Frequency-Aware Vision Transformers for High-Fidelity Super-Resolution of Earth System Models

Ehsan Zeraatkar*

Computer Science, Texas State University, San Marcos, Texas, US

Salah A Faroughi

Chemical Engineering, The University of Utah, Salt Lake City, UT, US

Jelena Tesic

Computer Science, Texas State University, San Marcos, Texas, US

Abstract

Super-resolution (SR) is crucial for enhancing the spatial fidelity of Earth System Model (ESM) outputs, allowing fine-scale structures vital to climate science to be recovered from coarse simulations. However, traditional deep *super-resolution* methods, including convolutional and transformer-based models, tend to exhibit spectral bias, reconstructing low-frequency content more readily than valuable high-frequency details. In this work, we introduce two frequency-aware frameworks: the Vision Transformer-Tuned Sinusoidal Implicit Representation (ViSIR), combining Vision Transformers and sinusoidal activations to mitigate spectral bias, and the Vision Transformer Fourier Representation Network (ViFOR), which integrates explicit Fourier-based filtering for independent low- and high-frequency learning. Evaluated on the E3SM-HR Earth system dataset across surface temperature, shortwave, and longwave fluxes, these models outperform leading CNN, GAN, and vanilla transformer baselines, with ViFOR demonstrating up to 2.6 dB improvements in PSNR and significantly higher SSIM. Detailed ablation and scaling studies highlight the benefit of full-field training, the impact of frequency hyperparameters, and the potential for generalization. The results establish ViFOR as a state-of-the-art, scalable solution for climate data downscaling. Future extensions will address temporal super-resolution, multimodal climate variables, automated parameter selection, and integration of physical conservation constraints to broaden scientific applicability.

Keywords: Earth System Models, Super-Resolution, Vision Transformers, Implicit Neural Representations, Spectral Bias, Climate Data

Email address: ehsanzeraatkar@txstate.edu (Ehsan Zeraatkar*)

1. Introduction

The accelerating impacts of climate change demand ever more precise, fine-grained predictions of atmospheric and Earth system phenomena. Earth System Models (ESMs) are the principal scientific tool for simulating the interactions of the atmosphere, oceans, land, ice, and biosphere, and they underpin our understanding of global and regional climate variability [1]. However, the computational demands of running these models at truly high spatial resolution remain prohibitive. As a result, even the most advanced global simulations produced by state-of-the-art ESMs are typically limited to coarse spatial grids, yielding outputs that lack the local detail essential for regional-scale climate assessment and decision making [2].

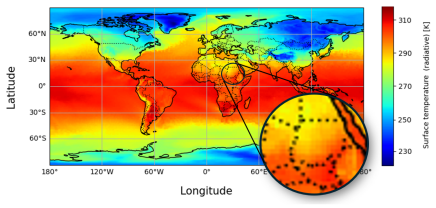

Figure 1: Global scale HR image yields a very LR output when limited to a country scale.

Figure 1 illustrates how detailed-appearing data at the planetary scale can degrade into low-resolution information when considered in terms of small areas. This crucial gap motivates the urgent development of super-resolution methods capable of transforming coarse ESM outputs into high-fidelity, high-resolution fields [3]. The *super-resolution* problem—where high-resolution (HR) images are reconstructed from low-resolution (LR) inputs—has become a core challenge in computer vision [4], with broad impact on medical imaging [5], wildlife tracking [6], security and surveillance [7], cultural preservation, and consumer electronics [8]. *super-resolution* methods must overcome the fundamental uncertainty of inferring missing details from coarse inputs, balancing computational tractability with accurate restoration of fine-scale structures [9, 10]. For ESM data, this difficulty is amplified: fine gradients and localized variations in climate variables make naive downscaling insufficient for regional applications [11]. Thus, advanced specialty *super-resolution* algorithms are required.

While deep learning approaches—especially convolutional neural networks (CNNs) and, more recently, vision transformers (ViTs)—have made impressive strides in *super-resolution* for natural imagery, they struggle when applied to Earth system data. A particularly persistent challenge is the problem of *spectral bias*: deep networks tend to favor smooth, low-frequency structures while failing to accurately reconstruct the fine-scale, high-frequency components that contain crucial physical information for climate science.

Recent breakthroughs in deep learning have delivered strong *super-resolution* models, including CNNs [12], ViTs [13], and generative diffusion techniques [14]. Vision Transformers are particularly promising due to their ability to model spatial long-range dependencies. Still, they remain plagued by difficulties in recovering sharp high-frequency details, as a result of *spectral bias* [15]. To address this, we first developed the *Vision Transformer–Tuned Sinusoidal Implicit*

Representation Network (ViSIR). ViSIR integrates sinusoidal activations in an implicit neural representation (INR) framework guided by ViTs, improving high-frequency detail recovery and partially mitigating spectral bias. While ViSIR demonstrates strong restoration of local details, it faces challenges in balancing high- and low-frequency content across diverse climate variables.

Responding to this problem, we introduce a frequency-aware framework for climate super-resolution, leveraging the global modeling strengths of ViTs while explicitly suppressing spectral bias. Our approach unfolds in two stages: first, integrating sinusoidal activations with transformer architectures in ViSIR; and second, expanding to explicit Fourier-based frequency separation in the Vision Transformer Fourier Representation Network (ViFOR). These novel architectures substantially enhance the reconstruction of high-frequency, physically meaningful gradients across key Earth system variables.

This work is motivated by the fact that bridging the resolution gap in ESM outputs will unleash new levels of fidelity in regional climate projections, empowering researchers and policymakers to make smarter, locally informed decisions. By directly confronting the core limitation of spectral bias in deep *super-resolution* architectures with targeted spectral innovations, we advance the state of the art in climate downscaling and provide a robust foundation for more accurate, reliable, and interpretable modeling of Earth's most complex systems.

Contributions: The paper introduces **ViSIR**, a hybrid Vision Transformer, an updated SIREN architecture that leverages sinusoidal activations to mitigate spectral bias in ESM super-resolution. We further present **ViFOR**, which extends ViSIR by incorporating Fourier-based activations, enabling explicit decomposition of low- and high-frequency components and achieving state-of-the-art results. Our evaluation on the E3SM-HR dataset demonstrates consistent and robust improvements in PSNR, SSIM, and MSE compared to CNN-, ViT-, and GAN-based baselines. Comprehensive ablation studies quantify the effect of sinusoidal frequencies in ViSIR and Fourier cutoffs in ViFOR, clarifying their strengths and limitations. Notably, ViFOR benefits markedly from full-image training, revealing the critical role of global context and establishing a scalable framework for ESM downscaling. The ViSIR and ViFOR are introduced as successive solutions to spectral bias in ESM SR. ViSIR establishes a frequency-sensitive SR context using sinusoidal INRs, while ViFOR advances performance by enabling explicit frequency component separation.

2. Related Work

Super-resolution has been a computer vision challenge over the years, extending from remote sensing and medical imaging to environmental and climate sciences. *super-resolution* is either single-image SR (SISR), in which one low-resolution (LR) input is restored into one high-resolution (HR) image, or multi-image SR (MISR), in which complementary information across a sequence of LR observations is used to estimate a better HR reconstruction [8]. While classical methods such as reconstruction-based and example-based

schemes [16, 17, 18, 19] yielded early improvement, they were limited by reliance on strong assumptions about image structure and smoothness, which made them unsuitable for general real-world cases.

Deep learning and, particularly, Convolutional Neural Networks (CNNs) revolutionized SR. The seminal SRCNN [20] brought end-to-end mapping from LR to HR images, beating interpolation-based baselines. Subsequent models such as VDSR [12], EDSR [21], and RDN [22] dove deeper into network architectures and employed residual and dense connections for improved reconstruction of fine-grained details. However, CNNs inherently possess a *spectral bias* towards low-frequency smooth features and undersampling of high-frequency ones like edges and textures [23]. This bias is particularly detrimental to Earth System Model (ESM) applications where high-contrast temperature and flux gradients in spatially localized areas harbor critical scientific information.

To address these problems, implicit neural representations (INRs), directly mapping continuous spatial coordinates to signal values, have been explored. The Sinusoidal Representation Network (SIREN) [24] demonstrated periodic activation functions that drastically minimize spectral bias so that high-frequency content could be recovered more effectively. Subsequent works such as Generalized INRs (GINR) [25] and Higher-Order INRs (HOIN) [26] extend these ideas by incorporating spectral graph embeddings and neural tangent kernels. In parallel, generative adversarial techniques such as SRGAN [27] aim at perceptual realism and have been applied to downscale coarse-resolution climate data. In contrast, multimodal architectures combining U-Nets with attention modules have been used to enhance regional temperature forecasts [28]. NeurOp-Diff [29] introduced a neural operator–based diffusion framework for continuous-scale super-resolution, demonstrating strong generalization to arbitrary scale factors with publicly available code. Similarly, PC-SRGAN [30] proposed a physically consistent SRGAN for transient scientific simulations, enforcing conservation properties while enhancing reconstruction quality.

Vision Transformers (ViTs) [13] have recently appeared as serious alternatives to the conventional CNNs in *super-resolution* tasks. By partitioning images into patches and applying multi-head self-attention, ViTs excel at dealing with long-range spatial relations and world context. This makes them very attractive for ESM datasets, where global-scale behavior must be preserved. However, ViTs also suffer from spectral bias, which means they have been found to collapse fine details and high-frequency features for all tasks [15]. Efforts to simplify transformer-based *super-resolution* models [31, 32] have improved efficiency but were unable to remove the innate problem of frequency imbalance. Most recently, TTRD3 [33] combined texture transfer and dual diffusion modeling to recover fine-grained remote sensing textures, underscoring the trend of frequency- and texture-aware *super-resolution* approaches.

These developments directly motivate our contributions. First, we introduced the *Vision Transformer–Tuned Sinusoidal Implicit Representation Network (ViSIR)*, which integrates ViTs and sinusoidal INRs to mitigate spectral bias and recover fine details in ESM data. While effective, ViSIR still falls short of balancing high- and low-frequency learning for heterogeneous climate

variables. From this, we propose the *Vision Transformer Fourier Representation Network (ViFOR)*, which expands the approach by adding Fourier-based activation functions. The design possesses an explicit high-pass and low-pass decoupling, more effectively alleviating spectral bias and achieving state-of-the-art results in ESM super-resolution.

3. Methodology

3.1. Problem Formulation

The super-resolution problem is formulated as learning a mapping function $f_\theta : I_{LR} \rightarrow I_{SR}$, where I_{LR} is a low-resolution (LR) image and I_{SR} is the reconstructed high-resolution (HR) output. For Earth System Model (ESM) data, LR inputs are generated by interpolating fine-resolution outputs from the E3SM-HR dataset onto a coarse $1^\circ \times 1^\circ$ grid, while HR references are the originals high-resolution ($0.25^\circ \times 0.25^\circ$) grid. Each image has normalized variables such as surface temperature, shortwave flux, and longwave flux for a 720×1440 grid. The *super-resolution* task is then imperative for the recovery of fine-scale gradients and localized variations from coarse-resolution ESM outputs. The quality of reconstruction is quantified in terms of Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM).

3.2. ViSIR: Vision Transformer–Tuned Sinusoidal Implicit Representation

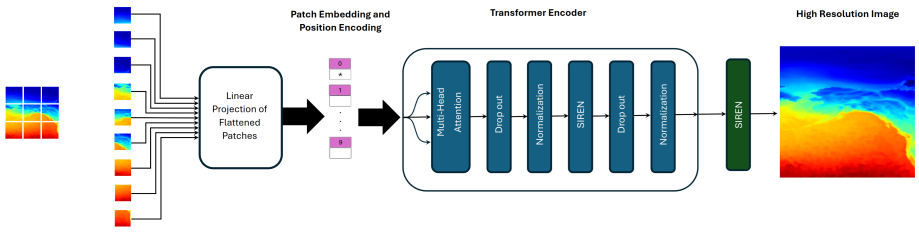

Figure 2: ViSIR divides the input image into patches, pre-processes them using embedding and position encoding, and feeds the input to a visual transformer followed by the SIREN architecture.

The Vision Transformer–Tuned Sinusoidal Implicit Representation (ViSIR) was introduced as a frequency-aware *super-resolution* method to mitigate *spectral bias*—the tendency of neural networks to overfit low-frequency components while neglecting fine details. ViSIR integrates a Vision Transformer (ViT) with Sinusoidal Representation Networks (SIRENs), combining global context modeling with high-frequency recovery.

ViSIR pipeline flow is illustrated in Figure 2. The ViT processes the input images in the proposed pipeline to capture global dependencies and essential contextual information for super-resolution. Given a low-resolution image I_{LR}, the patch embedding operation can be described as:

$$E_i = W_p \cdot \text{patch}(I_{\text{LR}})_i + b_p, \quad i = 1, \ldots, N, \tag{1}$$

where W_p and b_p are the patch embedding weights and bias, and N is the number of patches. These embeddings are then processed through multiple transformer layers, yielding contextualized features:

$$T = \text{ViT}(I_{\text{LR}}). \tag{2}$$

The SIREN then uses this information to address the spectral bias in the input image, resulting in higher-resolution output, as outlined in Eq. 3

$$f(x) = \sin\left(\omega_0 W x + b\right), \tag{3}$$

The ω_0 is a hyperparameter controlling the frequency of the sinusoidal function. In ViSIR, the final reconstruction is performed by feeding the transformer's output into one or more SIREN layers, as outlined in the Eq 4, where W_r and b_r are the weights and bias of the SIREN module.

$$R = f(T) = \sin\left(\omega_0 W_r T + b_r\right), \tag{4}$$

Next, ViSIR integrates ViT's global context modeling capabilities with SIREN's high-frequency representation strength. Algorithm 1 presents a concise pseudo-code overview of the proposed ViSIR methodology, detailing the steps from patch extraction and token embedding through transformer-based processing to the final reconstruction using a SIREN layer.

Algorithm 1 ViSIR Super-Resolution

Input: Low-resolution image LR, patch size P, number of layers L, number of heads H, embedding dimension D, frequency ω_0
Output: Reconstructed high-resolution image HR
Divide LR into non-overlapping patches of size $P \times P$.
for each patch **do**
    Compute a linear embedding to obtain token $x \in \mathbb{R}^D$.
Form token sequence $X = \{x_1, x_2, \ldots, x_N\}$ and add positional encodings.
for $\ell = 1$ to L **do**
    Apply multi-head self-attention on X using H heads.
    : Add a residual connection and perform layer normalization.
    Process the result through a SIREN network.
Aggregate token features (e.g., via average pooling) to obtain feature vector F.
Pass F through a SIREN layer:

$$HR = \sin\Big(\omega_0 \cdot (WF + b)\Big)$$

return HR.

While ViSIR effectively improves the reconstruction of high-frequency patterns, it remains limited by its uniform sinusoidal activations, which do not explicitly distinguish between low- and high-frequency components. This motivates the design of the improved ViFOR architecture.

3.3. ViFOR: Vision Transformer Fourier Representation Network

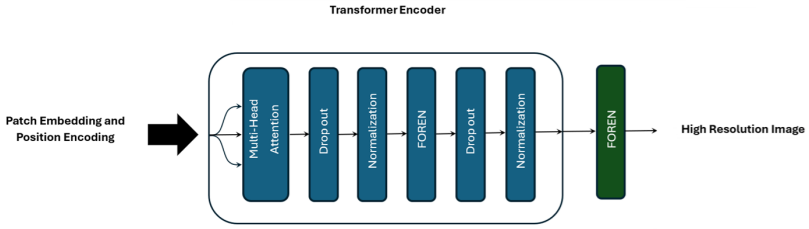

Figure 3: ViFOR pipeline: SIREN is replaced by the Fourier-based activation function Network in the transformer and output sections

The Vision Transformer Fourier Representation Network (ViFOR) generalizes ViSIR by involving explicit Fourier-based activation functions that allow for independent learning of the low-frequency and high-frequency components, as shown in Figure 3. The default MLP layers in the ViT are replaced with a Fourier Representation Network (FOREN), which uses low-pass and high-pass filters as nonlinear activations. This architecture tackles spectral bias more explicitly by separating frequency learning.

According to Figure 3, illustrating the ViFOR pipeline, first ViFOR segments the input image into patches $\mathbf{P}_i \in \mathbb{R}^{p\times p\times C}$. Each patch is flattened and embedded:

$$\mathbf{z}_i = \mathbf{W}_e \cdot \mathrm{vec}(\mathbf{P}_i) + \mathbf{E}_i$$

where $\mathbf{W}_e \in \mathbb{R}^{d\times p^2 C}$ is a learnable projection matrix and $\mathbf{E}_i$ is the positional encoding.

The sequence of embeddings $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ is passed through the transformer encode, which utilizes an FFT-based activation function at the MPL layers:

$$\mathbf{Z}' = \mathrm{TransformerEncoder}(\mathbf{Z})$$

$$\mathrm{FOREN}(x) = \mathcal{F}^{-1}(\mathcal{F}(x) \cdot H_f)$$

where H_f can be:

- Low-pass filter: attenuates frequencies above cutoff f_l
- High-pass filter: attenuates frequencies below cutoff f_h

We define two parallel FOREN branches:

$$\hat{\mathbf{X}}_{\text{low}} = \text{FOREN}_{f_l}(\mathbf{Z}')$$
$$\hat{\mathbf{X}}_{\text{high}} = \text{FOREN}_{f_h}(\mathbf{Z}')$$

The final output is a weighted fusion:

$$\hat{\mathbf{X}}_{HR} = \alpha \cdot \hat{\mathbf{X}}_{\text{low}} + (1 - \alpha) \cdot \hat{\mathbf{X}}_{\text{high}}, \quad \alpha \in [0, 1]$$

The model trains by minimizing a combination of MSE and frequency-domain losses:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \text{MSE}(\hat{\mathbf{X}}_{HR}, \mathbf{X}_{HR}) + \lambda_2 \cdot \text{FreqLoss}(\hat{\mathbf{X}}_{HR}, \mathbf{X}_{HR})$$

where λ_1, λ_2 are weights for loss terms.

Algorithm 2 ViFOR: Super-Resolution via Vision Transformer and Fourier Filters

Require: Low-resolution image $\mathbf{X}_{LR}$, patch size p, frequencies f_l, f_h, fusion weight α

Split $\mathbf{X}_{LR}$ into patches $\{\mathbf{P}_i\}$
for each patch $\mathbf{P}_i$ **do**
    Compute embedding: $\mathbf{z}_i \leftarrow \text{Embed}(\mathbf{P}_i) + \text{PosEnc}(\mathbf{P}_i)$
Form sequence $\mathbf{Z} \leftarrow [\mathbf{z}_1, \ldots, \mathbf{z}_n]$
$\mathbf{Z}' \leftarrow \text{TransformerEncoder}(\mathbf{Z})$
$\hat{\mathbf{X}}_{\text{low}} \leftarrow \text{IFFT}(\text{FFT}(\mathbf{Z}') \cdot H_{f_l})$
$\hat{\mathbf{X}}_{\text{high}} \leftarrow \text{IFFT}(\text{FFT}(\mathbf{Z}') \cdot H_{f_h})$
$\hat{\mathbf{X}}_{HR} \leftarrow \alpha \cdot \hat{\mathbf{X}}_{\text{low}} + (1 - \alpha) \cdot \hat{\mathbf{X}}_{\text{high}}$
Compute loss $\mathcal{L}_{\text{total}}$
Update model parameters via backpropagation

Ensure: Output: $\hat{\mathbf{X}}_{HR}$

This methodology provides a systematic approach to high-fidelity super-resolution that accounts for global structure and local texture in Earth System Model data.

3.4. Dataset and Preprocessing

We trained our models on the *Energy Exascale Earth System Model High-Resolution (E3SM-HR)* dataset. The dataset corresponds to a 30-year control simulation, with outputs bilinearly interpolated from the native cubed-sphere grid onto a standard longitude–latitude grid of $0.25^\circ \times 0.25^\circ$. We focused on three significant climate variables shown in Figure 4, Surface Temperature (TS), Shortwave Flux (FSW), and Longwave Flux (FLW). All of the variables were mapped and normalized to RGB channels to create multi-variable image representations. For scalability of the model, we employed two settings: (1) *sub-image training*, where an image was divided into eight non-overlapping patches, and (2) *full-image training*, where global fields were utilized in their entirety.

4. Experimental Setup

4.1. Baseline Models

The ViSIR and ViFOR methods are compared to FIVE established SR methods:

1. **SRCNN** is an early CNN-based *super-resolution* model with end-to-end learning [20].
2. **PC–SRGAN** is a physically consistent SRGAN variants that augment the SRGAN objective with physics-based constraints (e.g., gradient/Laplacian consistency and conservation surrogates) to enforce scientific plausibility in transient simulation [30].
3. **SRGAN** is a super-resolution GAN-based model emphasizing perceptual realism [9].
4. **ViT** Vision Transformer with patch embedding and self-attention [13].
5. **SIREN** uses an implicit representation with sinusoidal activations to mitigate spectral bias. [24].

This comparison ensures that both convolution- and transformer-based paradigms are represented, allowing a fair evaluation of ViSIR and ViFOR.

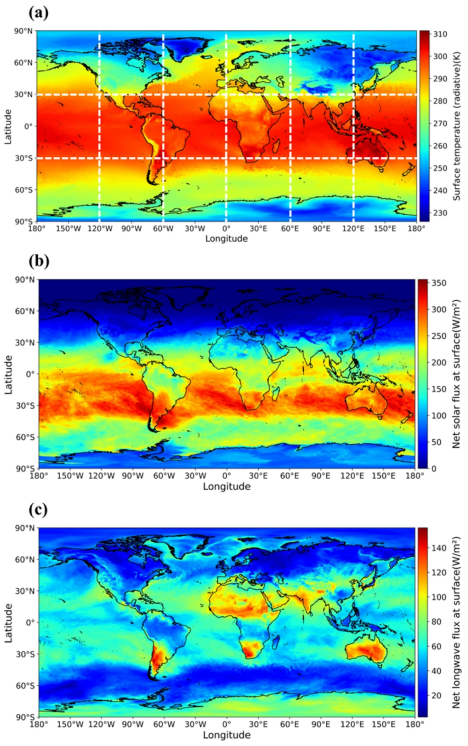

Figure 4: Panels (a), (b), and (c) show surface temperature, shortwave heat flux, and longwave heat flux, respectively, for the first month of year one obtained from the global fine-resolution configuration of E3SM.

4.2. Evaluation Metrics

To evaluate the algorithm's performance, we employ three complementary metrics. Mean Squared Error (MSE) to quantify the average pixel-wise difference between the reconstructed HR image and the ground truth, Peak Signal-to-Noise Ratio (PSNR) to measure reconstruction quality in decibels, where higher values indicate less distortion, and Structural Similarity Index (SSIM) to assess perceived image similarity, emphasizing structural fidelity[34].

These metrics capture both numerical accuracy (MSE, PSNR) and perceptual quality (SSIM), which are critical for assessing ESM data reconstructions.

4.3. Training Environment and Hyperparameters

Both ViSIR and ViFOR were trained on the E3SM-HR dataset, which includes 360 monthly images spanning ten years across three variables: surface

temperature, shortwave flux, and longwave flux. Models were optimized using the Adam optimizer with an initial learning rate of 10^{-4} and a cosine decay schedule. Mean squared error was used as the primary loss function, with additional perceptual loss experiments conducted to evaluate qualitative improvements.

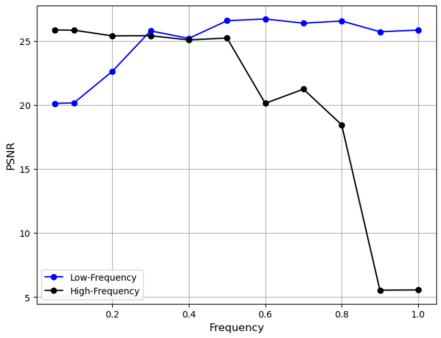

Figure 5: PSNR across different cutoff frequencies f_c for ViFOR. Optimal performance was achieved at $f_c = 0.3$ Hz.

For ViFOR, cutoff frequencies between 0.01 Hz and 1 Hz were investigated, with 0.3 Hz identified as a fixed optimal frequency for both low-pass and high-pass filters. This configuration ensured stable recovery of both global and fine-scale features without introducing artifacts. For ViSIR, we investigated the PSNR values with different sinusoidal activation function frequencies ranging from 10 to 60 Hz while varying the number of hidden layers in the SIREN from 1 to 6. Figure 6 illustrates the effect of different parameter combinations. Based on the results shown in Figure 6, the best PSNR value for EMS images is chosen to be two hidden layers with a frequency of 20. These are the hyperparameters used for the rest of the methods in frequency and layers to perform a fair comparison.

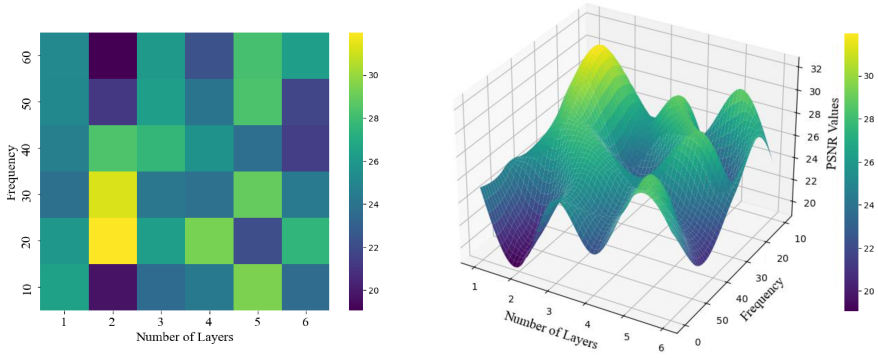

Figure 6: 2D (left) and 3D (right) illustration of the PSNR values for different Frequencies and different numbers of hidden layers used in the proposed ViSIR applied to the E3SM dataset

5. Experimental Results

5.1. Quantitative Comparison

We evaluated ViSIR and ViFOR against state-of-the-art baselines, including CNN-based methods (SRCNN), the Vision Transformer (ViT), Sinusoidal Representation Networks (SIREN), PC- and SRGAN. Table 1 summarizes the performance in terms of PSNR and MSE across three ESM variables: surface temperature (TS), shortwave flux (FSW), and longwave flux (FLW). The quantitative results in Table 1 highlight the exceptional reconstruction capacity of the novel ViSIR and ViFOR frameworks compared to baselines. Conventional CNN-based approaches, such as SRCNN and SRGAN, the transformer-based ViT, and the implicit SIREN models, show peak signal-to-noise ratio (PSNR) results generally below 24–27 dB and structural similarity index (SSIM) results around 0.6–0.7. These findings demonstrate finite capability in reconstructing high-frequency spatial details that are important for Earth System Model (ESM) variables. The physics-constrained PC-SRGAN has better reconstruction quality with PSNRs as high as 28.9 dB and SSIMs of 0.78 for longwave flux via the addition of physical consistency losses that encourage gradient preservation. Nevertheless, ViSIR features a noticeable performance improvement over all the baseline models by mitigating spectral bias through sinusoidal activation functions with PSNR improvements of around 5–6 dB over ViT and 1.0–1.5 dB over PC-SRGAN, with a more than 50% reduction of mean squared error (MSE) of all the variables considered. The improved ViFOR model extends these strengths by incorporating Fourier-based activation filtering and transformer attention to enable balanced learning of low- and high-frequency components. ViFOR achieves 29.3 dB, 29.5 dB, and 30.1 dB for Source Temperature, Shortwave Flux, and Longwave Flux, respectively, and SSIM up to 0.80 and MSE as low as 0.1%. ViFOR consistently outperforms ViSIR and all other models, with greater fidelity and structural integrity of reconstructed ESM fields with physically relevant spatial gradients and variability intact.

5.2. Ablation Studies

Effect of Sinusoidal Frequency Parameter in ViSIR.. We examined the influence of the SIREN frequency parameter ω_0. Larger values of ω_0 improved recovery of high-frequency details but at the cost of training instability. An intermediate selection ($\omega_0 = 30$) proved optimal, similar to previous work [24].

Effect of Fourier Cutoff Frequency in ViFOR.. For ViFOR, cutoff frequencies f_c between 0.01 and 1.0 Hz were tested. Figure 5 shows that $f_c = 0.3$ Hz offered maximum PSNR on both the low-pass and high-pass branches. Cutoffs higher than 0.3 Hz degraded performance through amplification of noise, while lower cutoffs limited recovery of high-frequency information.

Table 1: The values of MSE (%), PSNR (dB), and SSIM [0-1] for original I_O and reconstructed I_R images for three measurements, Source Temperature, shortwave heat flux, and longwave heat flux, and FOUR different models.

	Sub-Image			**Full Image**		
Models ↓ / Measures →	**MSE %**	**PSNR dB**	**SSIM [0,1]**	**MSE %**	**PSNR dB**	**SSIM [0,1]**
	Source Temperature					
ViT	0.42	22.43	0.62	0.49	23.27	0.64
SIREN	1.18	19.32	0.52	0.93	20.21	0.57
SRGAN	0.42	23.72	0.66	0.61	21.43	0.60
SRCNN	0.14	27.51	0.74	0.14	27.42	0.74
PC-SRGAN	0.19	27.04	0.73	0.17	27.28	0.74
ViSIR	0.13	28.60	0.76	0.13	28.50	0.76
ViFOR	0.13	28.62	0.76	**0.11**	**29.32**	**0.78**
	Shortwave heat flux					
ViT	0.47	23.12	0.64	0.40	23.92	0.66
SIREN	0.94	20.12	0.56	0.91	20.57	0.58
SRGAN	0.41	22.84	0.65	0.67	21.22	0.61
SRCNN	0.17	27.12	0.73	0.20	26.89	0.72
PC-SRGAN	0.14	28.36	0.75	0.17	28.12	0.74
ViSIR	0.14	28.32	0.75	0.15	28.01	0.74
ViFOR	0.13	28.67	0.77	**0.10**	**29.51**	**0.79**
	Longwave heat flux					
ViT	0.93	20.12	0.57	0.91	20.55	0.58
SIREN	0.90	20.69	0.59	0.67	21.20	0.60
SRGAN	0.64	21.32	0.60	0.91	20.57	0.58
SRCNN	0.21	26.42	0.71	0.22	26.21	0.71
PC-SRGAN	0.11	28.88	0.78	0.13	28.56	0.76
ViSIR	0.15	27.64	0.73	0.08	30.50	0.81
ViFOR	0.14	28.42	0.76	**0.09**	**30.13**	**0.80**

5.3. Comparison of Sub-image vs. Full-image Training

To assess the influence of spatial context on model performance, additional experiments were conducted by training the networks on $90^\circ \times 180^\circ$ regional sub-images and on full 720×1440 global fields. The results indicate that both ViSIR and ViFOR benefit from exposure to the full-image domain, with ViFOR exhibiting a more pronounced improvement. In particular, ViFOR achieved 29.3 dB PSNR and 0.78 SSIM on full global images compared to 28.6 dB and 0.76 SSIM on sub-image training for the Source Temperature variable. This substantial gain underscores the importance of retaining large-scale spatial dependencies inherent in Earth System Model (ESM) data. Owing to its transformer-based backbone and Fourier-modulated representations, ViFOR effectively exploits global contextual information, enabling the model to reconstruct high-frequency patterns that are spatially coherent with large-scale climate structures. In contrast, ViSIR exhibits moderate but consistent improvement under full-image training, reflecting its enhanced ability to model local variations through sinusoidal activations, albeit with less global context awareness than ViFOR.

6. Discussion

The progression from ViSIR to ViFOR demonstrates a systematic enhancement in frequency-aware super-resolution for Earth System Model (ESM) data. ViSIR combines sinusoidal implicit representations with Transformer self-attention,

effectively recovering high-frequency spatial details commonly lost in downscaling and significantly reducing spectral bias compared to CNN and vanilla ViT models. Quantitative results show ViSIR consistently achieves PSNR values up to 28.6 dB and SSIM scores as high as 0.76, representing a notable improvement over prior baselines.

ViFOR further advances this approach by explicitly separating low- and high-frequency components using Fourier-based activations. This architecture enables more balanced reconstructions of both large-scale and local climate patterns, as reflected in PSNR rates exceeding 29.0 dB and SSIM scores approaching 0.80 on full-image training sets. The model also consistently lowers MSE across all tested variables. Ablation studies confirm that optimal choice of sinusoidal and Fourier cutoff frequencies is essential for stable and precise reconstruction; full-image training provides observable gains, emphasizing the importance of global spatial context in climate modeling.

Despite these successes, several limitations are evident. Computational cost increases with full-field transformer models, but scalability has been empirically validated on conventional CPU infrastructure. The cutoff frequency (f_c) remains a sensitive hyperparameter, and current experiments address only spatial downscaling; temporal SR for intra-monthly or weekly fields remains unexplored in this study.

7. Conclusion and Future Work

This work establishes ViSIR and ViFOR as state-of-the-art frequency-aware architectures for super-resolution of ESM data. Both models are shown to recover fine-scale spatial variability lost by conventional deep learning approaches, with ViFOR providing the best overall accuracy, perceptual coherence, and scalability among all tested models.

Future research will focus on several fronts: (1) exploring model efficiency strategies to reduce computational overhead, such as model compression or transformer simplification; (2) developing systematic methods for automated hyperparameter tuning of frequency parameters; (3) expanding the models to spatio-temporal SR by addressing temporal downscaling and benchmarking across additional climate variables and more diverse domains; (4) investigating integration of physically-consistent loss functions to enhance scientific reliability; and (5) applying ViFOR to larger datasets to further assess scalability and practical deployment. These steps are anticipated to improve further the fidelity and applicability of frequency-aware SR in climate science, addressing current limitations and broadening impact.

8. Ethics approval and consent to participate

Ethics approval and consent to participate were not required for this study, as it used only publicly available datasets and did not involve interaction with human or animal subjects.

9. Competing interest

The authors declare that they have no competing interests.

10. Availability of data and material

The dataset used in this study is publicly available at [35].

11. Funding

S.A.F. and J. T. would like to acknowledge support by the Department of Energy's Biological and Environmental Research (BER) program (award no. DE-SC0023044)

References

[1] W. D. Collins, C. M. Bitz, M. L. Blackmon, G. B. Bonan, C. S. Bretherton, J. A. Carton, P. Chang, S. C. Doney, J. J. Hack, T. B. Henderson, J. T. Kiehl, W. G. Large, D. S. McKenna, B. D. Santer, R. D. Smith, The community climate system model version 3 (ccsm3), Journal of Climate 19 (2006) 2122–2143. doi:10.1175/JCLI3761.1.

[2] C. Heinze, V. Eyring, P. Friedlingstein, C. Jones, Y. Balkanski, W. Collins, T. Fichefet, S. Gao, A. Hall, D. Ivanova, et al., Esd reviews: Climate feedbacks in the earth system and prospects for their evaluation, Earth Syst. Dynam. 10 (2019) 379–452. URL: https://doi.org/10.5194/esd-10-379-2019. doi:10.5194/esd-10-379-2019.

[3] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, K. E. Taylor, Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization, Geoscientific Model Development 9 (2016) 1937–1958. URL: https://doi.org/10.5194/gmd-9-1937-2016. doi:10.5194/gmd-9-1937-2016.

[4] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of neural networks, in: International conference on machine learning, PMLR, 2019, pp. 5301–5310.

[5] A. Bashir, V. A. Larsen, M. Ziebell, K. Fugleholm, I. Law, Improved detection of postoperative residual meningioma with [68ga]ga-dota-toc pet imaging using a high-resolution research tomograph pet scanner, Clinical Cancer Research 27 (2021) 2216–2225. URL: https://doi.org/10.1158/1078-0432.CCR-20-3362. doi:10.1158/1078-0432.CCR-20-3362.

[6] I. Kuzmanić, I. Vujović, M. Vujović, Application of computer vision in security and emergency actions, in: Proceedings of the 14th Annual Conference of the International Emergency Management Society, TIEMS, Split, 2007, pp. 336–345–x.

[7] S. Grosche, F. Brand, A. Kaup, A novel end-to-end network for reconstruction of non-regularly sampled image data using locally fully connected layers, in: 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), 2021, pp. 1–6. doi:10.1109/MMSP53017.2021.9733541.

[8] B. C. Maral, Single image super-resolution methods: A survey, https://doi.org/10.48550/arXiv.2202.11763 (2022).

[9] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681–4690. URL: https://arxiv.org/abs/1609.04802.

[10] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017) 1132–1140. URL: https://api.semanticscholar.org/CorpusID:6540453.

[11] T. Vandal, E. Kodra, A. R. Ganguly, Deepsd: Generating high resolution climate change projections through single image super-resolution, arXiv preprint arXiv:1703.03126 (2017). URL: https://arxiv.org/abs/1703.03126, accessed: 2024-09-02.

[12] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646–1654. doi:10.1109/CVPR.2016.182.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[14] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 2256–2265.

[15] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, W. Liu, Improving vision transformers by revisiting high-frequency components, in: European Conference on Computer Vision, Springer, 2022, pp. 1–18.

[16] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Transactions on Image Processing 19 (2010) 2861–2873. doi:10.1109/TIP.2010.2050625.

[17] W. Freeman, T. Jones, E. Pasztor, Example-based super-resolution, IEEE Computer Graphics and Applications 22 (2002) 56–65. doi:10.1109/38.988747.

[18] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, Soft edge smoothness prior for alpha channel super resolution, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8. doi:10.1109/CVPR.2007.383028.

[19] H. Aly, E. Dubois, Image up-sampling using total-variation regularization with a new observation model, IEEE Transactions on Image Processing 14 (2005) 1647–1659. doi:10.1109/TIP.2005.851684.

[20] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2016) 295–307. doi:10.1109/TPAMI.2015.2439281.

[21] J. Kim, J. K. Lee, K. M. Lee, Enhanced deep residual networks for single image super-resolution, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 136–144.

[22] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481. doi:10.1109/CVPR.2018.00262.

[23] X. Zhang, Z. Zhang, S. Wu, Z. Zhang, Residual networks behave like ensembles of relatively shallow networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 1311–1325.

[24] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, G. Wetzstein, Implicit neural representations with periodic activation functions, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, volume 33 of *NIPS '20*, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 7462–7473.

[25] D. Grattarola, P. Vandergheynst, Generalised implicit neural representations, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, volume 35 of *NIPS '22*, Curran Associates Inc., Red Hook, NY, USA, 2022, pp. 30446–30458.

[26] Y. Chen, R. Wu, Y. Liu, C. Zhu, Hoin: High-order implicit neural representations, 2024. URL: `https://arxiv.org/abs/2404.14674`. `arXiv:2404.14674`.

[27] N. Shidqi, C. Jeong, S. Park, E. Zeller, A. B. Nellikkattil, K. Singh, Generating high-resolution regional precipitation using conditional diffusion model, 2023. URL: https://arxiv.org/abs/2312.07112. arXiv:2312.07112.

[28] S. Ding, X. Zhi, Y. Lyu, Y. Ji, W. Guo, Deep learning for daily 2-m temperature downscaling, Earth and Space Science 11 (2024). doi:10.1029/2023EA003227.

[29] Z. Zhang, W. Liu, H. Chen, M. Li, Y. Wang, Neuropdiff: Continuous remote sensing image super-resolution via neural operator diffusion, arXiv preprint arXiv:2501.09054 (2025). URL: https://arxiv.org/abs/2501.09054, code available at https://github.com/zerono000/NeurOp-Diff.

[30] A. Kumar, J. Li, R. Zhao, X. Wang, Pc-srgan: Physically consistent super-resolution generative adversarial network for general transient simulations, arXiv preprint arXiv:2505.06502 (2025). URL: https://arxiv.org/abs/2505.06502.

[31] X. Zhong, F. Du, L. Chen, Z. Wang, H. Li, Investigating transformer-based models for spatial downscaling and correcting biases of near-surface temperature and wind-speed forecasts, Quarterly Journal of the Royal Meteorological Society 150 (2023) 275–289. doi:10.1002/qj.4596.

[32] K. Karwowska, D. Wierzbicki, Modified esrgan with uformer for video satellite imagery super-resolution, Remote Sensing 16 (2024) 1926. doi:10.3390/rs16111926.

[33] S. Liu, K. Wang, Z. Hu, L. Chen, Ttrd3: Texture transfer residual denoising dual diffusion model for remote sensing image super-resolution, arXiv preprint arXiv:2504.13026 (2025). URL: https://arxiv.org/abs/2504.13026.

[34] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.

[35] E3SM Project, Energy Exascale Earth System Model (E3SM), [Computer Software], 2024. doi:10.11578/E3SM/dc.20240301.3.