear Interpolation

Despite improved resolution generalization compared to baselines, DFU still suffers from coherency issues when sampled at a sufficiently high resolution. However, image fidelity remains largely

unaffected. In contrast, simple zero-shot super-resolution generation techniques such as bilinear interpolation maintain image coherency at the cost of greatly reduced fidelity. For example, see the left two panels of Figure 5 and compare the images sampled via DFU and bilinear interpolation.

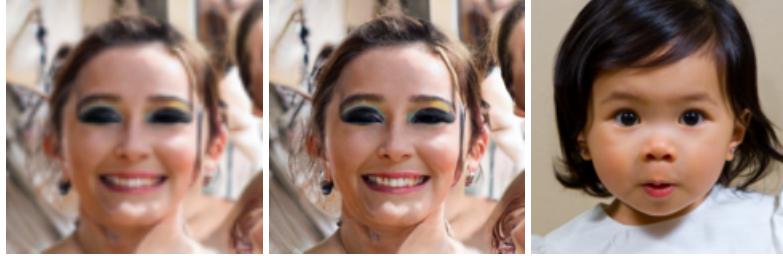


Figure 5: Comparison of pre-trained DFU to bilinear upsampling. **Left:**  $r = 96$  image bilinearly upsampled to 160x160. **Middle:** DFU sampled at 160x160. **Right:** Ground truth 160x160 image included as a reference for quality. The bilinearly upsampled image is able to maintain good coherency but lacks fidelity. In contrast DFU maintains both coherence and fidelity comparable to the ground truth.

**Setup** To improve DFU for large resolutions, we leverage bilinear interpolation’s ability to maintain coherence when upsampling. To do so, we first upsample our maximum resolution training dataset  $\mathcal{D}_{r_{\max}}$  to a target resolution  $R > r_{\max}$  producing our fine-tuning dataset  $\mathcal{D}_R$ . We then take our model pre-trained on lower resolutions and fine-tune on  $\mathcal{D}_R$  to improve image coherence at resolution  $R$ . Note this bilinear interpolation itself is a zero-shot super-resolution technique. Therefore our model is not being exposed to ground-truth higher-resolution images.

While fine-tuning, we need to do so without destroying the superior fidelity learned by DFU. To preserve fidelity, we use two independent tricks. First, we fine-tune on a mixture of resolutions instead of just on  $\mathcal{D}_R$ . We assign mix weight  $w_R$  to  $\mathcal{D}_R$  and the rest of the mix weight uniformly assigned to the pre-training resolutions as  $w_r = \frac{1-w_R}{5}$  for  $r \in \{32, 48, 64, 80, 96\}$ . This prevents our model from overfitting to the poor fidelity of  $\mathcal{D}_R$  by simultaneously re-training on high fidelity low-resolution images. We also experiment with freezing the spatial convolutions of the pre-trained model when fine-tuning. This is motivated by the use of spatial convolutions to learn high-frequency details (providing good fidelity) with the unfrozen spectral convolutions learning low-frequency structure to improve coherency.

	32x32	64x64	96x96	128x128	160x160
Pre-trained Model	1.47	2.92	4.96	7.85	14.8
Vanilla fine-tune ( $w_R = 1.0$ )	20.92	12.94	14.96	32.21	27.1
Mix w/ $w_R = 0.2$	1.41	<b>2.44</b>	4.66	7.69	12.4
Frozen spatial conv	1.46	2.85	4.76	7.62	11.9
Frozen spatial conv + $w_R = 0.2$	<b>1.36</b>	2.46	<b>4.54</b>	<b>7.06</b>	<b>11.3</b>

Table 3: FIDs on FFHQ across resolutions for different fine-tuning schemes. Our best result combines mixed resolution fine-tuning and parameter freezing.

**Results** We sweep over the mixing parameter  $w_R$ , the learning rate, and the number of fine-tuning steps while additionally freezing/un-freezing parts of the architecture. Results are recorded in table 4. Our best results are achieved with  $w_R = 0.2$ ,  $lr = 1e-4$ , while fine-tuning for an additional 10k steps with all the spatial convolutions frozen except those on the bottom layer of DFU.

We find both mixed resolution fine-tuning with a small  $w_R$  and freezing spatial convolutions to be important for preserving image fidelity at higher resolutions. The choice of  $w_R$  naturally trades-off how much fidelity is learned from lower resolutions versus how much coherence is learned from the upsampled dataset  $\mathcal{D}_R$ . Higher  $w_R$  corresponding to improved coherence. Beyond a certain threshold (e.g.,  $w_R \geq 0.5$ ) image fidelity degrades to a quality similar to  $\mathcal{D}_R$ . We see a full fine-tune

with  $w_R = 1$  degrades lower resolution performance. Additionally, fine-tuning for too many steps (more than 15k) allows the model to overfit to the poor fidelity data in  $\mathcal{D}_{256}$ . This adversely trades-off quality for coherence.

Freezing the spatial convolution blocks in combination with mixed resolution training further improves image fidelity allowing us to obtain our best zero-shot 11.3 FID on  $r = 160$ . When fine-tuning on a mixed resolution dataset with frozen spatial conv blocks we found it best to unfreeze the model on lower resolution samples, only freezing parameters on batches from  $\mathcal{D}_R$ . Fine-tuning with frozen blocks without a mixed dataset improved fidelity over the baseline vanilla fine-tune, but not as much as with mixed-resolution training. See Figure 9 for a comparison of various methods.

Fine-tuning on low-quality high resolution data also has the added benefit of further improving FID on lower resolutions. For example this improves our  $r = 32$  FID to a remarkable 1.36, despite fine-tuning on data at 5 times this resolution. This suggests learning structural details to improve coherence at higher resolutions allows our model to improve its score approximation across all resolutions. Such improvement gives more evidence that DFU well approximates the underlying score operator.



Figure 6: A visual comparison between fine-tuning methods. **Left:** 160x160 image sampled from naively-finetuned model. The image is over-smoothed. **Middle:** 160x160 image sampled from frozen + mixed resolution fine-tuned model. We have sacrificed some fidelity (but not too much) for coherency. **Right:** 160x160 image sampled from pre-trained model (i.e. DFU trained on resolution up to 96x96, without fine-tuning). This image has good fidelity but poor coherence.

## D FID across resolutions

In this section we add all the FID values for different models across resolutions for both FFHQ and LSUN-Church.

	32x32	48x48	64x64	80x80	96x96	112x112	128x128	144x144	160x160
Single-res UNet	1.75	2.65	3.92	4.85	6.43	N/A	N/A	N/A	N/A
Single-res Dual-FNO UNet	2.05	3.21	4.38	5.46	7.11	N/A	N/A	N/A	N/A
Single-res UNet (96x96) + Bi-linear	2.9	4.4	5.3	5.5	6.4	11.6	14.5	23.2	31.9
Multi-res UNet	2.5	3.6	4.3	5.8	7.5	17.9	25.2	36.4	45.5
FNO UNet	12.1	17.4	27.4	29.1	31.4	36.8	44.5	48.9	52.1
Uniformly Pre-trained Dual-FNO UNet	1.47	2.24	2.92	3.86	4.96	6.43	7.85	11.6	14.8
Weighted Pre-trained Dual-FNO UNet	2.56	3.55	3.94	3.72	4.80	6.01	7.12	10.3	12.1
Vanilla fine-tune ( $w_R = 1.0$ )	20.92	16.43	12.94	13.37	14.96	21.52	32.21	35.2	27.1
Mix w/ $w_R = 0.2$	1.41	1.94	2.44	3.54	4.66	6.31	7.69	10.2	12.4
Frozen spatial conv	1.46	2.45	2.85	3.77	4.76	5.95	7.62	9.92	11.9
Frozen spatial conv + $w_R = 0.2$	<b>1.36</b>	<b>1.89</b>	2.46	<b>3.48</b>	<b>4.54</b>	<b>5.87</b>	<b>7.06</b>	<b>9.77</b>	<b>11.3</b>

Table 4: FIDs on **FFHQ** across resolutions for all model types. Note the single-res UNet/Dual-FNO UNet rows train and evaluates models on each resolution independently. Evaluation done at more than  $96 \times 96$  is **super-resolution**.

Model	32x32	64x64	96x96	128x128	160x160	196x196
Dual FNO Unet	<b>1.31</b>	2.40	3.43	<b>4.86</b>	<b>8.78</b>	<b>20.00</b>
Multi-res Unet	1.52	<b>2.29</b>	3.62	<b>6.67</b>	<b>11.95</b>	<b>20.69</b>
Single res Unet	127.28	27.94	<b>3.31</b>	<b>15.44</b>	<b>32.06</b>	<b>NA</b>
Single res Unet + Bilinear Interpolation	2.96	3.26	3.31	<b>7.15</b>	<b>15.08</b>	<b>23.71</b>

Table 5: FIDs on **LSUN Church** across resolutions for different models. The single res Unet model is trained on  $96 \times 96$  data, the other models are trained on a uniform mixture of  $32 \times 32$ ,  $48 \times 48$ ,  $64 \times 64$  and  $96 \times 96$  data. Evaluation done at more than  $96 \times 96$  is **super-resolution**.

## E Dependence on Data Distribution

We observe better results on the LSUN-Church dataset compared to FFHQ. We attribute this difference to the relatively more rigid global structure of FFHQ. LSUN-Church contains a broader spectrum of potential global structures, as churches can assume various forms. Additionally, the level of detail surrounding the churches is more lenient, allowing for greater flexibility. Conversely, the FFHQ dataset is less flexible, as the global structure of a face must be preserved consistently, and the details must closely resemble the actual features (eyes, noses, etc.). Thus distributions with rigid global structure like FFHQ highlight the ability of DFU to well learn this structure. A full table of FIDs for both FFHQ and LSUN-Church can be found in the appendix along with samples from LSUN-Church.

## F LSUN-Church samples across resolutions

In this section we provide additional samples generated on the LSUN-Church dataset using Dual FNO Unet. We train our model using the same configuration as for the FFHQ dataset and the same multi-resolution training regime. No fine tuning was performed for this dataset.

## G FFHQ samples across resolutions

We provide additional samples generated by DFU, the MultiRes baseline, and the SingleRes baseline on FFHQ. Figure 10 provides a direct visual comparison of the methods. Figures 11, 12, and 13 contain more samples from each method respectively.



Figure 7: Samples from Dual FNO Unet at  $160 \times 160$  resolution. The model was trained in resolutions up to  $96 \times 96$



Figure 8: Samples from Dual FNO Unet at  $196 \times 196$  resolution. The model was trained in resolutions up to  $96 \times 96$



Figure 9: Samples from Dual FNO Unet at  $256 \times 256$  resolution. The model was trained in resolutions up to  $96 \times 96$



Figure 10: Samples from DFU, SingleRes + Bilinear Interpolation, and MultiRes UNet respectively up to  $160 \times 160$  resolution. The models were trained in resolutions up to  $96 \times 96$



Figure 11: **Left:** Samples from DFU over  $32 \times 32$  to  $128 \times 128$ . **Right:** DFU at  $160 \times 160$



Figure 12: **Left:** Samples from MultiRes UNet over  $32 \times 32$  to  $128 \times 128$ . **Right:** MultiRes UNet at  $160 \times 160$

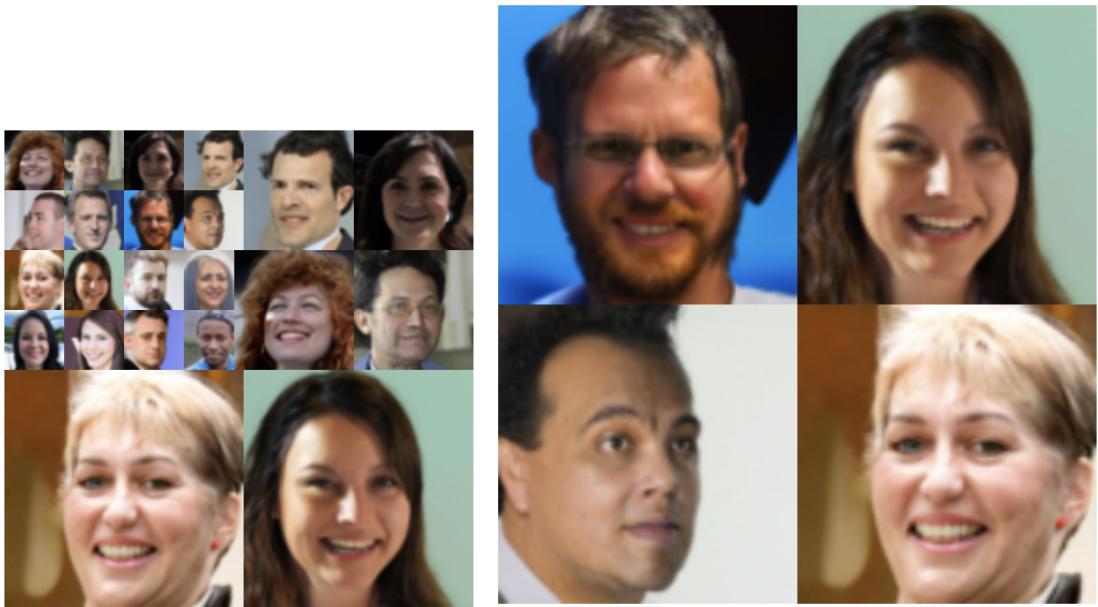


Figure 13: **Left:** Samples from SingleRes UNet + bilinear interpolation over  $32 \times 32$  to  $128 \times 128$ .  
**Right:** SingleRes UNet + bilinear interpolation at  $160 \times 160$