ONE RULE FOR EVERY SCALE: RECOVERING FINE DETAIL WITH SELF-SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks used for modeling natural signals often correctly learn coarse spatial scales but struggle to resolve fine scales, leading to blur, missing detail, or rollout drift. This attenuation of fine detail has been observed independently across many domains, from image reconstruction to neural PDE solvers. We observe that many of these natural signals exhibit approximate self-similarity, meaning that their fine scales are largely predictable from their coarse scales. In this work, we propose to embed self-similarity as an architectural inductive bias to better resolve fine scales. We formalize this inductive bias as a recurrence with a shared operator across scales and prove when this operator can be parameterized efficiently with parameter count independent of the number of fine scales predicted, allowing it to be appended to existing models as a small module. Using this result, we introduce self-similar cascade decoders, a family of such modules that admit domain-specific realizations. Our Fourier basis instantiation, called the self-similar spectral cascade (S3C), improves performance on turbulent Navier-Stokes forecasting by up to 54% across backbones, resolutions, and time horizons—while contributing 0.1% of total parameters. Our wavelet instantiation, S2WC, outperforms several strong baselines for deterministic image super-resolution.

1 INTRODUCTION

A common challenge for deep neural networks modeling natural signals (e.g., images, audio, weather) is learning and resolving fine spatial scales. This difficulty manifests as loss of detail, blurring, and rollout drift across tasks and domains. For instance, naive coordinate MLPs struggle to fit high-frequency detail (Tancik et al., 2020). Neural partial differential equation (PDE) solvers and deterministic weather models often neglect low-amplitude spatial frequencies, with errors that compound over long rollouts (Lippe et al., 2023; Lam et al., 2023). Similar behavior has been documented across a wide range of fields, including physics-informed neural networks (Wang et al., 2022), speech synthesis (Ren et al., 2022), video models (Mathieu et al., 2015), and MRI reconstructions (Muckley et al., 2021). See Figure 1b for an illustration.

These observations have been reported independently and under various names. We refer to this general phenomenon as **spectral attenuation**: the systematic decay of model output spectra compared to the target. Various domains have attributed spectral attenuation to different mechanistic causes and have developed their own remedies. For instance, Fourier features (Tancik et al., 2020) and sinusoidal activations (Sitzmann et al., 2020) alleviate spectral bias (Rahaman et al., 2019; Xu et al., 2019) for coordinate MLPs. New layers have been introduced to combat architectures biased towards coarser modes (Li et al., 2020; Liu-Schiaffini et al., 2024), and generative approaches have been proposed to prevent blur that arises in one-to-many mappings (Price et al., 2025).

However, we observe a commonality among many of these domains affected by spectral attenuation: the natural signals they model are often statistically or approximately self-similar across scales (Figure 1a). For instance, natural images empirically follow a $1/f^2$ power spectrum (Field, 1987; Ruderman & Bialek, 1993), and image patches recur across scales within a single image, sufficiently reliably to perform single-image super-resolution (Glasner et al., 2009). Turbulent fluids and atmospheric fields exhibit predictable spectral scaling laws (Obukhov, 1941; Gage, 1985). Similar laws pervade the natural sciences, such as in waves (Phillips, 1958), heartbeat periods (Kobayashi & Musha, 1982), and brain activity (Novikov et al., 1997).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

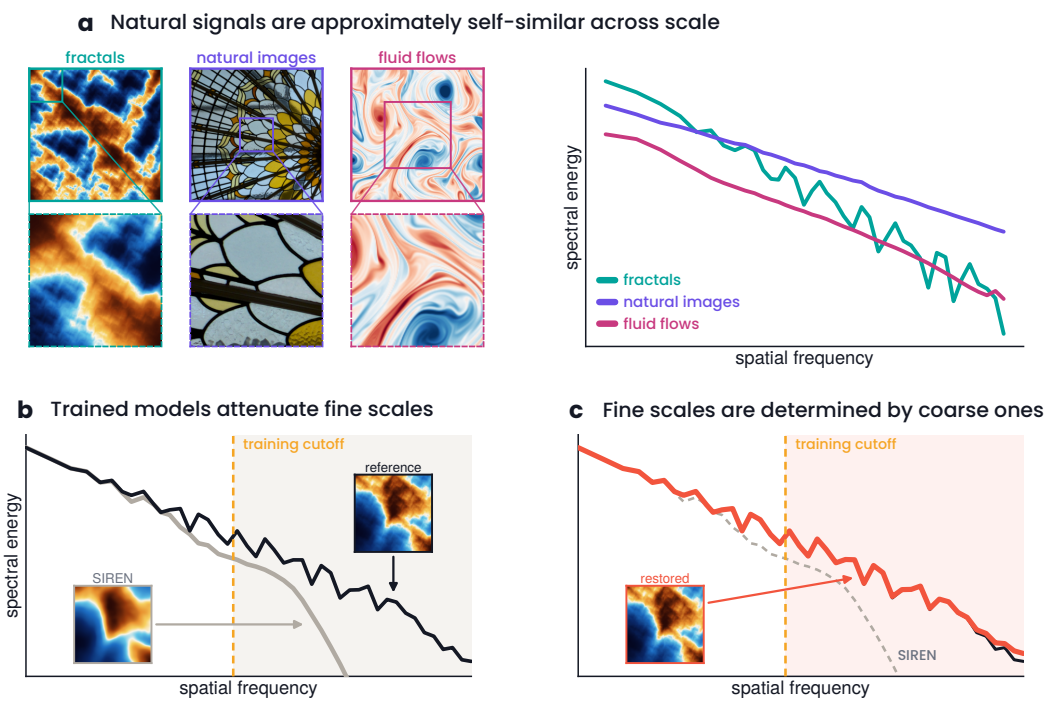


Figure 1: (a) Natural signals exhibit approximate self-similarity across scales, which can often be observed through power-law decays in the power spectrum. (b) Neural networks often attenuate the fine scales, particularly near and beyond the training cutoff, leading to blur and lack of fine detail. (c) Under self-similarity, fine scales can be approximately determined by coarse scales. We use self-similarity as a prior to recover these fine scales. Power spectra and model predictions in (b-c) are from an illustrative toy fractal problem. Details in Appendix A.2.

In this paper, we incorporate **self-similarity** as an inductive bias to mitigate spectral attenuation in any existing architectural backbone. We formalize this as a recurrence in scale, with a single shared operator mapping from coarser to finer scales. However, we prove that naive parameter sharing across scales (e.g., instantiated as a kernel integral or spatially-varying convolution) requires the parameter count to grow in order to produce progressively finer detail in the scale-recurrence. We prove that the shared operator must instead relate scales geometrically, acting on a rescaled copy of its input (e.g., via a fixed coordinate dilation). Under this condition, one shared parameter-efficient operator suffices to produce arbitrarily fine detail in the scale-recurrence. Self-similarity thus mitigates spectral attenuation with scale-independent parameter cost.

We introduce *self-similar cascade decoders* as a family of modules that correct attenuated fine scales in neural network outputs while preserving the accurate coarse scales. We show that classical transposed convolutions and fractal image coding can be recovered as special cases of this family. Instantiating the module as a scale-recurrence in Fourier space, we introduce the *self-similar spectral cascade (S3C)*, a tiny module that learns and extrapolates the power spectrum decay of neural network outputs. S3C achieves up to 54% improvement on fluid flow temporal evolution while only accounting for 0.1% of model parameters and is capable of zero-shot generalization to unseen Fourier bands, with performance improvements that compound over autoregressive rollouts. We instantiate the module in wavelet space as the *self-similar wavelet cascade (S2WC)*, an expressive attention-based module that learns to map wavelet decompositions between scales, outperforming several strong baselines on deterministic image super-resolution benchmarks.

2 Self-similarity for efficient scale recurrence

In this section, we frame self-similarity as a recurrence in scale with some shared operator S. We consider parameterizing S using a simple generalization of the convolution integral with a spatially-

varying kernel and a dilated translation, and show that the parameter-free dilation is the crucial component that enables fine detail production with a scale-independent kernel. We then derive special cases of S to recover standard transposed convolutions and fractal image coding, while we instantiate S in Fourier and wavelet space to derive the modules in Section 3.

2.1 PRELIMINARIES

Consider the function Hilbert space $\mathcal{F} = L^2(\mathcal{X})$, with domain $\mathcal{X} \subseteq \mathbb{R}^d$, as the model space for natural signals (e.g., images or solutions to PDEs). Consider the nested sequence of closed subspaces $V_0 \subset V_1 \subset V_2 \subset \cdots \subset \mathcal{F}$ whose union is dense in $\mathcal{F}$ and their corresponding orthogonal projection operators $P_k : \mathcal{F} \to V_k$. We define a sequence of *detail spaces* $\{W_k\}_{k\in\mathbb{N}}$ by $W_k = (P_{k+1} - P_k)\mathcal{F}$ and define $P_{W_k} \coloneqq P_{k+1} - P_k$. For any $v_{k+1} \in V_{k+1}$, we have $v_{k+1} = v_k + w_k$ for some $v_k \in V_k$ and $w_k \in W_k$ (Mallat, 2008). Thus, any $f \in \mathcal{F}$ can be uniquely decomposed as

$$f = v_0 + \sum_{k=0}^{\infty} w_k \tag{1}$$

where $v_0 = P_0 f$ and $w_k = P_{W_k} f$. We characterize *spectral attenuation* as the empirical phenomenon where a model prediction of f is only reliable on $V_0, W_0, W_1, \ldots, W_K$. Mitigating spectral attenuation is the process of increasing the reliability window K (Fig. 1b-c). While this is generally not possible for arbitrary classes of signals, we show that self-similarity provides sufficient structure. We define *self-similarity* as an explicit recurrence over the decomposition in Eq. 1. Let $S : \mathcal{F} \to \mathcal{F}$ be a fixed operator and set $\tilde{S}_k \coloneqq P_{W_k} S$. Starting from $v_0 =: w_{-1}$, we define

$$w_k \coloneqq \tilde{S}_k w_{k-1} = \tilde{S}_k \tilde{S}_{k-1} \cdots \tilde{S}_0 v_0, \qquad f = v_0 + \sum_{k=0}^{\infty} \tilde{S}_k \tilde{S}_{k-1} \cdots \tilde{S}_0 v_0, \tag{2}$$

so S is the self-similarity operator mapping each scale to the next. By learning operator S, we can reconstruct the fine scales $w_{K+1}, w_{K+2}, \ldots$ of a self-similar signal f from $v_0, w_0, \ldots, w_K$.[1] To see this, suppose we have S and f up to scale K, that is $f^{(K)} = v_0 + \sum_{k=0}^{K} w_k$ then for any $J > K$,

$$f^{(J)} = f^{(K)} + \sum_{k=K+1}^{J} \tilde{S}_k w_{k-1}.$$

where $w_K = P_{W_k} f^{(K)}$. We propose a model class for S which admits efficient parameterization.

2.2 PARAMETERIZING THE SELF-SIMILARITY OPERATOR S

Suppose $\mathcal{X} = \mathbb{R}^d$ and $\phi \in W_{k-1}$. A common primitive for parameterizing operators in deep learning is the convolution

$$(S_{\text{conv}}\phi)(x) = \int_{\mathcal{X}} \alpha(\omega)\, \phi(x - \omega)\, d\omega, \tag{3}$$

for some spatially homogeneous kernel α. However, by the convolution theorem, it is well-known that $S_{\text{conv}}\phi$ can never have Fourier[2] support wider than that of ϕ. Thus, S_{conv} is a poor candidate to learn a mapping $W_{k-1} \to W_k$. As such, we consider two potential generalizations of S_{conv} to learn the self-similarity operator: (1) making α spatially-varying $\alpha(x, \omega)$ (a spatially-varying convolution (Liu et al., 2018)) and (2) modulating the translation $x - \omega$ by a dilation factor λ,

$$(S\phi)(x) = \int_{\mathcal{X}} \alpha(x, \omega)\, \phi(\lambda x - \omega)\, d\omega. \tag{4}$$

We now analyze when S is capable of synthesizing frequencies higher than those in its input ϕ.

Fix $\gamma > 1$ and define the sequences $\{V_k\}_{k\in\mathbb{N}}$ and $\{W_k\}_{k\in\mathbb{N}}$,

$$V_k = \left\{ f \in \mathcal{F} : \operatorname{supp} \widehat{f} \subseteq \bigcup_{l=0}^{k} B_l \right\}, \qquad W_k = \left\{ f \in \mathcal{F} : \widehat{f} = 0 \text{ a.e. outside } B_k \right\}, \tag{5}$$

[1] Appendix C.2 derives the L^2 convergence conditions for this representation.

[2] We default to Fourier W_k due to their widespread use across engineering and applied mathematics. However, our construction is general, and the choice of W_k derives architectures for particular domains (Section 3).

where $\widehat{f}$ denotes the Fourier transform of f and the sequence $\{B_k\}_{k\in\mathbb{N}}$ denotes the Fourier bands

$$B_k = \{\xi \in \mathbb{R}^d : \gamma^k \le \|\xi\| \le \gamma^{k+1}\}.$$

Further, for $0 < \varepsilon < \gamma - 1$, we define the sub-band

$$I_k^\varepsilon = \{\xi \in \mathbb{R}^d : (1+\varepsilon)\gamma^k \le \|\xi\| \le \gamma^{k+1}\} \subset B_k.$$

Theorem 1 (informal)**.** *Given operator S in Eq. 4 and $P_x(\alpha)$ the spatial frequency support of α,*

$$\operatorname{supp} \widehat{S\phi} \subseteq \lambda \operatorname{supp} \widehat{\phi} + P_x(\alpha), \tag{6}$$

i.e., the bandlimit of $S\phi$ is the sum of the geometric dilation of λ on the bandlimit of ϕ and the spatial bandlimit of α. See Appendix C and Eq. 21 for the full statement and proof.

Specifically, for $\phi_{k-1} \in W_{k-1}$, if $\lambda \le 1$, then for $S\phi_{k-1}$ to produce nonzero energy in $I_k^\varepsilon \subset B_k$, α must have x-bandlimit $R \ge \varepsilon\gamma^k$. Conversely, $\lambda > 1$ suffices to produce energy at arbitrarily high frequencies[3] with α independent of k and fixed x-bandlimit R.

Parameterizing S entails learning kernel α with a finite number of parameters. Below, we explicitly connect the theorem's statements on bandlimit R to parameter count for signals defined on the torus.

Example 1 (Learning a power law)**.** *Suppose $\mathcal{X} = \mathbb{T}$ is the torus[4] with $\gamma = 2$ in Eq. 5, and the target spectrum decays as a power law: $|\hat{f}(\xi)| \propto |\xi|^{-\beta}$. If $\alpha(x,\omega) = \chi(x)\cdot\alpha_0(\omega)$ then*

$$\widehat{S\phi}(\xi) = \sum_{\zeta\in B_{k-1}} \hat{\chi}(\xi - \lambda\zeta)\cdot\hat{\alpha}_0(\zeta)\cdot\hat{\phi}(\zeta).$$

For a power law, the gain is approximately $\hat{\alpha}_0(\zeta) = 2^{-\beta}$, a single parameter independent of ζ. Fully parameterizing χ with bandlimit R requires $2R+1$ parameters as we work in one dimension.

If $\lambda = 2$, for $\xi \in B_k$, there exists a frequency $\zeta \in B_{k-1}$ such that $|\xi - 2\zeta| \le 1$, so $R = 1$ suffices. If $\lambda \le 1$, $|\xi - \lambda\zeta| \ge 2^{k-1}$ for all $\zeta \in B_{k-1}$ and $\xi \in I_k^{1/2}$, so $R \ge 2^{k-1}$, requiring $2^k + 1$ Fourier coefficients. Thus, for $\lambda = 2$, the required number of parameters is 4 for all scales k. For $\lambda \le 1$, the minimum number of parameters grows quickly, reaching 1026 for $k = 10$.

2.3 Special cases of the self-similarity operator

We now show that S can be elegantly expressed in well-known bases. Let $\{\varphi_m^{(k-1)}\}_{m=1}^M$ and $\{\varphi_n^{(k)}\}_{n=1}^N$ be orthonormal bases for W_{k-1} and W_k, respectively. For W_{k-1}, we can express $w_{k-1} = \sum_{m=1}^M c_m^{(k-1)}\varphi_m^{(k-1)}(x)$. Applying projection operator P_{W_k}, we can express the coefficients of Sw_{k-1} in W_k space, under basis $\{\varphi_n^{(k)}\}_{n=1}^N$,

$$c_n^{(k)} = \left\langle \varphi_n^{(k)}, Sw_{k-1}\right\rangle = \sum_{m=1}^M c_m^{(k-1)} \underbrace{\left[\int_{\mathcal{X}}\int_{\mathcal{X}} \varphi_n^{(k)*}(x)\,\alpha(x,\omega)\,\varphi_m^{(k-1)}(\lambda x - \omega)\,d\omega\,dx\right]}_{A_{n,m}^{(k)}},$$

where $A^{(k)} \in \mathbb{C}^{N\times M}$ represents the structured linear transformation that maps the coefficients of W_{k-1} to those of W_k. In practice, α, and hence sequence $\{A^{(k)}\}_{k\in\mathbb{N}}$, is encoded by finite parameters, which explicitly represents the shared mapping between scales (i.e., the self-similar structure of the target). Full derivation is given in Appendices C.2 and C.3.

Fourier basis Suppose we are on the torus $\mathcal{X} = \mathbb{T}^d$ and have defined W_k as in Eq. 5. Setting $\lambda = 2$ and selecting the Fourier basis $\varphi_m^{(k)}(x) = (2\pi)^{-d/2}e^{i\xi_n x}$, matrix $A^{(k)}$ has the form

$$A_{n,m}^{(k)} = \widetilde{\alpha}(\xi_n - 2\xi_m, \xi_m), \tag{7}$$

where $\widetilde{\alpha}$ is the full Fourier transform of kernel $\alpha(x,\omega)$ from Eq. 4.

[3] See Appendix A.1 for an illustrative example.

[4] Theorem 1 can be extended to the torus using standard techniques (Stein & Weiss, 1971).

Wavelet basis In the wavelet basis with mother wavelet ψ, $\varphi_n^{(k)}(x) = 2^{k/2}\,\psi(2^k x - n)$, we can once again set $\lambda = 2$. For the classical Haar wavelet, we obtain

$$A_{n,m}^{(k)} = \frac{1}{\sqrt{2}} \underbrace{\int_{\mathcal{X}} \alpha\left(x, \frac{n-m}{2^{k-1}}\right) |\varphi_n^{(k)}(x)|^2 \, dx}_{=:\,\bar{\alpha}_n\left(\frac{n-m}{2^{k-1}}\right)} = \frac{1}{\sqrt{2}}\,\bar{\alpha}_n\left(\frac{n-m}{2^{k-1}}\right), \tag{8}$$

where the interpretation of $\bar{\alpha}_n(\omega)$ is a function that routes energy between W_{k-1} and W_k:

$$\bar{\alpha}_n(\omega) = \mathbb{E}_{x\sim|\varphi_n|^2}[\alpha(x,\omega)]. \tag{9}$$

Transpose convolutions A similar calculation can be used to show that standard stride-2 transpose convolutions can be derived under a zeroth order B-spline basis with $\lambda = 2$ (Eq. 4). The resulting discretized kernel is constructed from an underlying continuous kernel and generalizes the resolution-dependent transposed convolution to function spaces. See Appendix C.3.

Fractal image coding Replacing the simplified $\lambda x - \omega$ with a general coordinate transformation $\tau(x,\omega)$ in Eq. 4, our construction also generalizes classical fractal image encoding Jacquin (1992) and its recent neural variant Poli et al. (2022). Both partition an image into range cells, matched to a larger domain cell (or weighted combination of them, in the neural case), and they can be recovered by integrating over the set of domain cells, letting α be piecewise constant dependent on range-domain pairs, and setting τ to encode range-domain matches. See Appendix C.3 for details.

3 Self-similar cascade decoders

In Section 2.3, we showed that self-similarity in scale admits a convenient coarse-to-fine interpretation under several canonical bases, such as Fourier series, wavelets, and B-splines. We use this analysis to characterize the design space for practical deep learning architectures.

State-dependent recurrence Real-world systems rarely exhibit exact self-similarity. As such, we condition the shared operator S (Eq. 4) on the current approximation v_{k-1}:

$$(S_k\phi)(x) = \int_{\mathcal{X}} \alpha(x,\omega; v_{k-1}) \cdot \phi(\lambda x - \omega)\, d\omega \qquad c^{(k)} = A^{(k)}(v_{k-1})\, c^{(k-1)}. \tag{10}$$

The convergence of Eq. 2 follows similarly to the state-independent case, where a uniform bound $\sup_v \|A(v)\|_{\text{op}} < 1$ guarantees convergence, with a limit depending on v_0 (see Appendix C.2).

Given the goal of correcting attenuated fine scales while preserving accurate coarse scales, we introduce the *self-similar cascade decoder* $\mathcal{D}$ appended after any deep architecture $\mathcal{E}$. In the continuum case, the network acts on $f \in \mathcal{F}$ in the state-dependent analog of Eq. 2:

$$(\mathcal{D}\circ\mathcal{E})f = \mathcal{E}f + \sum_{k=0}^{\infty} \left(\tilde{S}_k \circ \cdots \circ \tilde{S}_1 \circ \mathcal{E}\right) f. \tag{11}$$

We now introduce two architectural modules that instantiate the theory developed in Section 2.3.

3.1 Self-similar spectral cascade

Self-similar spectral cascade (S3C) is the minimal instantiation of this architecture into practice, developed primarily for PDE applications. S3C intuitively learns to extrapolate the power spectrum profile from reliable low frequencies to arbitrarily high frequencies.

Concretely, suppose $\mathcal{X} = \mathbb{T}^d$, the torus. S3C is instantiated in the Fourier basis, setting $\lambda = 2$ in Eq. 4 and restricting $\alpha(x,\omega) = \alpha(\omega)$ to a spatially homogeneous kernel. The bands of analysis are thus dyadic (via $\lambda = 2$), as defined in Eq. 5. Since α is restricted in this manner, by Theorem 1, the operator S (Eq. 4) maps between exact dyadic harmonics. Note that since $\lambda > 1$, Theorem 1 enables S to expand the frequency support of the input despite α being spatially homogeneous. As such, $(S\phi)(x) = \int_{\mathcal{X}} \alpha(\omega)\,\phi(2x-\omega)\,d\omega$ in Fourier space can be expressed as

$$\widehat{S\phi}(\xi) \propto \widehat{\alpha}(\xi/2) \cdot \widehat{\phi}(\xi/2). \tag{12}$$

Without a particular directional prior in the general case, we further simplify $\widehat{\alpha}$ to be a function of the spectral radius $|\xi|$. These choices of α, λ, and basis mean that S3C can be expressed in Fourier coefficient space with a sequence of matrices $A^{(k)}$, following Eq. 7. However, in practice such an approach overrides Fourier phase information at higher frequencies, replaced by the diagonal harmonic rule imposed by the spatially-invariant $\alpha(\omega)$. We find that the more effective practical approach is to restrict S3C to Fourier amplitudes instead, retaining the model's predicted phase.

S3C is a highly parameter-efficient module appended to the end of a large backbone architecture $\mathcal{E}$ (e.g., FNO, U-Net, or Transformer), either in the latent space before channel projection or directly in the output space. Let hyperparameter ξ^* be the highest frequency for which the backbone's predictions are reliable, and define $Z = \widehat{\mathcal{E}f}$ with channel dimensionality C for input $f \in \mathcal{F}$.

1-d case For clarity, we first consider the 1-d case. S3C proceeds in a scale-recurrent manner on the Fourier amplitudes (acting independently per channel), first initializing the state as the log amplitude slope:

$$s_0 = \log|Z_{\xi^*}| - \log|Z_{\xi^*/2}| \in \mathbb{R}^C.$$

An MLP $g_\theta : \mathbb{R}^C \to \mathbb{R}^C$ shared across scales progresses s_k and predicts the log-gain of s_{k+1}:

$$s_{k+1} = -\text{softplus}\left(-(s_k + g_\theta(s_k))\right), \qquad L_k = \sum_{j=1}^{k} s_j. \tag{13}$$

For $\xi > \xi^*$ and $\xi \in B_k$ (i.e., ξ in dyadic band k spanning $(\xi_k, \xi_{k+1}]$), we define

$$\log \mathcal{A}^*_\xi = \log|Z_{\xi^*}| + L_{k-1} + s_k \frac{\log\xi - \log\xi_k}{\log\xi_{k+1} - \log\xi_k},$$

interpolating amplitudes in each band. Since α in equation 4 is x-independent, the operator maps each mode to exactly twice its wavenumber, leaving much of the bands B_k empty. We thus perform log-linear interpolation between the endpoints of the band to avoid introducing spectral artifacts. The final output preserves the original predicted phases,

$$Z_\xi = \begin{cases} Z_\xi & |\xi| \le \xi^* \\ \mathcal{A}^*_\xi \cdot \frac{Z_\xi}{|Z_\xi| + \varepsilon} & |\xi| > \xi^* \end{cases}. \tag{14}$$

N-d case For higher dimensions (particularly the 2-d instantiation presented in Section 4), we radially bin $|Z|$ into shells by $|\xi|$ and perform all computation analogously to the 1-d case, with the exception that amplitudes of individual modes $|Z_\xi|$ are replaced by shell-wise average amplitudes.

3.2 Self-similar wavelet cascade

To showcase our theory in different bases, we also present the *self-similar wavelet cascade (S2WC)* in a wavelet basis. As in S3C, wavelet cascade shares parameters between scales and can be applied in the output or lifted latent space after a larger neural backbone. Unlike S3C, the scale decomposition is explicit. Given the current approximation $v_k \in \mathbb{R}^{H\times W\times C}$, S2WC decomposes it by one level using the discrete wavelet transform (DWT) into a coarse approximation $a_k \in \mathbb{R}^{H/2\times W/2\times C}$ and three detail spaces $d_k \in \mathbb{R}^{H/2\times W/2\times 3C}$. The state-dependent kernel $\alpha(x, \omega; a_k, d_k)$ is defined to be a cross-attention operator, with corresponding matrix representation $A(a_k, d_k)$ that implements the coarse-to-fine mapping in Eq. 8. The queries are the current approximation v_k, the keys are the wavelet coefficients $[a_k, d_k]$, and the are values the current-scale details d_k. The next scale is

$$d_{k+1} = A(a_k, d_k)\, d_k \qquad\qquad v_{k+1} \propto \text{IDWT}(v_k, d_{k+1}) \in \mathbb{R}^{2H\times 2W\times C}, \tag{15}$$

up to an energy-preserving normalization constant. Since $A(a_k, d_k)$ maps between coarse and finer detail spaces, each pixel on the former grid corresponds to four pixels on the latter. The cross-attention operator thus attends a $P \times P$ patch on the coarse grid ($H/2 \times W/2$) to the values of the four new pixels in the output detail space.

We use a 2-d axial RoPE (Su et al., 2024; Fang et al., 2024) positional encoding within each patch and, in this case, do not constrain the cross-attention matrix $A(a_k, d_k)$ to enforce L^2 convergence. Since we introduce this architecture primarily for image applications, we relax the convergence condition because (1) natural images may lack monotonically-decreasing power spectra and (2) convergence in L^2 may itself not be desirable or necessary for many imaging applications.

Table 1: Turbulent Navier-Stokes autoregressive forecasting error on 64×64 data. All errors are reported $\times 10^{-2}$. Backbones have $O(10^7)$ parameters; S3C adds $O(10^4)$.

		rel. L^2		rel. H^1		rel. HF_{32}	
Backbone	Steps	Base	**+S3C**	Base	**+S3C**	Base	**+S3C**
FNO	1	10.0	**9.8**	17.3	**17.1**	58.8	**56.5**
	2	12.7	**12.5**	22.6	22.6	65.2	**63.6**
	4	16.3	**16.1**	29.8	**29.7**	71.2	**70.3**
	6	19.5	**19.3**	35.9	**35.6**	76.2	**75.9**
	8	23.0	**22.7**	42.2	**41.7**	**81.9**	82.0
	10	26.7	**26.5**	48.7	**48.2**	**88.0**	88.5
U-Net	1	8.5	**6.9**	14.4	**11.4**	46.5	**40.0**
	2	12.8	**9.9**	22.8	**17.2**	56.3	**49.2**
	4	19.5	**14.6**	35.3	**26.1**	68.8	**60.2**
	6	24.9	**18.8**	45.0	**33.6**	79.2	**69.0**
	8	30.1	**22.9**	53.7	**40.9**	88.8	**77.8**
	10	35.2	**27.2**	62.0	**48.2**	97.3	**86.3**
LocalNO	1	6.1	**4.8**	9.2	**7.3**	42.9	**33.2**
	2	8.0	**6.6**	13.1	**10.9**	49.0	**40.6**
	4	11.4	**9.4**	20.1	**16.6**	57.2	**47.8**
	6	15.1	**11.8**	27.2	**21.4**	68.6	**52.8**
	8	18.7	**14.2**	33.7	**25.8**	79.8	**57.8**
	10	22.7	**16.6**	40.6	**30.1**	90.3	**63.0**

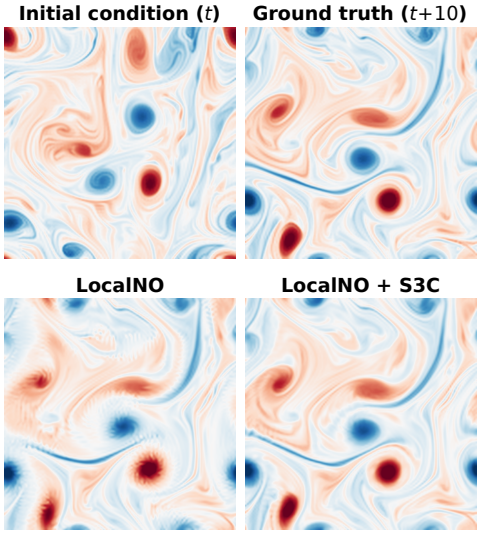


Figure 2: Navier-Stokes 10-step forecasting visualization with LocalNO backbone. Best viewed on screen. Interpolated for clarity.

4 Experimental Results

To validate our spectral cascade decoders, we evaluate our architectures on autoregressive forecasting and spectral extrapolation tasks. For the former, we consider autoregressive turbulent Navier-Stokes forecasting to demonstrate that spectral improvements via S3C can compound over autoregressive steps. For spectral extrapolation, we evaluate S3C on the PDE smoothing problem and S2WC on deterministic image super-resolution.

4.1 Navier-Stokes autoregressive forecasting

We consider autoregressive forecasting of the 2-d Navier-Stokes equations on the torus $(0, 2\pi)^2$. In particular, we consider the Kolmogorov flow form, given by

$$\partial_t u + u \cdot \nabla u - \frac{1}{\text{Re}} \Delta u = -\nabla p + \sin(4y)\hat{x}, \qquad \nabla \cdot u = 0, \tag{16}$$

where u, p indicate velocity and pressure, respectively, and Re represents the Reynolds number. We use the data and setup (e.g., converting to vorticity form) from Li et al. (2022) and select $\text{Re} = 5000$, which corresponds to a challenging turbulent flow. Models trained on this turbulent dataset have been observed to systematically attenuate high frequencies (Liu-Schiaffini et al., 2024). Given initial condition $u(\cdot, 0) = u_0$, we learn the evolution operator $u_0 \mapsto u(\cdot, h)$ for some fixed h. The learned evolution operator can be applied $k \in \mathbb{N}$ times to predict the state at arbitrary time kh.

We apply S3C to three different popular backbones, Fourier neural operator (FNO) (Li et al., 2020), U-Net (adapting the architecture used in (Gupta & Brandstetter, 2022; Berner et al., 2026)), and the local neural operator (LocalNO) (Liu-Schiaffini et al., 2024). We train separate models on resolutions 64×64 and 128×128. We evaluate using relative L^2, H^1, and a filtered L^2 loss HF_k that sums squared error terms only over all the modes with wavenumber larger than k.

From Table 1 and Figure 2, we see that S3C improves the backbone performance across almost all losses and backbones. Figure 2 visualizes how gains at higher frequencies at one-step prediction propagate to correct coarse structure at later steps. We observe that the gains are greater for LocalNO and U-Net since both have explicit local layers that aid in capturing high-frequencies. On the other hand, FNO's explicit mode truncation gives it an inductive bias towards lower frequencies. As such, S3C is not as effective at inducing the model to learn high frequencies. While S3C is backbone-agnostic, its performance capacity is ultimately determined by the backbone architecture. Our full results can be found in Appendix D.1, where *S3C brings up to 54% L^2 improvement*.

Table 2: Quantitative comparison of super-resolution methods with a SwinIR backbone. We report PSNR (dB) at $\times 2$ and $\times 4$; higher is better. **Bold** and underline denote the best and second-best result in each column, out of selected baselines. Numbers for prior methods taken from Cao et al. (2023), where † denotes entries reproduced by Lee & Jin (2022).

	DIV2K-val		Set5		Set14		B100		Urban100		Manga109	
Method	$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$
SwinIR	—	—	**38.35**	32.72	34.14	28.94	32.44	27.83	33.40	27.07	39.60	31.67
SwinIR-MetaSR	35.15†	29.33†	38.26	32.47	34.14	28.85	32.39	27.75	33.29	26.76	39.46	31.37
SwinIR-LIIF	35.17†	29.46†	38.28	32.73	34.14	28.98	32.39	27.84	33.36	27.15	39.57	31.71
SwinIR-LTE	35.24	29.51	38.33	**32.81**	**34.25**	**29.06**	32.44	27.86	33.50	27.24	**39.63**	**31.79**
SwinIR-S2WC (ours)	**35.28**	**29.54**	38.31	32.74	34.20	29.02	**32.47**	**27.88**	**33.52**	**27.28**	39.49	31.76

4.2 PDE SMOOTHING PROBLEM

We consider the PDE smoothing problem (Calvello et al., 2026) as a well-defined deterministic spectral extrapolation problem. We train 1-d and 2-d variants of S3C on the 1-d Kuramoto-Sivashinsky (KS) equation (see Appendix D.2) and 2-d Navier-Stokes from Section 4.1, respectively.

For these settings, let $u(\cdot, t)$ denote the true infinite-dimensional state at time t and let $v(\cdot, t)$ be a truncation of u onto a finite set of Fourier bases. Formally the smoothing problem consists of learning the mapping $v(\cdot, [0, t]) \mapsto u(\cdot, t)$. For tractability, we instead learn the approximation $v(\cdot, [t_0, t]) \mapsto u(\cdot, t)$ by conditioning the model on a previous finite history of discrete time steps. Practically, we consider mapping Fourier-downsampled low resolution to high resolution solutions of the KS and Navier-Stokes equations. We train models on the resolution $32 \mapsto 128$ map for KS and on resolution $32 \times 32 \mapsto 64 \times 64$ for Navier-Stokes.

Across several backbones, we find that S3C improves the performance on the smoothing problem, seen as more accurate Fourier spectra and sharper high-frequency details that were suppressed by the base model (See Appendix D.2 for more results). In capturing the challenging exponential decay of the KS equation, *we find S3C capable of extrapolating the tail beyond its training domain,* whereas the base FNO's spectrum tail quickly diverges beyond the training regime (Figure 4). Our full results, including the challenging Navier-Stokes setting, can be found in Appendix D.2.

4.3 IMAGE SUPER-RESOLUTION

To validate S2WC, we consider the arbitrary-scale deterministic super-resolution lineage of Chen et al. (2020); Lee & Jin (2022); Cao et al. (2023). Following their setup, we train on DIV2K-train (Agustsson & Timofte, 2017), containing 800 2K-resolution images. We also evaluate on the standard test sets DIV2K-val, Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2010), B100 (Martin et al., 2001), Urban100 (Huang et al., 2015), and Manga109 (Matsui et al., 2017). We train our model using the SwinIR (Liang et al., 2021) backbone without upsampling layers.

In our experiments, S2WC achieves competitive results against prior methods on $2\times$ and $4\times$ super-resolution (Table 2). From Figure 3, we also observe that S2WC is capable of reproducing the small details that are not immediately visible in the low resolution input (e.g., holes and mailbox numbers in the images). We want to emphasize that these image super-resolution benchmarks are highly competitive and achieving state-of-the-art performance depends on the choice of backbone and on heavily-engineered architectures. Our goal with S2WC is to demonstrate the expressive range of the cascade decoder design space, showing that advanced modules such as cross attention fit into our framework. As such, our aim is not to achieve state-of-the-art results but to perform competitively against well-established baselines. We compare with MetaSR (Hu et al., 2019), LIIF (Chen et al., 2020), and LTE (Lee & Jin, 2022), three foundational architectures in the arbitrary-scale super-resolution literature. Full experimental details are in Appendix D.3.

4.4 FREQUENCY LOSS ABLATION

We conduct an ablation study to compare S3C against an analogous Fourier amplitude loss term, which computes the MSE between the Fourier amplitudes for modes outside a radius k. We train

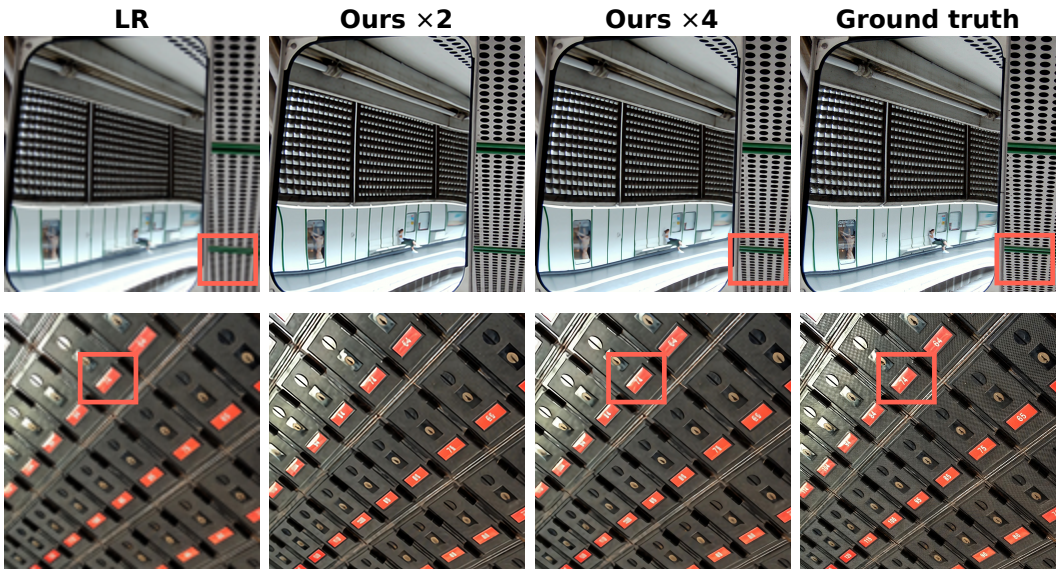


Figure 3: S2WC predictions on Urban100 center crops. Red boxes added for emphasis. Best viewed on screen and zoomed-in.

Table 3: Comparison against a high-frequency loss baseline (Freq.) with a LocalNO backbone on turbulent autoregressive Navier-Stokes forecasting at 64×64. All errors are reported $\times 10^{-2}$.

	rel. L^2		rel. H^1		rel. HF_{32}	
Steps	Freq.	**S3C**	Freq.	**S3C**	Freq.	**S3C**
1	7.9	**4.8**	12.4	**7.3**	52.8	**33.2**
2	10.6	**6.6**	17.8	**10.9**	60.1	**40.6**
4	15.5	**9.4**	27.2	**16.6**	69.8	**47.8**
6	21.3	**11.8**	37.9	**21.4**	82.9	**52.8**
8	27.3	**14.2**	48.8	**25.8**	97.7	**57.8**
10	33.0	**16.6**	58.6	**30.1**	110.0	**63.0**

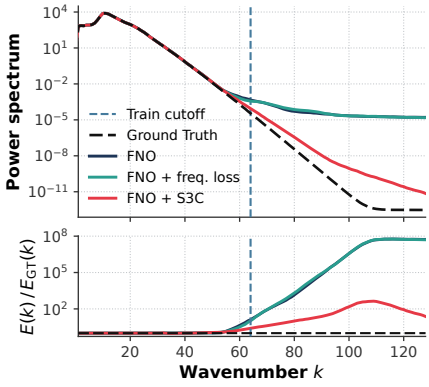


Figure 4: KS spectral extrapolation.

the model with a weighted combination of standard L^2 loss and this additional term. The results on NS forecasting are shown in Table 3 and on KS spectral extrapolation in Figure 4. S3C substantially outperforms the frequency loss across all metrics and autoregressive steps. While the frequency loss helps match the ground truth spectrum slightly in the training regime compared to the baseline LocalNO, it is unable to generalize to unseen modes. We also found S3C substantially easier to train and tune than the high-frequency loss for which one must additionally tune the loss weight. We provide thorough ablations for the spectral cascade decoder design space in Appendix F.

5 Conclusion and Future Work

In this paper, we argue that spectral attenuation in neural network outputs can be mitigated by incorporating the self-similarity of natural signals as an architectural inductive bias. We propose *self-similar cascade decoders*, a family of lightweight scale-recurrent modules that iteratively correct attenuated fine scales. We develop S3C and S2WC, two self-similar cascade decoder modules, and apply them to PDE and imaging problems. In future work, it would be interesting to further explore this design space for more expressive domain-specific architectures and new, parameter-efficient generative architectures. Please refer to Appendix B for more discussion and related works.

AI USE STATEMENT

In this work, we used generative AI tools for assistance in formalizing and checking proofs, as well as aiding in implementing methods and experimental pipelines. We have not used generative AI tools for developing conceptual frameworks, refining hypotheses, or interpreting results, and the remaining required disclosure tasks (e.g., assisting with translation) are not applicable to this work. We have reviewed AI-assisted work by checking and verifying LLM-generated code and thoroughly reviewing claims and proofs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We made substantial efforts to ensure reproducibility of our theoretical and empirical results. While the claims and theorems in the main text are, at times, stated informally (these places are carefully marked) for the sake of presentation, our appendix includes full assumptions, statements, and proofs. Our empirical results rely on several established PDE and image datasets. We followed published data processing and evaluation pipelines as closely as possible (reporting when deviations were necessary) to ensure reproducibility and comparisons with prior works. Our source code will be publicly released alongside the camera-ready version of the paper.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1122–1131. IEEE, 2017.

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024.

Julius Berner, Miguel Liu-Schiaffini, Jean Kossaifi, Valentin Duruisseaux, Boris Bonev, Kamyar Azizzadenesheli, and Anima Anandkumar. Principled approaches for extending neural architectures to function spaces for operator learning. *Nature Machine Intelligence*, pp. 1–9, 2026.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British machine vision conference (BMVC)*, 2012.

Edoardo Calvello, Elizabeth Carlson, Nikola Kovachki, Michael N Manta, and Andrew M Stuart. Operator learning for smoothing and forecasting. *arXiv preprint arXiv:2603.20359*, 2026.

Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1796–1807. IEEE, 2023.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. *arXiv preprint arXiv:2012.09161*, 2020.

Zixin Jessie Chen, Zhuo Chen, Archer Wang, Jeff Gore, William T Freeman, Congyue Deng, and Marin Soljačić. Everything at every scale: Scale-invariant diffusion with continuous super-resolution. *arXiv preprint arXiv:2605.26032*, 2026.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Sander Dieleman. Diffusion is spectral autoregression, 2024. *URL https://sander.ai/2024/09/02/spectral-autoregression. html*, 12, 2024.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

KS Gage. A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci*, 42:950–960, 1985.

Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pp. 349–356. IEEE, 2009.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

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.

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

Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1575–1584. IEEE, 2019.

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*, pp. 5197–5206, 2015.

Yuanhui Huang, Weiliang Chen, Wenzhao Zheng, Yueqi Duan, Jie Zhou, and Jiwen Lu. Spectralar: Spectral autoregressive visual generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15842–15852. IEEE, 2025.

Arnaud E Jacquin. Image coding based on a fractal theory of iterated contractive image transformations. *IEEE Transactions on image processing*, 1(1):18–30, 1992.

Ylva Jansson and Tony Lindeberg. Scale-invariant scale-channel networks: Deep networks that generalise to previously unseen scales. *Journal of Mathematical Imaging and Vision*, 64(5):506–536, 2022.

Siavash Khodakarami, Vivek Oommen, Aniruddha Bora, and George Em Karniadakis. Mitigating spectral bias in neural operators via high-frequency scaling for physical systems. *Neural Networks*, pp. 108027, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Masanori Kobayashi and Toshimitsu Musha. 1/f fluctuation of heartbeat period. *IEEE transactions on Biomedical Engineering*, (6):456–457, 1982.

Jean Kossaifi, Nikola Kovachki, Zongyi Li, David Pitt, Miguel Liu-Schiaffini, Robert Joseph George, Boris Bonev, Kamyar Azizzadenesheli, Julius Berner, Valentin Duruisseaux, et al. A library for learning neural operators. *arXiv preprint arXiv:2412.10354*, 2024.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1919–1928. IEEE, 2022.

Tianhong Li, Qinyi Sun, Lijie Fan, and Kaiming He. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. *Advances in Neural Information Processing Systems*, 35:16768–16781, 2022.

Zongyi Li, Samuel Lanthaler, Catherine Deng, Michael Chen, Yixuan Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. Scale-consistent learning for partial differential equations. *arXiv preprint arXiv:2507.18813*, 2025b.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. *Advances in Neural Information Processing Systems*, 36:67398–67433, 2023.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018.

Miguel Liu-Schiaffini, Julius Berner, Boris Bonev, Thorsten Kurth, Kamyar Azizzadenesheli, and Anima Anandkumar. Neural operators with localized integral and differential kernels. *arXiv preprint arXiv:2402.16845*, 2024.

S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2008. ISBN 9780080922027.

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 eighth IEEE international conference on computer vision. ICCV 2001*, volume 2, pp. 416–423. Ieee, 2001.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications*, 76(20):21811–21838, 2017.

Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE transactions on medical imaging*, 40(9):2306–2317, 2021.

Evgeny Novikov, A Novikov, David Shannahoff-Khalsa, B Schwartz, and J Wright. Scale-similar activity in the brain. *Physical Review E*, 56(3):R2387, 1997.

AM Obukhov. On the energy distribution in the spectrum of a turbulent flow. In *Dokl. Akad. Nauk SSSR*, volume 32, pp. 22–24, 1941.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

Oliver M Phillips. The equilibrium range in the spectrum of wind-generated waves. *Journal of Fluid Mechanics*, 4(4):426–434, 1958.

Michael Poli, Winnie Xu, Stefano Massaroli, Chenlin Meng, Kuno Kim, and Stefano Ermon. Self-similarity priors: Neural collages as differentiable fractal representations. *Advances in Neural Information Processing Systems*, 35:30393–30405, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Shaoxiang Qin, Fuyuan Lyu, Wenhui Peng, Dingyang Geng, Ju Wang, Xing Tang, Sylvie Leroyer, Naiping Gao, Xue Liu, and Liangzhu Leon Wang. Toward a better understanding of fourier neural operators from a spectral perspective. *arXiv preprint arXiv:2404.07200*, 2024.

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*, pp. 5301–5310. PMLR, 2019.

Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Revisiting over-smoothness in text to speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8197–8213, 2022.

Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

Mansi Sakarvadia, Kareem Hegazy, Amin Totounferoush, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W Mahoney. The false promise of zero-shot super-resolution in machine-learned operators. In *International Conference on Learning Representations*, volume 2026, pp. 69635–69688, 2026.

Assaf Shocher, Nadav Cohen, and Michal Irani. Zero-shot super-resolution using deep internal learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3118–3126. IEEE, 2018.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019.

Elias M Stein and Guido Weiss. *Introduction to Fourier analysis on Euclidean spaces*, volume 1, chapter 7. Princeton university press, 1971.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pp. 711–730. Springer, 2010.

A INTUITION

A.1 REPRESENTING THE WEIERSTRASS FUNCTION

The Weierstrass function is a well-known example of a *fractal curve* which everywhere continuous but nowhere differentiable. Intuitively, its roughness is an example the type of signal representable by our self-similar scale-recurrent procedure.

From Eq. 1, we can represent a signal f as

$$f = v_0 + \sum_{k=0}^{\infty} w_k = v_0 + w_0 + \sum_{k=1}^{\infty} S w_{k-1},$$

with the repeated self-similarity operator S. Take the special case when the integral operator S (see formal definition in Eq. 20) is defined with respect to the Dirac measure at 0, so that

$$(S\phi)(x) = \alpha(x)\phi(\lambda x).$$

Let $\mathcal{X} = [-1, 1] \subset \mathbb{R}$, fix $\alpha(x) = a \in (0, 1)$, and set $B(x) = \cos(x)$, and let $v_0 = w_{-1} = B$ such that $w_k = S^{k+1}B$. Since $\|S\|_{\text{op}} = a\lambda^{-1/2} < 1$ (for $\lambda > 1$), the affine map $T\phi = S\phi + B$ is a contraction and thus converges to a unique fixed point by the Banach fixed point theorem:

$$\Phi = \sum_{n=0}^{\infty} S^n B = \sum_{n=0}^{\infty} a^n \cos(\lambda^n x).$$

By Hardy's extension of Weierstrass's construction, this series is continuous everywhere and nowhere differentiable precisely when $a\lambda \geq 1$. Since $a < 1$, then $\lambda > 1/a > 1$. Thus, for the fixed point to maintain this "rough" property requires $\lambda > 1$, corroborating the result in Theorem 1.

A.2 DETAILS ON SYNTHETIC FRACTAL APPROXIMATION PROBLEM

Inspired by the Weierstrass function, we create a synthetic setting to demonstrate the benefits of a self-similarity architectural prior as well as the spectral attenuation of standard neural networks. We create a problem with the special case of self-similarity: exact self-similarity across scales. In other words, finer scales that *can be perfectly reconstructed* from the coarse ones. We define the ground truth analytically as a cascading series of arbitrary depth, with fine scales constructed from coarser ones. We consider the function approximation problem using implicit neural representations. Specifically, we construct a sequence of fields $\{\psi_0, \psi_1, \ldots, \psi_k, \ldots\}$ such that subsequent fields exhibit higher-frequency content, motivated by the Weierstrass function. We use a baseline SIREN, a k-scale-conditioned SIREN, and a self-similar SIREN (developed from our theory) to motivate self-similarity as a prior.

Problem setup In particular, we fix a smooth phase field $\Phi : [0, 1)^2 \to \mathbb{R}$, drawn as a single instance of a Gaussian random field with periodic boundary conditions. The field has power spectrum $\propto |\xi|^{-\alpha}$, with $\alpha = 5.5$ and sampled on a 2048×2048 grid, normalized to zero mean and unit variance and then scaled to standard deviation 0.15. In $x = (x_1, x_2) \in \mathbb{R}^2$, define the base

$$\Psi_0(x) = \cos(2\pi[x_1 + \Phi(x_1, x_2)]) \cos(2\pi[x_2 + \Phi(x_1, x_2)]). \tag{17}$$

The K-level target is given by

$$\Psi_K(x) = \sum_{k=0}^{K} c^k \Psi_0(2^k x) = \sum_{k=0}^{K} \psi_k, \tag{18}$$

with periodic boundary conditions on $[0, 1)^2$. We set $c = \frac{1}{2}$.

This target exhibits the following properties:

1. **Exact scale invariance:** The level-k target is the base w_0 evaluated at dyadically dilated coordinates and scaled by c^k, making the target field exhibit exact self-similarity in scale as $k \to \infty$.

2. **Power-law decay:** The decaying c^k amplitude drives the power spectrum to decay as a power law.
3. **Non-trivial structure:** For trivial $\Phi = 0$, the base reduces to a separable product of cosines, and the target Ψ_K becomes a fixed sum of axis-aligned harmonics. We initialize a non-trivial Φ to produce harder structure for the underlying SIREN to learn.

We train on scales $K \in \{0, 1, 2\}$ on a uniform grid of resolution 64×64 and evaluate the models on unseen scales $K > 2$ at the higher resolution of 1024×1024. This disentangles the resolution from the scale axis.

We compare three variants of SIREN (Sitzmann et al., 2020): (1) a baseline SIREN trained on $K = 2$, (2) a SIREN K-conditioned trained on $K \in \{0, 1, 2\}$, and (3) a self-similar SIREN derived from the theory introduced in Section 2. The SIREN architecture uses four sinusoidal layers of width 256 with $\omega_0 = 30$ and standard SIREN initialization. Coordinates are mapped to $[-1, 1]$.

Self-similar SIREN The network learns to estimate Ψ_0, the base generator of Ψ_K. Given the prediction $\widehat{\Psi}_0$, the predicted output is given by

$$\widehat{\Psi}_K(x) = \sum_{k=0}^{K} c^k \mathcal{P}\left[\widehat{\Psi}_0\right](2^k x),$$

where $\mathcal{P}$ is a low-pass filter. This takes the same series form as introduced in Section 2 and which defines ground truth operator (Eq. 18).

Following the form in Eq. 4, the detail fields $\psi_k := c^k \Psi_0(2^k x)$ satisfy $\psi_{k+1} = S\psi_k$, where

$$\alpha(x, \omega) = cp(\omega) \qquad (S\phi)(x) = c(p * \phi)(2x), \tag{19}$$

with $\lambda = 2$ and p a low-pass filter. Equivalently, the partial sums Ψ_K satisfy the affine recurrence $\Psi_{K+1} = T\Psi_K$ with $T\phi = S\phi + \Psi_0$, which admits a unique fixed point under the Banach fixed point theorem.

In practice, we implement this in Fourier space. Given Ψ_k, the operator S can be implemented as

1. Take the full discrete Fourier transform of Ψ_k, of shape $N \times N$.
2. Apply a low-pass filter in Fourier space by taking the centered $N/2 \times N/2$ frequencies.
3. Dilate and zero-insert add odd frequencies to shape $N \times N$.
4. Take the inverse Fourier transform and multiply by c to return Ψ_{k+1}.

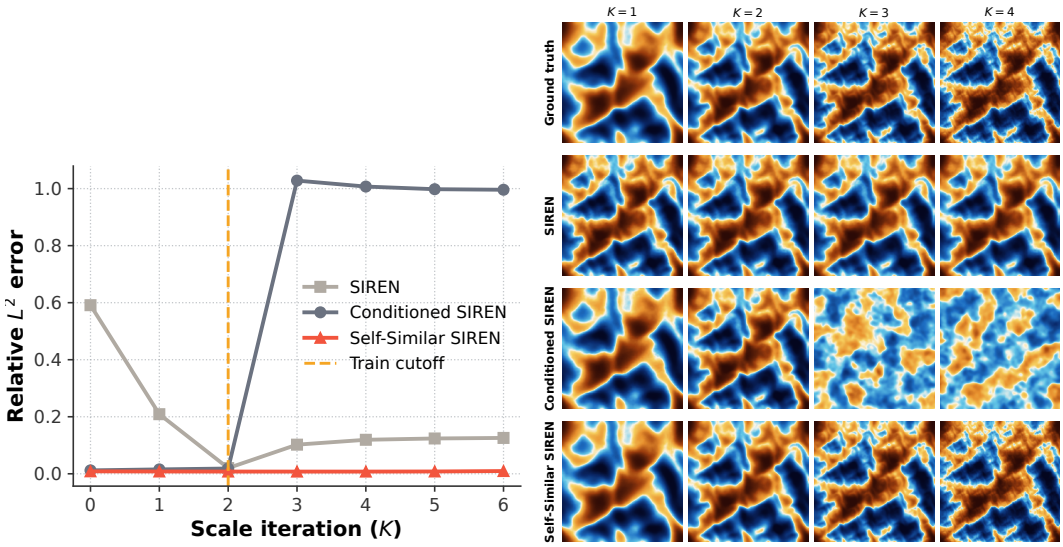
Conditional SIREN A third input channel is passed in containing $2K/K_{\text{max}} - 1$, a scale encoding, with $K_{\text{max}} = 2$ during training. This feature normalization maps the input to the range $[-1, 1]$. The conditional SIREN is never trained on $K > 2$ and is thus not expected to generalize outside of the training range for K. We include it for reference, as explicit conditioning is a common architectural trick for out-of-distribution generalization in neural architectures.

Training details We train for 20,000 steps using Adam on L^2, with initial learning rate of 10^{-3}, which is annealed to 10^{-7} during training. We use a weight decay of 10^{-6}.

Results and discussion In our experiments, we see that the self-similar SIREN achieves near-zero error across all scales, including those $K > 2$ unseen during training. In contrast, conditional SIREN is unable to generalize to unseen scales. See Figure 5.

In this toy experiment, we acknowledge that the comparison between the three models is not equivalent: self-similar SIREN structurally incorporates the series upon which the ground truth operator is defined. In contrast, conditional SIREN cannot reasonably be expected to generalize to unseen scales because the scale conditioning takes the model into a regime unseen during training. Likewise, the baseline SIREN (without scale conditioning) is trained only on $K = 2$ and thus will never output more high frequency content than what is contained at $K = 2$.

We thus emphasize that the results of these experiments are merely meant to be illustrative as a toy example. The conditional SIREN comparison shows that simple scale-conditioning cannot be



(a) (b)

Figure 5: Quantitative (a) and qualitative (b) comparisons between SIREN, conditional SIREN, and self-similar SIREN on the toy fractal problem. Best viewed on screen.

expected to generalize to arbitrarily fine scales. We provide the standard SIREN as a baseline to show the high-frequency features that models miss when their predicted spectra diverge from the truth beyond the training resolution, which we consistently observe among the baseline architectures in the PDE problems we consider in the main experiments (Section 4).

B Discussion and Related Works

B.1 Related Works

Coarse-to-fine scale hierarchies are well-explored in deep learning architectures. In generative modeling, a coarse-to-fine generation procedure has been explicitly embedded into generative pipelines (Ho et al., 2022; Li et al., 2025a; Tian et al., 2024), including across different bases (Huang et al., 2025; Chen et al., 2026). Traditional diffusion models have been observed to implicitly generate outputs in a coarse-to-fine manner (Dieleman, 2024). In this paper, we join the coarse-to-fine lineage but focus on self-similarity across scales as an explicit architectural prior, instantiated as a parameter-efficient and model-agnostic decoder module.

The scale-equivariant CNN literature (Cohen & Welling, 2016; Sosnovik et al., 2019; Worrall & Welling, 2019) is rich and closely related to our coarse-to-fine operator via filter dilation. Some prior work has also considered generalization to unseen scales (Jansson & Lindeberg, 2022). While scale-equivariant architectures primarily target scale equivariance of internal representations with respect to the inputs, we instead explore the problem of synthesizing unseen finer modes from a smooth input dominated by coarser modes.

Self-similarity has been employed for natural image super-resolution both classically (Glasner et al., 2009) and in neural architectures (Shocher et al., 2018). This prior is sufficiently strong to also enable image encoding and compression (Jacquin, 1992; Poli et al., 2022). In neural PDE solvers, scale-consistency has also been used for generalization across scales and PDE parameters (Li et al., 2025b).

Neural operators (Kovachki et al., 2023; Azizzadenesheli et al., 2024) generalize neural networks to function-to-function mappings. Several works have identified spectral attenuation at high frequencies (Qin et al., 2024; Khodakarami et al., 2025), particularly those unseen during training. While accurately predicting unseen modes is ill-posed in general (Sakarvadia et al., 2026), we show that a self-similarity inductive bias enables extrapolation to unseen modes for particular problems.

B.2 Discussion and Limitations

In this paper, we identify *spectral attenuation* as a consistent phenomenon across fields and domains where neural networks struggle to resolve fine spatial scales. We observe that this phenomenon is inconsistent with the approximate self-similarity exhibited by many natural signals. As such, we consider incorporating self-similarity across scales as an explicit architectural inductive bias. Our motivation is not to replace existing architectures, but rather to supplement them in a way that their attenuated fine scales can be corrected without affecting the accurately predicted coarse scales.

We present this perspective as orthogonal and complementary to the successful prior techniques for mitigating spectral attenuation. We take our Navier-Stokes autoregressive forecasting case in Section 4 and Appendix D.1 as an example. Comparing the performance of FNO and LocalNO, we see that the latter outperforms the former on one-step prediction. Indeed, LocalNO (Liu-Schiaffini et al., 2024) itself introduces two parallel branches to FNO to enable it to better capture high frequency detail. In our experiments, we observe that S3C improves LocalNO by a much larger margin than it does for FNO. As such, the performance of S3C depends, of course, on the performance and characteristics of its backbone. This is an example of how the architectural modifications of LocalNO from FNO complement with S3C to compound the performance improvements each had over the baseline FNO architecture.

Naturally, since the self-similar cascade decoders are a family of decoder modules attached to a backbone predictor, they cannot guarantee performance improvements independently of their backbones. Indeed, the self-similarity prior assumes that coarse scales are correctly predicted by the backbone. While the spectral attenuation phenomenon is empirically observed to correctly learn coarse scales at the expense of fine scales, a severely underfitting backbone that cannot learn the coarse scales will likely not see a significant performance boost with the proposed modules.

While most natural signals do not exhibit exact self-similarity, the theoretical and practical relaxations we introduce (e.g., state-dependent recurrence and S3C acting on amplitudes only) still enable substantial performance improvements over baselines in our experiments. However, our current deterministic architectures may not be most useful in mitigating spectral attenuation coming from one-to-many mappings, as a generative framework is often the natural scheme in such a case. A generative extension of our proposed self-similar cascade decoders is an interesting direction we leave for future work.

C Full theoretical results

C.1 Full statement and proof of Theorem 1

C.1.1 Preliminaries

Recall that the Fourier transform of valid ϕ on $\mathbb{R}^d$ is given by

$$\widehat{\phi}(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \phi(x)\, e^{-i\xi\cdot x}\, dx.$$

We write the full Fourier transform of (kernel) $\alpha(x,\omega)$ on $\mathbb{R}^d \times \mathbb{R}^d$ as

$$\widetilde{\alpha}(\eta_x, \eta_\omega) = (2\pi)^{-d} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \alpha(x,\omega)\, e^{-i\eta_x\cdot x} e^{-i\eta_\omega\cdot\omega}\, d\omega\, dx.$$

For $\gamma > 1$ and $k \in \mathbb{N}$, we formally define the Fourier bands and detail spaces:

$$B_k = \{\xi \in \mathbb{R}^d : \gamma^k \le \|\xi\| \le \gamma^{k+1}\}$$
$$W_k = \{f \in L^2(\mathbb{R}^d) : \operatorname{ess\,supp} \widehat{f} \subseteq B_k\}.$$

For $0 < \varepsilon < \gamma - 1$, also recall:

$$I_k^\varepsilon = \{\xi : (1+\varepsilon)\gamma^k \le \|\xi\| \le \gamma^{k+1}\}.$$

Finally, recall the definition of our generalized operator

$$(S_{\alpha,\lambda}\phi)(x) = \int_{\mathbb{R}^d} \alpha(x,\omega)\phi(\lambda x - \omega)\, d\mu(\omega), \tag{20}$$

where we may drop the dependence of S on α, λ if it is clear from context. For the primary results and the less formal presentation in the main text, we default to the Lebesgue measure $d\mu(\omega) = d\omega$, but some cases we prove in this appendix rely on non-Lebesgue choices of $d\mu(\omega)$.

Finally, we define the frequency support of α in the spatial dimensions by

$$P_x(\alpha) = \overline{\{\eta_x : \exists\ \eta_\omega \text{ s.t. } \widetilde{\alpha}(\eta_x, \eta_\omega) \neq 0\}}. \tag{21}$$

We say that α is R-bandlimited if

$$P_x(\alpha) \subseteq \overline{B}(0, R),$$

the closed ball of radius R.

For the results below, we select the convenient sufficient regularity assumption that $\alpha \in \mathcal{S}(\mathbb{R}^{2d})$, the Schwartz space on $\mathbb{R}^{2d}$. For $\phi \in W_k$, then $\hat{\phi}$ is supported on a finite-measure set and is in L^1 by Cauchy-Schwarz. With the regularity assumptions on α, we have that $\alpha \in L^1$ as well, so these conditions on $\hat{\phi}$ and α makes the resulting integrals absolutely integrable and enables the use of Fubini's theorem.

Although we are formally considering $\alpha \in \mathcal{S}(\mathbb{R}^{2d})$, the following results also hold for broader kernel classes whenever $S_{\alpha,\lambda}$ and the Fourier transforms can be appropriately defined. Furthermore, we formulate the analysis here on $\mathbb{R}^d$ for simplicity. However, in practical cases we may consider $\mathcal{X} \subset \mathbb{R}^d$. In these cases, we can consider $\phi = F|_{\mathcal{X}}$, where F is an ambient function defined over $\mathbb{R}^d$. The operator acting on f can also be defined by the restriction of the operator acting on F to $\mathcal{X}$. The results thus follow on the ambient signal F with no distinction. If one instead wishes to use $L^2(\mathcal{X})$ as the intrinsic space for the underlying signal, one may wish to replace the Fourier transform with a different domain-dependent operator, whereby the resulting theory is analogous but not identical verbatim to this Euclidean setting.

C.1.2 Primary results

We now present and prove the formal statement for Theorem 1.

Lemma 1. *Let $\alpha \in \mathcal{S}(\mathbb{R}^{2d})$ and suppose that $\phi \in L^2(\mathbb{R}^d)$ has finite-measure Fourier support. Then,*

$$\widehat{S\phi}(\xi) = \int_{\mathbb{R}^d} \widetilde{\alpha}(\xi - \lambda\zeta, \zeta)\widehat{\phi}(\zeta)\, d\zeta.$$

Proof. We have that

$$\phi(\lambda x - \omega) = (2\pi)^{-d/2} \int \widehat{\phi}(\zeta)\, e^{i\lambda\zeta\cdot x} e^{-i\zeta\cdot\omega}\, d\zeta.$$

Substituting this into the definition of $S\phi$, taking the Fourier transform, and applying Fubini's theorem, we get

$$\begin{aligned}\widehat{S\phi}(\xi) &= (2\pi)^{-d} \int \widehat{\phi}(\zeta) \left[\int\int \alpha(x,\omega)\, e^{-i(\xi-\lambda\zeta)\cdot x} e^{-i\zeta\cdot\omega} d\omega\, dx\right] d\zeta \\ &= \int \widetilde{\alpha}(\xi - \lambda\zeta, \zeta)\widehat{\phi}(\zeta)\, d\zeta,\end{aligned}$$

by definition of the joint transform of α. □

Next, consider some measurable source and target frequency sets $E, F \subseteq \mathbb{R}^d$, respectively. Define

$$d_{\lambda,E}(\xi) = \operatorname{dist}(\xi, \lambda E) := \inf_{\zeta\in E} \|\xi - \lambda\zeta\|$$

and the routing gaps

$$\begin{aligned}\Delta_{\text{some}}(F, E; \lambda) &= \operatorname{ess\,inf}_{\xi\in F} d_{\lambda,E}(\xi) \\ \Delta_{\text{all}}(F, E; \lambda) &= \operatorname{ess\,sup}_{\xi\in F} d_{\lambda,E}(\xi).\end{aligned}$$

Then, we can prove the following theorem.

Theorem 2. *Let* $\operatorname{ess\,supp} \widehat{\phi} \subseteq E$ *and let* α *be* R*-bandlimited in* x*. Then,* $\widehat{S_{\alpha,\lambda}\phi}(\xi) = 0$ *for almost every* ξ *such that* $d_{\lambda,E}(\xi) > R$*. Equivalently,*

$$\operatorname{ess\,supp} \widehat{S_{\alpha,\lambda}\phi} \subseteq \overline{\lambda E + \overline{B}(0,R)}.$$

Consequently,

1. *If* $\|\mathbb{1}_F \widehat{S\phi}\|_{L^2} > 0$*, then* $R \geq \Delta_{some}(F, E; \lambda)$
2. *If* $\widehat{S\phi} \neq 0$ *almost everywhere on* F*, then* $R \geq \Delta_{all}(F, E; \lambda)$.

Proof. If $d_{\lambda,E}(\xi) > R$, then for every $\zeta \in E$ we have

$$\|\xi - \lambda\zeta\| > R,$$

and since α is R-spatially bandlimited, we have

$$\widetilde{\alpha}(\xi - \lambda\zeta, \zeta) = 0.$$

Using this and the fact that $\operatorname{ess\,supp} \widehat{\phi} \subseteq E$, from Lemma 1 we thus have that $\widehat{S_{\alpha,\lambda}\phi}(\xi) = 0$. We can then show the two lower bounds on R by contradiction:

1. If instead $R < \Delta_{\text{some}}(F, E; \lambda)$, then $R < d_{\lambda,E}(\xi)$ for almost every $\xi \in F$. Thus, $\widehat{S\phi}(\xi) = 0$ for almost every $\xi \in F$, so $\|\mathbb{1}_F \widehat{S\phi}\|_{L^2} = 0$, a contradiction. Hence, $R \geq \Delta_{\text{some}}(F, E; \lambda)$.
2. If instead $R < \Delta_{\text{all}}(F, E; \lambda)$, then by definition of the essential supremum, the set $F_R = \{\xi \in F : d_{\lambda,E}(\xi) > R\}$ has positive measure. Thus, for every $\xi \in F_R$, we have $\widehat{S\phi}(\xi) = 0$, which contradicts the claim that $\widehat{S\phi} \neq 0$ almost everywhere on F. Hence, $R \geq \Delta_{\text{all}}(F, E; \lambda)$.

□

This leads us to an important corollary.

Corollary 1. *Suppose that* $0 < \lambda \leq 1$ *and let* $\phi \in W_{k-1}$ *with* α R*-bandlimited spatially. If*

$$\|\mathbb{1}_{I_k^\varepsilon} \widehat{S\phi}\|_{L^2} > 0,$$

then $R \geq \varepsilon\gamma^k$.

Proof. For $\zeta \in B_{k-1}$,

$$\|\lambda\zeta\| \leq \lambda\|\zeta\| \leq \gamma^k.$$

For $\xi \in I_k^\varepsilon$,

$$\|\xi\| \geq (1 + \varepsilon)\gamma^k.$$

Then,

$$\|\xi - \lambda\zeta\| \geq \|\xi\| - \lambda\|\zeta\| \geq \varepsilon\gamma^k.$$

Applying Theorem 2 with $E = B_{k-1}$, $F = I_k^\varepsilon$ and recognizing that

$$\Delta_{\text{some}}(F, E; \lambda) \geq \varepsilon\gamma^k$$

proves the corollary, since $R \geq \Delta_{\text{some}}(F, E; \lambda)$. □

In particular, this gives us the special case of standard $\alpha(x, \omega) = \alpha(\omega)$ and "spatially-varying" convolutions. The former is incapable of routing energy to the target frequency band, while the latter requires geometrically scaling x-bandwidth.

So far, we have treated the scaling factor γ in the Fourier bands B_k to be potentially distinct from the λ scaling factor in the operator definition $S_{\alpha,\lambda}$. Now, we prove that these must coincide to similarly enable the kernel to remain bandlimited. First, we define

$$D_k(\lambda) = \sup_{\xi \in B_k} \operatorname{dist}(\xi, \lambda B_{k-1}).$$

Proposition 1. *With the above definition of $D_k(\lambda)$, we have that*

$$D_k(\lambda) = \max\left\{\gamma^{k-1}(\lambda-\gamma)_+, \gamma^k(\gamma-\lambda)_+\right\},$$

with sequence $\{D_k(\lambda)\}_k$ only bounded in k if and only if $\lambda = \gamma$.

Proof. The distance from a point of radius ρ to an annulus with inner and outer radii u, v respectively is

$$(u-\rho)_+ + (\rho - v)_+,$$

where $(\cdot)_+$ denotes $\max\{\cdot, 0\}$. $D_k(\lambda)$ takes a maximum of this expression over B_k. This maximum must occur at one of the radial endpoints of B_k: either γ^k or γ^{k+1}. Substituting in $u = \lambda\gamma^{k-1}$ and $v = \lambda\gamma^k$ and checking both of these endpoints against cases $\lambda = \gamma, \lambda > \gamma, \lambda < \gamma$, we get

$$D_k(\lambda) = \max\{\gamma^{k-1}(\lambda-\gamma)_+, \gamma^k(\gamma-\lambda)_+\},$$

as desired.

If $\lambda < \gamma$, then the second term above grows as γ^k. If $\lambda > \gamma$, the first term grows as γ^{k-1}. Hence $D_k(\lambda)$ remains bounded if and only if $\lambda = \gamma$. □

Remark 1. Treating I_k^ε as the target instead of B_k enables a slight relaxation of γ within an ε-dependent ball around λ.

Corollary 2. *Suppose a single $R < \infty$, independent of k, is used by kernel α, and for every k there exists some $\phi_k \in W_{k-1}$ such that $\widehat{S_{\alpha,\lambda}\phi_k} \neq 0$ almost everywhere on B_k. Then, $\gamma = \lambda$.*

Proof. Theorem 2 gives that $R \geq D_k(\lambda)$ for all k. Proposition 1 shows that $D_k(\lambda)$ is unbounded unless $\gamma = \lambda$. □

Finally, we will now prove the existence of a kernel α with spatial bandwidth $R > 0$ that satisfies the desired spectral routing properties.

Theorem 3. *Let $\lambda > 1$ and suppose the target family F_k satisfies $F_k \subseteq \lambda B_{k-1}$ for all k. Then, for any $R > 0$, there exists some kernel $\alpha_R \in \mathcal{S}(\mathbb{R}^{2d})$ with R-spatial bandlimit, such that for every k there exists some $\phi_k \in W_{k-1}$ with*

$$\widehat{S_{\alpha_R,\lambda}\phi_k}(\xi) > 0$$

for almost every $\xi \in F_k$. In particular, this can be applied to $F_k = B_k$.

Proof. Let us choose some real, even $h \in C_c^\infty(\mathbb{R}^d)$, the space of infinitely differentiable functions with compact support, such that $h \geq 0$, $\operatorname{supp} h \subseteq \overline{B}(0, R)$, and $h(p) > 0$ for $\|p\| < R/2$. If we let $\chi = \mathcal{F}^{-1}h$, with $\mathcal{F}$ denoting the Fourier transform, and let $\alpha(\omega)$ be a Schwartz function whose Fourier transform is strictly positive everywhere (e.g., a Gaussian). Define the factorized kernel

$$\alpha_R(x, \omega) = \chi(x)\alpha(\omega),$$

and thus

$$\widetilde{\alpha}_R(p, \zeta) = h(p)\widehat{\alpha}(\zeta),$$

which has spatial bandwidth at most R. For each k, define ϕ such that

$$\widehat{\phi}_k = \mathbb{1}_{B_{k-1}}.$$

From Lemma 1, we thus have that

$$\widehat{S_{\alpha_R,\lambda}\phi_k}(\xi) = \int_{B_{k-1}} h(\xi - \lambda\zeta)\widehat{\alpha}(\zeta)\, d\zeta.$$

For almost every $\xi \in F_k \subseteq \lambda B_{k-1}$, $\zeta_0 = \xi/\lambda$ lies within B_{k-1}. Thus, a sufficiently-small positive-measure neighborhood of ζ_0 remains in B_{k-1} and satisfies $\|\xi - \lambda\zeta\| < R/2$. By assumption, both h and $\widehat{\alpha}$ are positive there, so the integral is positive. Hence, $S_{\alpha_R,\lambda}$ routes positive energy for almost all target $\xi \in F_k$. □

Thus, we have shown that a dilation factor of $\lambda > 1$ is necessary for an R-bandlimited kernel α to cover the full next band B_k from an input in B_{k-1}. We have also shown that there exists such a valid kernel α that transports the energy in this way.

Since this result is independent of k, in the scale-recurrent paradigm, we share the kernel α (and hence the operator $S_{\alpha,\lambda}$) between scales.

C.2 SCALE RECURRENCE DERIVATIONS AND CONVERGENCE PROOF

Expanding on Section 2.3, we first show the derivation of the coefficients of Sw_{k-1} in W_k space, which shows that we can express the operator S at scale k as a matrix $A^{(k)}$ mapping between the coefficients between W_{k-1} and W_k with respect to a basis of each space.

Given the basis $\{\varphi_m^{(k-1)}\}$ for W_{k-1}, we can express $w_{k-1} = \sum_{m=1}^{M} c_m^{(k-1)} \varphi_m^{(k-1)}(x)$ for some coefficients $\{c_m^{(k-1)}\}$. Substituting this into the definition of operator S, we get

$$\begin{aligned}(Sw_{k-1})(x) &= \int_{\mathcal{X}} \alpha(x,\omega) \left[\sum_{m=1}^{M} c_m \varphi_m^{(k-1)}(\lambda x - \omega)\right] d\omega \\ &= \sum_{m=1}^{M} c_m^{(k-1)} \left[\int_{\mathcal{X}} \alpha(x,\omega) \varphi_m^{(k-1)}(\lambda x - \omega)\, d\omega\right].\end{aligned}$$

Applying the projection operator P_{W_k} by taking the inner product with $\{\varphi_n^{(k)}\}_{n=1}^{N}$, basis of W_k, we find the new coefficients in W_k space:

$$\begin{aligned}c_n^{(k)} = \left\langle \varphi_n^{(k)}, Sw_{k-1} \right\rangle_{L^2} &= \int_{\mathcal{X}} \varphi_n^{(k)*}(x)(Sw_{k-1})(x)\, dx \\ &= \sum_{m=1}^{M} c_m^{(k-1)} \underbrace{\left[\int_{\mathcal{X}} \int_{\mathcal{X}} \varphi_n^{(k)*}(x)\, \alpha(x,\omega)\, \varphi_m^{(k-1)}(\lambda x - \omega)\, d\omega\, dx\right]}_{A_{n,m}^{(k)}}, \quad (22)\end{aligned}$$

where $A^{(k)} \in \mathbb{C}^{N \times M}$ represents $\tilde{S}_k = P_{W_k} S$, the structured linear transformation that maps the coefficients of W_{k-1} to those of W_k.

We now show sufficient conditions on $A^{(k)}$ for the series (Eq. 2) to converge in L^2. Let $c^{(k)} \in \mathbb{C}^N$ be a vector of coefficients for $w_k \in W_k$ when projected onto the orthonormal basis $\{\varphi_n^{(k)}\}_{n=1}^{N}$, and let $c^{(0)}$ be a vector of coefficients for the base signal $v_0 \in V_0$. We know by Parseval's identity that

$$\|w_k\|_{L^2} = \|c^{(k)}\|_{\ell_2}.$$

Given matrix $A^{(k)}$, the finite-dimensional operation of the operator S over the coefficient space at scale k, we can bound $\|c^{(k)}\|_{\ell_2}$ by

$$\|c^{(k)}\|_{\ell_2} \leq \|A^{(k)}\|_{\text{op}} \cdot \|c^{(k-1)}\|_{\ell_2},$$

and unrolling recursively,

$$\|c^{(k)}\|_{\ell_2} \leq \left(\prod_{j=1}^{k} \|A^{(j)}\|_{\text{op}}\right) \|c^{(0)}\|_{\ell_2}.$$

For the series $\sum_k w_k$ to converge absolutely, we require

$$\begin{aligned}\sum_{k=1}^{\infty} \|w_k\|_{L^2} &= \sum_{k=1}^{\infty} \|c^{(k)}\|_{\ell^2} < \infty \\ \sum_{k=1}^{\infty} \|c^{(k)}\|_{\ell^2} &\leq \|c^{(0)}\|_{\ell_2} \sum_{k=1}^{\infty} \left(\prod_{j=1}^{k} \|A^{(j)}\|_{\text{op}}\right),\end{aligned}$$

so by the ratio test the sufficient condition for convergence in L^2 is

$$\limsup_{k \to \infty} \|A^{(k)}\|_{\text{op}} < 1. \quad (23)$$

The convergence proof in L^2 in coefficient space is easily adapted for input-dependent operators $A(v_k)$. The state-independent proof does not directly apply in this case because it holds only for

a predefined trajectory of $(v_0, v_1, \dots)$ and $(A^{(1)}, A^{(2)}, \dots)$. However, if we institute the uniform bound

$$\sup_{v \in \mathcal{F}} \|A(v)\|_{\text{op}} \leq \beta < 1, \tag{24}$$

then we get that

$$\|c^{(k)}\|_{\ell_2} \leq \|A(v_{k-1})\|_{\text{op}} \cdot \|c^{(k-1)}\|_{\ell_2} \leq \beta \|c^{(k-1)}\|_{\ell_2} \leq \beta^k \|c^{(0)}\|_{\ell_2},$$

and thus the series $\sum_{k=1}^{\infty} \|w_k\|_{L^2}$ converges.

C.3 Derivations of scale-recurrent special cases

C.3.1 Fourier basis

For the sake of presentation, let $\mathcal{X} = \mathbb{T}^d$. Suppose we define the detail space W_k as being induced by a dyadic set of frequency bands in the L^2 basis. Then, the elements of the basis of W_{k-1} can be expressed as $\varphi_m^{(k-1)}(x) = (2\pi)^{-d/2} e^{i\xi_m x}$.

If we define $\lambda = 2$ in the definition of operator S (Eq. 4), then

$$\varphi_m^{(k-1)}(2x - \omega) = (2\pi)^{-d/2} e^{i\xi_m(2x-\omega)} = (2\pi)^{-d/2} e^{-i\xi_m \omega} e^{i(2\xi_m)x} = e^{-i\xi_m \omega} \varphi_n^{(k)}(x),$$

where $\xi_n = 2\xi_m$. As such, the matrix $A^{(k)} \in \mathbb{C}^{N \times M}$ takes the following form:

$$\begin{aligned}
A_{n,m}^{(k)} &= (2\pi)^{-d} \int_{\mathcal{X}} \int_{\mathcal{X}} \varphi_n^{(k)*}(x)\, \alpha(x, \omega)\, e^{-i\xi_m \omega} e^{i(2\xi_m)x}\, d\omega\, dx \\
&= (2\pi)^{-d} \int_{\mathcal{X}} \int_{\mathcal{X}} e^{-i\xi_n x}\, \alpha(x, \omega)\, e^{-i\xi_m \omega} e^{i(2\xi_m)x}\, d\omega\, dx \\
&= (2\pi)^{-d} \int_{\mathcal{X}} \underbrace{\left(\int_{\mathcal{X}} \alpha(x, \omega) e^{-i\xi_m \omega}\, d\omega \right)}_{\widehat{\alpha}(x,\, \xi_m)} e^{-i(\xi_n - 2\xi_m)x}\, dx.
\end{aligned}$$

We observe that the full Fourier transform of $\alpha(x, \omega)$ is given by

$$\widetilde{\alpha}(\xi_x, \xi_\omega) := (2\pi)^{-d} \int_{\mathcal{X}} \int_{\mathcal{X}} \alpha(x, \omega) e^{-i\xi_x x} e^{-i\xi_\omega \omega}\, d\omega\, dx,$$

which takes the precise form of $A_{n,m}^{(k)}$ above. Aligning terms, we get that

$$A_{n,m}^{(k)} = \widetilde{\alpha}(\xi_n - 2\xi_m, \xi_m). \tag{25}$$

C.3.2 Wavelet basis

Given mother wavelet $\psi(x)$, in the wavelet basis, we have

$$\varphi_m^{(k-1)}(x) = 2^{(k-1)/2}\, \psi(2^{k-1}x - m) \qquad \varphi_n^{(k)} = 2^{k/2}\, \psi(2^k x - n).$$

We once again set $\lambda = 2$ in Eq. 4, but we now take μ in Eq. 20 to be the counting measure on the spatial nodes at scale $k-1$, and define $\omega_\Delta = \Delta / 2^{k-1}$ for $\Delta \in \mathbb{Z}$. As such, we see that

$$\begin{aligned}
\varphi_m^{(k-1)}(2x - \omega_\Delta) &= 2^{(k-1)/2}\, \psi\left(2^{k-1} \left(2x - \frac{\Delta}{2^{k-1}} \right) - m \right) \\
&= \frac{1}{\sqrt{2}} \left[2^{k/2}\, \psi(2^k x - (\Delta + m) \right] \\
&= \frac{1}{\sqrt{2}}\, \varphi_{\Delta+m}^{(k)}(x).
\end{aligned}$$

Given this, we can now derive the entries of matrix $A^{(k)} \in \mathbb{R}^{N \times M}$:

$$\begin{aligned}
A_{n,m}^{(k)} &= \int_{\mathcal{X}} \varphi_n^{(k)}(x) \left[\sum_{\Delta \in \mathbb{Z}} \alpha\left(x, \frac{\Delta}{2^{k-1}} \right) \frac{1}{\sqrt{2}} \varphi_{\Delta+m}^{(k)}(x) \right] dx \\
&= \frac{1}{\sqrt{2}} \sum_{\Delta \in \mathbb{Z}} \int_{\mathcal{X}} \alpha\left(x, \frac{\Delta}{2^{k-1}} \right) \varphi_n^{(k)}(x) \varphi_{\Delta+m}^{(k)}(x)\, dx.
\end{aligned} \tag{26}$$

In the case of orthogonal wavelets with mutually exclusive compact supports (i.e., Haar), this operator becomes sparse:

$$A^{(k)}_{n,m} = \frac{1}{\sqrt{2}} \underbrace{\int_{\mathcal{X}} \alpha\left(x, \frac{n-m}{2^{k-1}}\right) |\varphi^{(k)}_n(x)|^2 \, dx}_{=: \, \bar{\alpha}_n\left(\frac{n-m}{2^{k-1}}\right)} = \frac{1}{\sqrt{2}} \, \bar{\alpha}_n\left(\frac{n-m}{2^{k-1}}\right), \tag{27}$$

where the interpretation of $\bar{\alpha}_n(\omega)$ is as a function that routes energy betwteen W_{k-1} to W_k:

$$\bar{\alpha}_n(\omega) = \mathbb{E}_{x \sim |\varphi_n|^2}[\alpha(x, \omega)]. \tag{28}$$

C.3.3 Transpose convolutions

For simplicity, suppose $\mathcal{X} = [0, 1)$. The results extend straightforwardly to higher spatial dimensions. To construct standard deep learning operators, we let the physical space be spanned by the zeroth order B-spline $\phi(x) = \mathbb{1}_{[0,1)}(x)$. The coarse and fine approximation spaces V_{k-1} and V_k are spanned by the following basis functions:

$$\varphi^{(k-1)}_m(x) = 2^{(k-1)/2} \phi(2^{k-1}x - m)$$
$$\varphi^{(k)}_n(x) = 2^{k/2} \phi(2^k x - n).$$

We select the same $\lambda = 2$ and counting measure μ as before and restrict $\omega_\Delta = \frac{\Delta}{2^{k-1}}$ for $\Delta \in \mathbb{Z}$. Furthermore, we enforce translation equivariance by restricting α to be a convolutional kernel:

$$\alpha(x, \omega) = \kappa(x - \omega).$$

We now slightly redefine the operator S as mapping between approximation spaces V_{k-1} and V_k instead of between detail spaces W_{k-1} and W_k, although this difference is mostly cosmetic to align with standard deep learning architectures for the sake of this illustration. To construct the routing matrix $A^{(k)}_{n,m}$, we must compute $\varphi^{(k-1)}_m(2x - \omega_\Delta)$:

$$\begin{aligned}
\varphi^{(k-1)}_m\left(2x - \omega_\Delta\right) = \varphi^{(k-1)}_m\left(2x - \frac{\Delta}{2^{k-1}}\right) &= 2^{(k-1)/2} \, \phi\left(2^{k-1}\left(2x - \frac{\Delta}{2^{k-1}}\right) - m\right) \\
&= 2^{(k-1)/2} \, \phi(2^k x - \Delta - m) \\
&= \frac{1}{\sqrt{2}} \, \varphi^{(k)}_{\Delta+m}(x),
\end{aligned}$$

which we find is proportional to the bases of V_k. Thus, $A^{(k)}_{n,m}$ takes the form

$$\begin{aligned}
A^{(k)}_{n,m} &= \frac{1}{\sqrt{2}} \int_{\mathcal{X}} \varphi^{(k)}_n(x) \left[\sum_{\Delta \in \mathbb{Z}} \kappa\left(x - \frac{\Delta}{2^{k-1}}\right) \varphi^{(k)}_{\Delta+m}(x)\right] dx \\
&= \frac{1}{\sqrt{2}} \, 2^k \int_{I_n} \kappa\left(x - \frac{n-m}{2^{k-1}}\right) \, dx
\end{aligned}$$

because $\varphi^{(k)}_n(x)\varphi^{(k)}_{\Delta+m}(x) = 0$ everywhere unless $\Delta = n - m$, and where $I_n = [n2^{-k}, (n+1)2^{-k})$ is the support of $\varphi^{(k)}_n$. Performing the change of variables $u = x - n2^{-k}$, we get

$$A^{(k)}_{n,m} = \frac{1}{\sqrt{2}} \, 2^k \int_0^{2^{-k}} \kappa\left(u - (n-2m)2^{-k}\right) \, du. \tag{29}$$

Intuitively, $A^{(k)}_{n,m}$ represents the energy transferred from pixel m on the coarse grid to pixel n on the fine grid. This energy is computed as an average over the fine grid pixel size of the continuous convolution kernel κ.[5]

[5]This has the benefit of reducing aliasing of the kernel κ, which point evaluations of continuous convolution kernels (e.g., Liu-Schiaffini et al. (2024)) are susceptible to.

If we define the discrete kernel K as

$$K^{(k)}[j] = \frac{1}{\sqrt{2}}\, 2^k \int_0^{2^{-k}} \kappa\left(u - j2^{-k}\right)\, du,$$

then the coefficients (which can be seen in this case as a discretization of the underlying function) $c_n^{(k)}$ that represent w_k space can be expressed as

$$c_n^{(k)} = \sum_m K^{(k)}[n - 2m] \cdot c_m^{(k-1)},$$

which corresponds precisely to a common deep learning transposed convolution with stride 2. However, as opposed to standard deep learning, the discrete kernel K is constructed from continuous kernel κ, thus generalizing the resolution-dependent transposed convolution to a resolution-agnostic *transposed convolution operator*.

Transposed convolutions have been shown to lead to "checkerboard" artifacts in computer vision applications. Odena et al. (2016) showed that these artifacts can be removed by using resize-convolutions, where a smooth interpolation is applied before the convolution in the transposed convolution. Under our construction, this merely corresponds to a different order of B-spline chosen for the basis functions $\varphi_n^{(k)}$.

C.3.4 Fractal image coding

If we generalize the $\lambda x - \omega$ dilation in Eq. 4 to a general coordinate transformation $\tau(x, \omega)$ and take the integral $d\mu(\omega)$ over some space Ω, we can recover the standard discrete construction from fractal image coding (Jacquin (1992) and its neural variant Poli et al. (2022)).

Specifically, let $\mathcal{X} = [0, 1)^d$ and an image/signal is represented by $\phi : \mathcal{X} \to \mathbb{R}$. Let $\{R_k\}_{k=1}^K$ and $\{D_n\}_{n=1}^N$ be the domain and range cells, respectively, both of which partition $\mathcal{X}$. To be concrete, suppose the side of each range cell has length h and the side of each domain cell has length λh. Let $k(x)$ be the index of the range cell that contains point $x \in \mathcal{X}$.

Define $\Omega = \{1, \ldots, N\}$ and define $\tau(x, n)$ to be the inverse of the map that rescales D_n to the size of $R_{k(x)}$ and places it at the range cell's location. Define $\alpha(x, n) = \gamma_{k(x),n} a_{k(x),n}$, which is piecewise constant in x (i.e., constant on each range cell) and indexed by n. Define the bias $B(x) = \sum_n \gamma_{k(x),n} b_{k(x),n}$, which is also piecewise constant.

This construction aligns with both the classical partitioned iterated function system (Jacquin, 1992) and the modern neural variant (Poli et al., 2022). In the classical setting, mixing weights $\gamma_{k,n}$ are binary, summing to 1 over all n. In the neural variant, these are dense and learned.

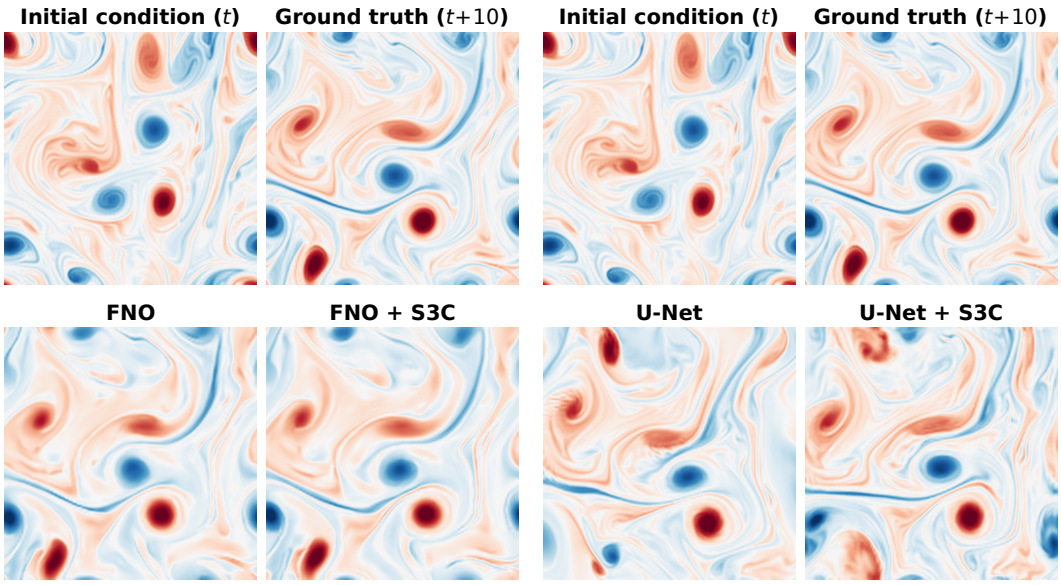
D Additional experimental results

In this section, we present our full experimental results for the Navier-Stokes autoregressive temporal evolution problem, the Navier-Stokes and Kuramoto-Sivashinsky (KS) smoothing problems, and the image super-resolution problem.

D.1 Navier-Stokes autoregressive forecasting

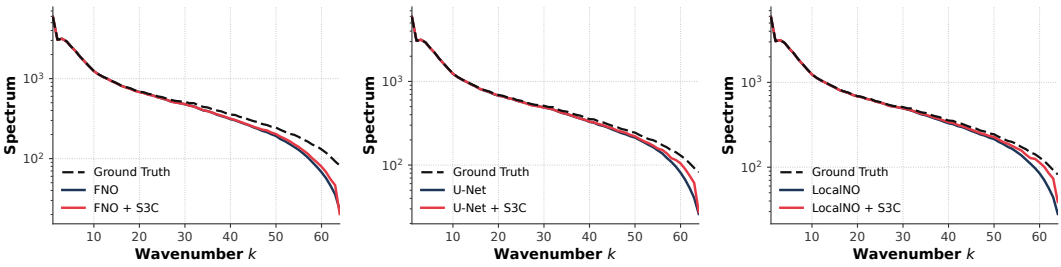
To complement the LocalNO results in Figure 2, Figure 6 shows the analogous 10-step field visualizations for FNO and U-Net at resolution 128×128. Across all three, we see that S3C's improvements at the fine scales propagates to the coarse scales on autoregressive rollout. We observe that U-Net underperforms on the 10-step rollout compared to FNO and LocalNO, although S3C does help to qualitatively improve the performance over the base model. Close inspection of Figure 2 also shows that S3C helps smooth and correct high-frequency noise that is present in the base LocalNO's 10-step prediction.

Figure 7 compares the power spectra of a one-step Navier-Stokes prediction with FNO, U-Net, and LocalNO backbones at 64×64. While all models capture the low frequencies well, the models supplemented with S3C show better convergence to the tail than the base models. However, this



(a) FNO (b) U-Net

Figure 6: Navier-Stokes 10-step forecasting visualization with FNO and U-Net backbones. Best viewed on screen. Interpolated for clarity.



(a) FNO (b) U-Net (c) LocalNO

Figure 7: Power spectra over a one-step Navier-Stokes autoregressive rollout at 64×64.

is also mediated by the base model's architectural limitations towards the high frequencies. For instance, from Figures 7a and 7c, we see that the base FNO performs worse on high frequencies than the base U-Net and LocalNO. This can be attributed to the architectural mode truncation in the FNO architecture, forcing the skip connections and nonlinearities to carry the high frequencies across layers (Liu-Schiaffini et al., 2024). In contrast, U-Net and LocalNO both contain local (hence spectrally global) operations, thus allowing the model to greater emphasize the high frequencies. These power spectra results also largely corroborate the results seen in Table 1, particularly in H^1 and HF_{32}. We find that LocalNO outperforms U-Net, possibly due to its disentangling of local and global feature learning through explicit local and global branches (Liu-Schiaffini et al., 2024).

Table 4 shows the numerical results for the Navier-Stokes rollout experiment trained at 128×128 resolution. At this more challenging resolution, we again find that the impact of S3C is bottlenecked by the base model's capacity. For instance, we observe that FNO does not benefit greatly from S3C, with the performance quite similar between the two at most step counts. U-Net does see a minor improvement from S3C that compounds as the number of autoregressive steps increases. In the FNO case, we attribute this to FNO's explicit mode truncation, which makes the high frequencies dependent primarily on nonlinearities and pointwise skip connections. We hypothesize that these prevent the FNO from producing correct mode phases for higher frequencies, thus leading to a gap in the high frequency loss that grows with autoregressive steps.

Table 4: Turbulent Navier-Stokes autoregressive forecasting error on 128×128 data. All errors are reported $\times 10^{-2}$. Backbones have $O(10^7)$ parameters; S3C adds $O(10^4)$.

		rel. L^2		rel. H^1		rel. HF_{64}	
Backbone	Steps	Base	**+S3C**	Base	**+S3C**	Base	**+S3C**
FNO	1	14.6	14.6	**35.6**	36.2	85.4	**78.5**
	2	**17.8**	18.0	**43.2**	44.5	90.3	**87.7**
	4	22.9	22.9	**54.4**	55.0	**96.0**	96.9
	6	28.3	**27.7**	65.4	**64.7**	**102.4**	104.4
	8	33.9	**32.9**	76.5	**74.6**	**108.0**	111.0
	10	39.3	**38.2**	86.4	**84.1**	**111.6**	115.8
U-Net	1	18.8	**18.1**	42.5	**41.8**	**81.2**	82.2
	2	31.7	**30.4**	68.7	**67.7**	**104.4**	106.8
	4	50.2	**46.6**	96.9	**94.2**	**121.2**	122.1
	6	63.9	**57.8**	111.5	**107.3**	128.3	**126.0**
	8	74.7	**66.7**	119.7	**115.1**	131.0	**126.7**
	10	83.7	**74.0**	124.3	**120.1**	131.3	**126.2**
LocalNO	1	9.0	**5.6**	20.2	**12.4**	53.9	**43.1**
	2	14.9	**8.0**	34.6	**18.5**	69.0	**50.4**
	4	27.2	**12.6**	60.3	**29.7**	95.1	**59.9**
	6	34.8	**17.9**	73.1	**42.0**	110.4	**71.4**
	8	43.0	**23.5**	86.7	**53.9**	125.1	**83.4**
	10	51.2	**29.5**	97.8	**64.8**	130.5	**92.8**

However, LocalNO (the strongest backbone tested in our experiments) showed significant improvements in all three metrics over all autoregressive steps. For instance, the relative L^2 between the base LocalNO and LocalNO + S3C dropped from $9.0 \cdot 10^{-2}$ to $5.6 \cdot 10^{-2}$. This pattern continued for all steps evaluated. In this setting, we observed S3C decrease L^2 error up to 54% compared to the base model, while only contributing about 0.1% of the total parameter count. We attribute this improvement entirely to S3C embedding a self-similarity prior across scales that otherwise has to be implicitly learned by the backbone model.

D.2 PDE SMOOTHING PROBLEM

The main text (Section 4.2) presents the smoothing problem results for the 1-d Kuramoto-Sivashinsky (KS) equation, defined by the initial value problem

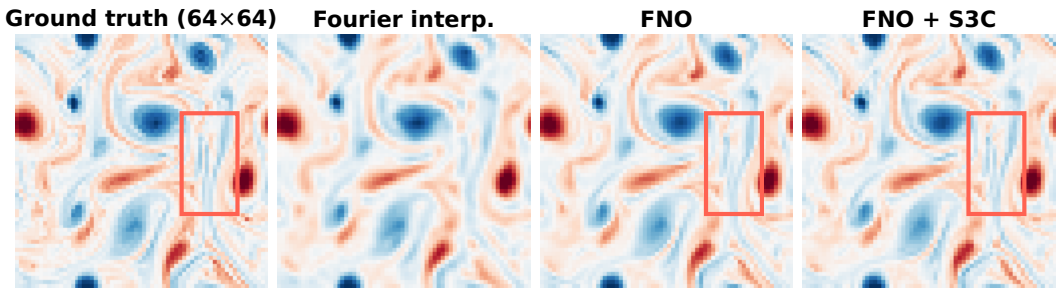
$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial x^4}, \tag{30}$$

where $x \in [0, 32\pi]$ with periodic boundary conditions.

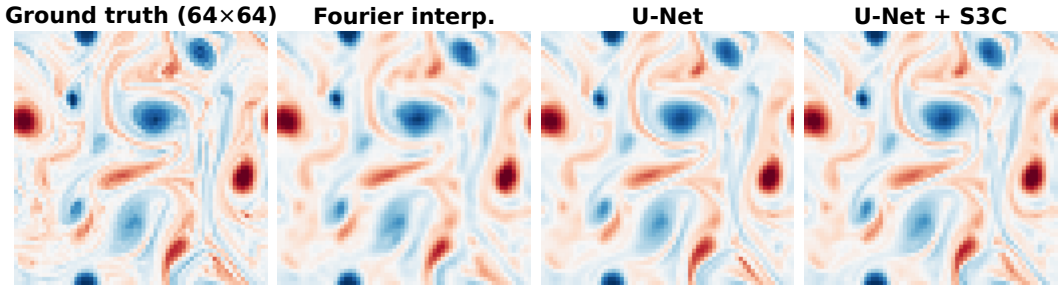
We demonstrated that S3C is not only capable of encouraging the baseline model to better match the ground truth power spectrum at seen resolutions, it is also capable of much stronger spectral extrapolation (sometimes referred to as zero-shot super-resolution) compared to the baselines.

In this section, we present results for the $32 \times 32 \to 64 \times 64$ Navier-Stokes smoothing problem. This setting is possibly the most challenging PDE problem we approach in this paper. This can be visually seen in the difference between the Fourier interpolation and ground truth channels in Figure 8. This figure shows an example uncurated field from FNO, U-Net, and LocalNO. The red boxes in Figure 8a are added for emphasis. We observe that S3C in general does help the base model to better capture high frequency content in the output space. While this is somewhat visible in the power spectra (Figure 9), the improvement is of smaller degree than for the KS equation and the Navier-Stokes temporal rollouts. For instance, for U-Net, the impact of S3C is negligible over the performance of the base model itself.

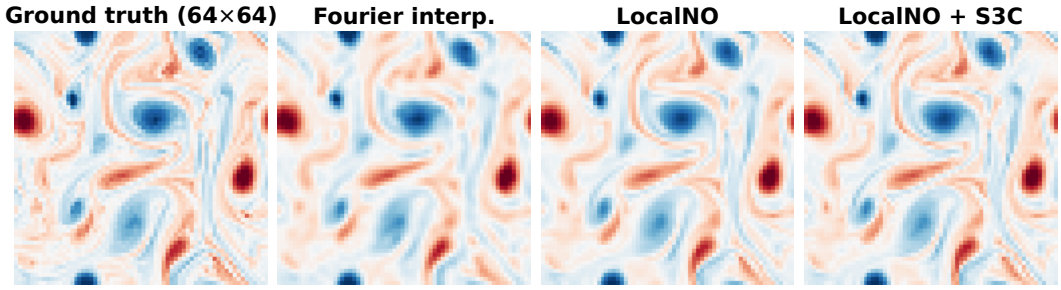
Due to the difficulty of this task, we emphasize that our goal is not to attain zero error on this task but to explore and characterize the settings where S3C can improve over base models. This result demonstrates that S3C cannot freely improve the spectral tail prediction of any model; its performance is dependent on that of its backbone. In this sense, it is *complementary* to other architectural choices used to mitigate spectral attenuation explored in the literature, not a substitute.



(a) FNO

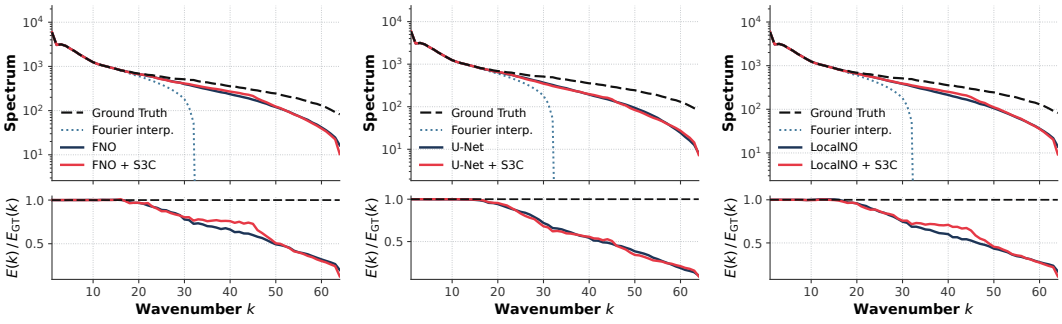


(b) U-Net



(c) LocalNO

Figure 8: Uncurated predicted fields for $32 \times 32 \to 64 \times 64$ Navier-Stokes smoothing problem. Red boxes added for emphasis.



(a) FNO (b) U-Net (c) LocalNO

Figure 9: Power spectra for $32 \times 32 \to 64 \times 64$ Navier-Stokes smoothing problem.

D.3 Image super-resolution

We closely follow the experimental setup and published code of Cao et al. (2023), including training for 1000 epochs on L1 loss and training on patch crops of the full 2K images. However, we make a

Table 5: Full quantitative comparison of super-resolution methods with SwinIR. We report PSNR (dB); higher is better. **Bold** and underline denote the best and second-best, respectively. Numbers for baselines taken from Cao et al. (2023), where † denotes entries reproduced by Lee & Jin (2022). Asterisk (*) denotes that our model was not explicitly trained $\times 3$ and these results are bicubic downsampling from $\times 4$.

	DIV2K-val			Set5			Set14			B100			Urban100			Manga109		
Method	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$
SwinIR	—	—	—	**38.35**	**34.89**	32.72	34.14	30.77	28.94	32.44	29.37	27.83	33.40	29.29	27.07	39.60	34.74	31.67
SwinIR-MetaSR	35.15†	31.40†	29.33†	38.26	34.77	32.47	34.14	30.66	28.85	32.39	29.31	27.75	33.29	29.12	26.76	39.46	34.62	31.37
SwinIR-LIIF	35.17†	31.46†	29.46†	38.28	34.87	32.73	34.14	30.75	28.98	32.39	29.34	27.84	33.36	29.33	27.15	39.57	34.68	31.71
SwinIR-LTE	35.24	**31.50**	29.51	38.33	**34.89**	**32.81**	**34.25**	**30.80**	**29.06**	32.44	**29.39**	27.86	33.50	**29.41**	27.24	**39.63**	**34.79**	**31.79**
SwinIR-S2WC (ours)	**35.28**	31.45*	**29.54**	38.31	34.73*	32.74	34.20	30.69*	29.02	**32.47**	29.36*	**27.88**	**33.52**	29.27*	**27.28**	39.49	34.44*	31.76

few minor exceptions. Since S2WC predicts each patch densely, we compute the loss over the full patch instead of subsampling. By construction, the wavelet transform in S2WC forces output scales to be powers of 2. As such, we train on $2\times$ and $4\times$ super-resolution (S2WC shares weights across each), and we consider only these two scales for evaluation. We do not consider generalization to scales beyond $4\times$ because we found the two-step autoregressive decoder loop to be too shallow in practice for the model to learn to generalize to unbounded loop iterations.

In this paper, we compare with MetaSR (Hu et al., 2019), local implicit image function (LIIF) (Chen et al., 2020), and local texture estimator (LTE) (Lee & Jin, 2022) as backbone-agnostic modules. While there are more recent and better-performing architectures (e.g., Cao et al. (2023)) these three architectures are foundational cornerstones of the arbitrary-scale super-resolution literature. Furthermore, in the pursuit of state-of-the-art results, we believe many recent architectures have become highly-engineered. While such advancements are valuable, in this paper, our goal is explicitly to avoid head-to-head comparisons with such methods. Our aim is to demonstrate the expressive range of the cascade decoder design space and show how a simple cross-attention instantiation in wavelet space (without heavy architectural engineering) can perform competitively (10/12 top-2 and 6/12 top-1 among the twelve columns in Table 2) with the foundational papers of the field. Furthermore, when comparing only to MetaSR and LIIF on $\times 2$ and $\times 4$ super-resolution, S2WC achieves top-1 performance 11/12 times.

Our full quantitative comparison against these baselines can be found in Table 5. As noted before, SwinIR-S2WC outperforms (or achieves highly competitive results against) many of the tested baselines, including on challenging datasets such as Urban100. For completeness, we also present $\times 3$ super-resolution performance on all test sets, despite the fact that S2WC was never explicitly supervised on data of this scale. The $\times 3$ predictions provided are computed by bicubic interpolation from native $\times 4$ predictions. Despite not being trained directly at $\times 3$ scale, S2WC still outperforms baseline models such as MetaSR several times.

Figures 10, 11, 12, and 13 show sample SwinIR-S2WC predictions for random DIV2K validation, Urban100, BSDS100, and Manga109 crops, respectively. We note that S2WC is capable of resolving small detail that is not immediately visible in the low-resolution input. For example, in Figure 3, at the low resolution input, the holes in the red box are not distinguishable from black vertical lines. However, in the $\times 2$ and $\times 4$ model outputs, the model is able to accurately predict that these are in fact holes (see, for instance, Figure 6 in Cao et al. (2023)). In the second row of Figure 3, our model is correctly able to reconstruct some of the numbers on the mailboxes.

E Implementation details

E.1 Navier-Stokes temporal rollout

We train all models for 100 epochs with an initial learning rate of 10^{-3} for the neural operators and 10^{-4} for the U-Net that anneals to zero following a cosine schedule during training. We use a weight decay of 10^{-4}, a batch size of 16, and Adam (Kingma & Ba, 2014) for optimization. Following Li et al. (2022), we truncate the first 100 seconds of each Navier-Stokes trajectory as the transient phase before the attractor. We train all models on L^2 loss.

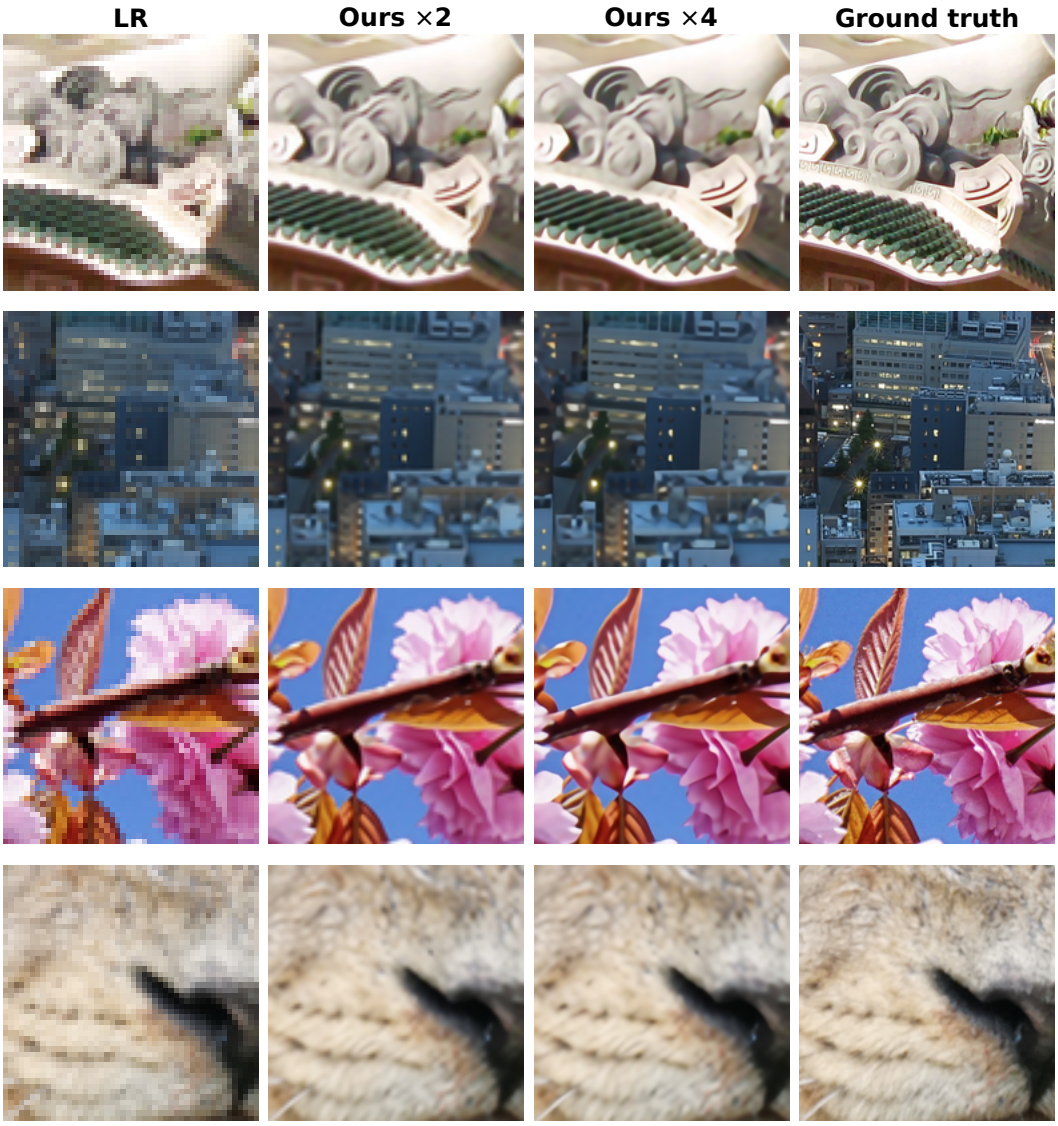


Figure 10: Sample S2WC predictions on random DIV2K validation center crops.

We use 4 layers for FNO and LocalNO, as well as 128 lifting and projection channels. We use 32 modes for FNO and 16 for LocalNO and 64 hidden channels for FNO and 128 for LocalNO, to approximately match their parameter counts. We concatenate the $[0, 1]^2$ grid as positional encoding for both models. For LocalNO, we use differential and local integral layers in parallel with the FNO-style global convolution in all four layers, following the hyperparameter defaults in Liu-Schiaffini et al. (2024). We use the FNO and LocalNO implementations in Kossaifi et al. (2024).

The U-Net architecture is implemented using the Berner et al. (2026) adaptation of the architecture used in Gupta & Brandstetter (2022). We use 40 hidden channels, GELU activations, and 3×3 kernels. We use one block per resolution, and downsample twice to reach the bottleneck layer. The channels at the intermediate and bottleneck resolutions are 80 and 160, respectively.

The S3C module is the same for all architectures: a low mode cutoff of 16, a 2-layer MLP with 32 hidden units per layer, and we apply the module in the latent space.

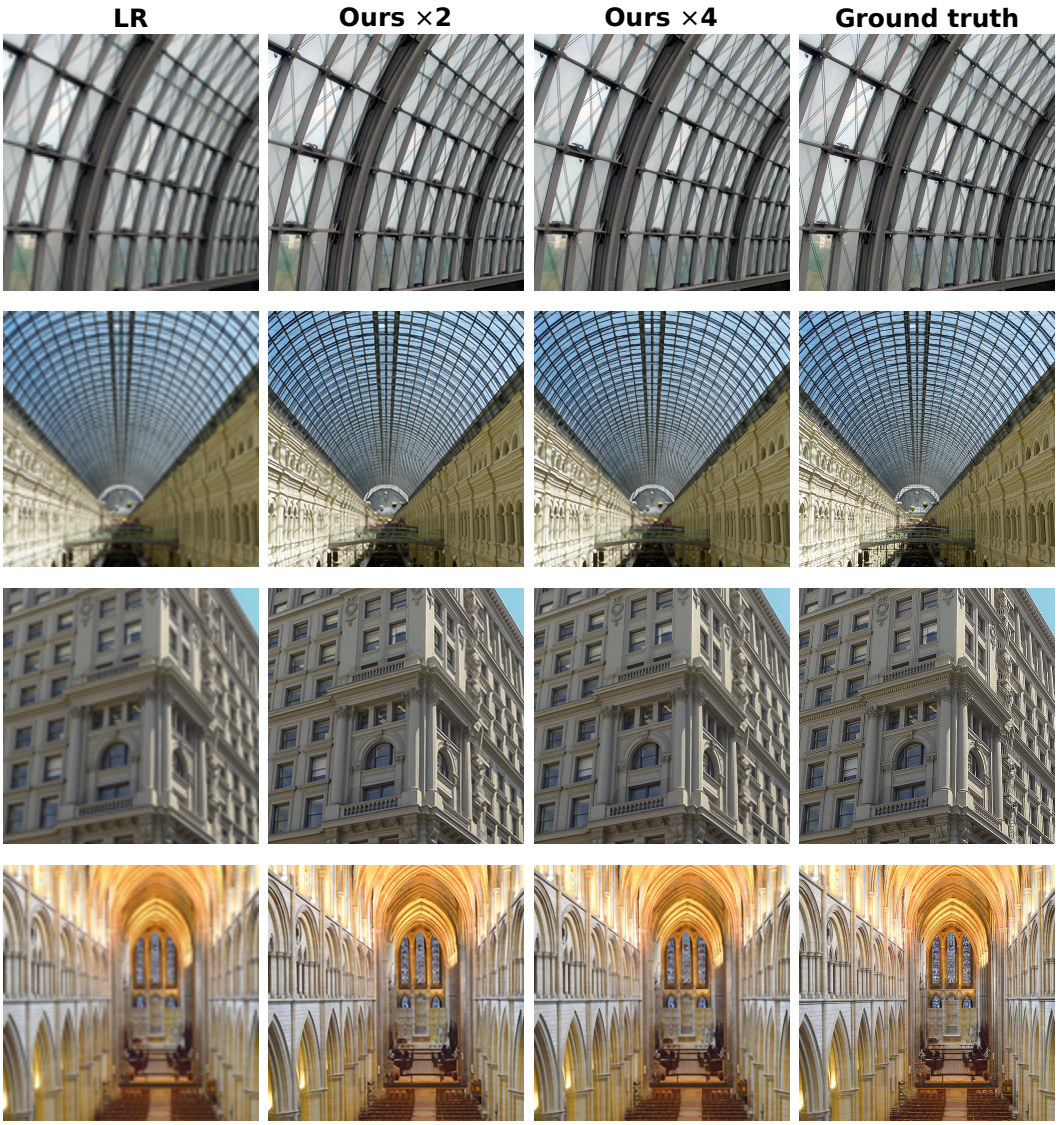


Figure 11: Sample S2WC predictions on random Urban100 center crops.

E.2 PDE SMOOTHING PROBLEM

We first present the hyperparameters for the Kuramoto-Sivashinsky smoothing problem. While the primary S3C module replaces the model's predicted amplitudes beyond the mode cutoff, we find that for Kuramoto-Sivashinsky's rapidly decaying power spectrum tail, relaxing this condition into a multiplicative modulation of the existing spectrum helps the model better learn the decay rate in practice. The backbone FNO architecture uses 32 modes, 64 channels. The S3C module is applied in the output space, using a 2-layer MLP with 64 hidden units to predict the multiplicative amplitude modulation for the next 16 modes. We train both models for 100 epochs with an initial learning rate of 10^{-3} annealed using a cosine scheduler. We use a weight decay of 10^{-4}, trained on L^2 with Adam.

The hyperparameters used for the Navier-Stokes smoothing problem are identical to those used for the Navier-Stokes temporal evolution problem, except that the S3C module is applied in the output space. For the neural operator architectures, we query the output at a higher resolution than the input resolution.

We condition all models using the last 4 time steps as history.

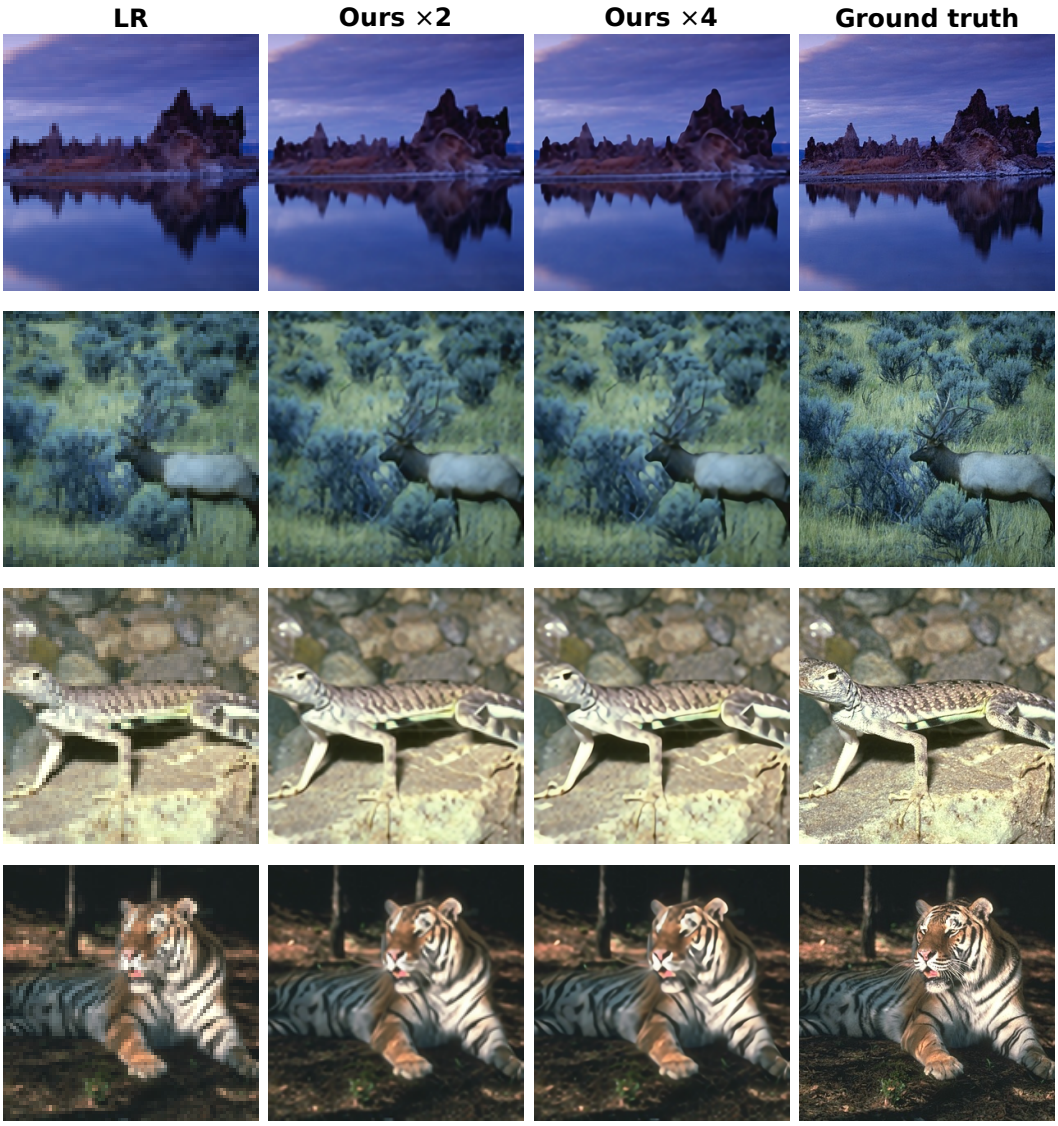


Figure 12: Sample S2WC predictions on random BSDS100 center crops.

E.3 IMAGE SUPER-RESOLUTION

We follow the DIV2K training protocol described in Cao et al. (2023), training on 48×48 low-resolution patches. We train our model using a SwinIR (Liang et al., 2021) backbone on $2\times$ and $4\times$ scales. On DIV2K-val we follow the evaluation protocol in LIIF (Chen et al., 2020), reporting PSNR. We generate low-resolution inputs using MATLAB-style bicubic downsampling at both training and evaluation time. On Set5, Set14, BSDS100, Urban100, and Manga109 we follow the published evaluation protocol in CiaoSR (Cao et al., 2023). Our evaluation protocol was confirmed with released LIIF checkpoints.

We use an embedding dimension of 180, 4 attention heads, and 8×8 attention patches. We make two minor engineering additions to our architecture: (1) We find it useful to compute queries, keys, and values by applying a 3×3 depthwise convolution with dilation dependent on prediction scale (e.g., larger dilation for $4\times$ than $2\times$), followed by a 1×1 pointwise convolution. (2) We also apply a squeeze-and-excitation (Hu et al., 2018) channel attention layer at the end of each self-similar decoder cascade block. We use a randomly-initialized SwinIR backbone with the upsampling layers removed, following (Cao et al., 2023).

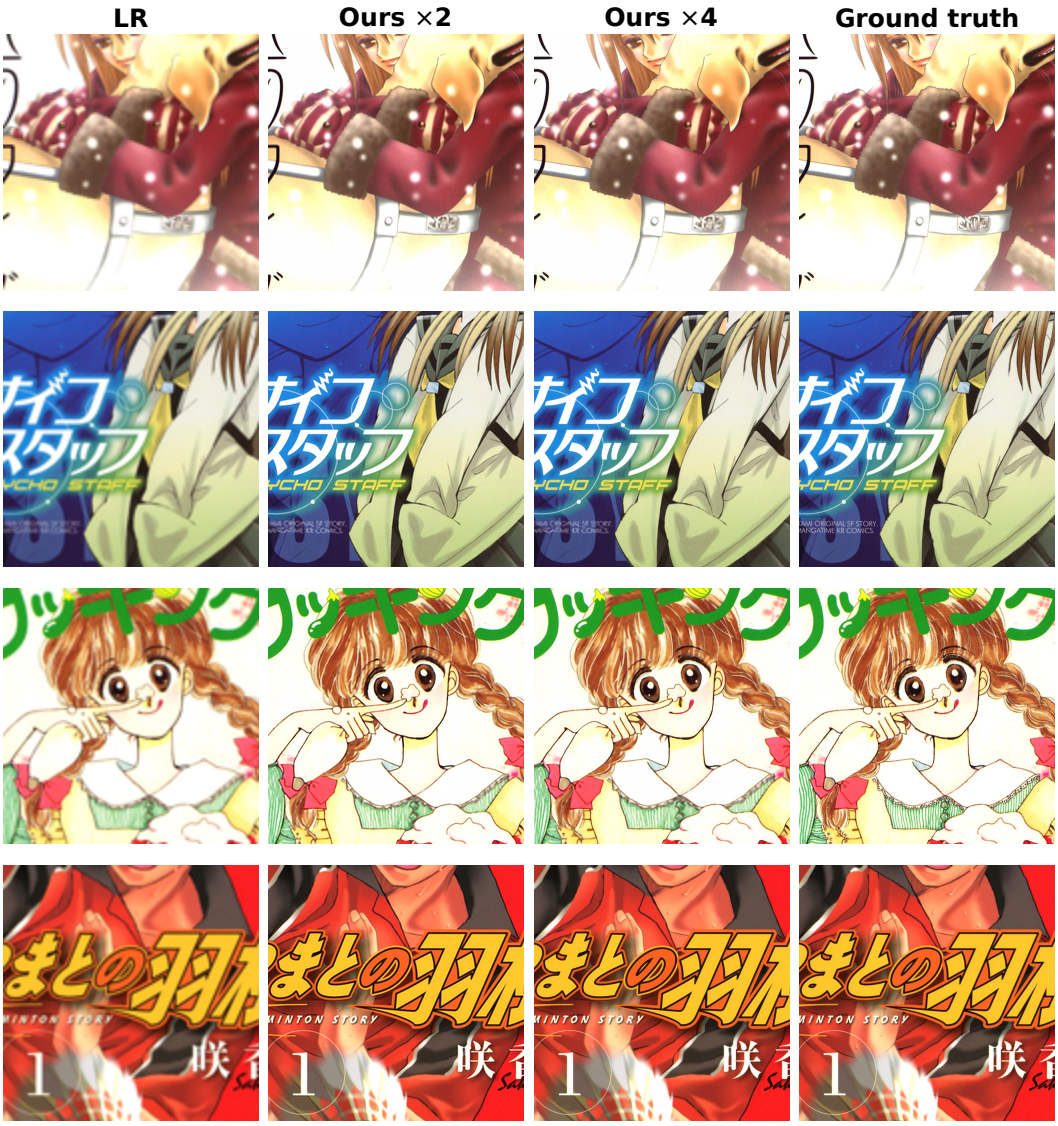


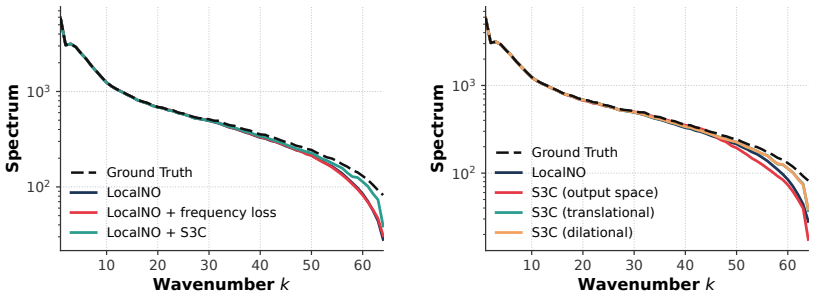
Figure 13: Sample S2WC predictions on random Manga109 center crops.

In S2WC, queries are formed from the current approximation v_k, whose $H \times W$ positions are precisely the four children of each of the $H/2 \times W/2$ coarse positions, i.e., each coarse position corresponds to four positions at the finer $H \times W$ scale. Keys and values are drawn from a $P \times P$ window on the coarse grid. For computational simplicity, we fix this $P \times P$ attention window across scales, meaning that its physical footprint shrinks across scales. We find this performs similarly to the stricter case where the number of positions in the attention footprint increases to match the increasing resolution across scales but is substantially faster and less memory-intensive.

We train for 1M steps on L^1 loss using Adam, starting with a learning rate of 10^{-4} which anneals to 10^{-8} during training.

F Full ablations

In the main text, we presented an important ablation comparing S3C with a frequency weighting loss and demonstrated that it is unable to quantitatively improve predictions in the Navier-Stokes temporal rollout, and it is unable to match the correct spectral tail in the Kuramoto-Sivashinsky equation for modes unseen during training.



(a) Frequency loss ablation (b) S3C ablations

Figure 14: S3C ablations evaluated on power spectra over a one-step Navier-Stokes forecasting at 64×64 with a LocalNO backbone.

In this section, we present our full ablation experiments. In particular, we consider the challenging temporal rollout problem for the 64×64 Navier-Stokes setting. We compare performing S3C on the output space instead of the latent space and we compare using a translational (MLP predicts mode amplitudes for a fixed number of modes at each step) versus a dilational (MLP predicts mode amplitudes for a dyadically growing number of modes at each step). We also explore embedding S3C modules at the end of each layer instead of at the end of the full model.

From Figure 14, we observe that the frequency loss performs similarly to the base LocalNO in terms of the one-step Navier-Stokes power spectrum. We also observe that the output space S3C performs even worse than the base LocalNO, whereas latent-space dilational versus translational models perform similarly. We choose to adopt the dilational approach for the Navier-Stokes case because it aligns in a more principled way with power law decay, whereas a translational mode increment better aligns with exponential decay, although we empirically determined in this Navier-Stokes setting that the underlying MLP can learn the distinction appropriately.

Table 6 shows full quantitative results on the Navier-Stokes temporal rollout problem. We observe that per-layer-S3C performs substantially worse than last-layer S3C. We hypothesize that the nonlinearities between each layer alters the high-frequency tail and prevents the modules from learning the appropriate spectral decay. Other possibilities are that per-layer S3C modules disrupt the optimization landscape during training. While dilational mode spacing outperforms translational mode spacing across almost all steps, the numerical difference is relatively small, indicating approximately equal performance.

Regarding the choice of placing S3C in output space or latent space, while it is true that the nonlinear pointwise projections alters the power spectrum, particularly at high frequencies, we also find substantial expressivity improvements in the richer latent space when S3C operates on multiple channels. We also find that on the Navier-Stokes temporal extrapolation problem, the additional expressivity provided by the latent space construction helped significantly at the seen modes. However, for the PDE smoothing problem (e.g., the KS problem in Figure 4), we sometimes wish to extrapolate beyond the training mode cutoff. In this regime, we find that latent-space S3C is particularly sensitive to the nonlinearities in the pointwise projection, and we resort to output-space S3C, which can help significantly (e.g., see KS problem).

Table 6: S3C ablations on turbulent Navier-Stokes at 64×64. All errors are reported $\times 10^{-2}$. (a) Applying S3C after every FNO layer (per-layer) versus once after the last layer (last-layer). (b) Applying the correction directly in output space versus in latent space before the pointwise projection, with either translational or dilational spacing between the mode bands.

(a) Per-layer vs. last-layer S3C (FNO backbone)

	rel. L^2			rel. H^1			rel. HF_{32}		
Steps	Base	Per-layer	Last-layer	Base	Per-layer	Last-layer	Base	Per-layer	Last-layer
1	10.0	10.4	**9.8**	17.3	17.2	**17.1**	58.8	63.5	**56.5**
2	12.7	13.1	**12.5**	22.6	22.9	22.6	65.2	68.6	**63.6**
4	16.3	17.1	**16.1**	29.8	31.1	**29.7**	71.2	73.8	**70.3**
6	19.5	20.7	**19.3**	35.9	37.9	**35.6**	76.2	78.5	**75.9**
8	23.0	24.5	**22.7**	42.2	44.7	**41.7**	**81.9**	84.0	82.0
10	26.7	28.6	**26.5**	48.7	51.8	**48.2**	**88.0**	90.0	88.5

(b) Correction space and mode spacing (LocalNO backbone)

	rel. L^2			rel. H^1			rel. HF_{32}		
Steps	Output	Transl.	Dilat.	Output	Transl.	Dilat.	Output	Transl.	Dilat.
1	6.7	4.8	4.8	11.1	7.5	**7.3**	43.0	33.4	**33.2**
2	8.9	6.7	**6.6**	15.6	11.1	**10.9**	50.5	40.8	**40.6**
4	13.1	9.5	**9.4**	23.9	16.8	**16.6**	63.5	48.2	**47.8**
6	17.0	11.9	**11.8**	31.3	21.5	**21.4**	76.9	53.6	**52.8**
8	20.3	14.2	14.2	37.1	25.9	**25.8**	85.0	58.8	**57.8**
10	23.4	16.7	**16.6**	42.5	30.2	**30.1**	90.6	64.0	**63.0**