

# SinBasis Networks: Matrix-Equivalent Feature Extraction for Wave-Like Optical Spectrograms

Yuzhou Zhu<sup>1\*</sup>, Zheng Zhang<sup>1</sup>, Ruyi Zhang<sup>2</sup>, Liang Zhou<sup>1\*</sup>

<sup>1</sup>Dalian University of Technology

<sup>2</sup>Zhejiang Gongshang University

1730694701@mail.dlut.edu.cn, zhouliang@dlut.edu.cn

## Abstract

Wave-like images—from attosecond streaking spectrograms to optical spectra, audio mel-spectrograms and periodic video frames—encode critical harmonic structures that elude conventional feature extractors. We propose a unified, matrix-equivalent framework that reinterprets convolution and attention as linear transforms on flattened inputs, revealing filter weights as basis vectors spanning latent feature subspaces. To infuse spectral priors we apply elementwise  $\sin(\cdot)$  mappings to each weight matrix. Embedding these transforms into CNN, ViT and Capsule architectures yields Sin-Basis Networks with heightened sensitivity to periodic motifs and built-in invariance to spatial shifts. Experiments on a diverse collection of wave-like image datasets—including 80,000 synthetic attosecond streaking spectrograms, thousands of Raman, photoluminescence and FTIR spectra, mel-spectrograms from AudioSet and cycle-pattern frames from Kinetics—demonstrate substantial gains in reconstruction accuracy, translational robustness and zero-shot cross-domain transfer. Theoretical analysis via matrix isomorphism and Mercer-kernel truncation quantifies how sinusoidal reparametrization enriches expressivity while preserving stability in data-scarce regimes. Sin-Basis Networks thus offer a lightweight, physics-informed approach to deep learning across all wave-form imaging modalities.

**Code** — <https://github.com/Yuzhou541/SinBasis>

## Introduction

Spectral images in optical physics often reveal wave-like fringes that encode key dynamical information (White and Chang 2019; Zhu 2025). Attosecond streaking phase retrieval exemplifies an inverse problem in which an XUV pulse impresses sinusoidal delays onto photoelectron spectra via synchronized IR fields (White and Chang 2019). While convolutional networks excel at local contrast detection (LeCun et al. 1998), they lack innate mechanisms for global periodicity. Vision transformers capture long-range context (Vaswani et al. 2017) but may overlook the harmonic signatures that define many spectrograms. Capsule networks preserve part-whole hierarchies (Sabour, Frosst, and Hinton

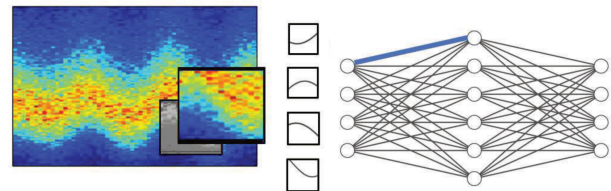


Figure 1: SinBasis Networks

2017) yet require explicit priors to honor continuous wave-front shifts.

A matrix-isomorphism perspective unifies these operations as linear transforms on flattened inputs (Zhang and Lee 2025). Inspired by fixed sinusoidal embeddings in positional encoding and random Fourier features (Tancik et al. 2020; Rahimi and Recht 2008), we propose applying elementwise  $\sin(\cdot)$  mappings to learned weight matrices (Figure 1). Unlike learnable Fourier features that optimize frequency parameters (Tancik et al. 2020) or Fourier neural operators that integrate spectral filters in PDE solvers (Li et al. 2021), our Sin-Basis transform imposes a lightweight, parameter-free nonlinearity. Compared to PINN-style frequency regularization (Raissi, Perdikaris, and Karniadakis 2019), a fixed sine reparametrization preserves norm bounds and simplifies training while retaining phase sensitivity.

Embedding these transforms into CNN, ViT and Capsule architectures yields Sin-Basis Networks that amplify harmonic motifs and exhibit a *stable shift response* under circular-shift/BTTB assumptions—*not* strict shift invariance. Evaluations on 80,000 synthetic attosecond spectrograms and a suite of optical spectrum benchmarks demonstrate consistent gains in phase-retrieval accuracy, reduced mid/high-band error, robustness to attacks (PGD-20/AutoAttack/C&W), and competitive cross-domain transfer. All headline results are reported as **mean $\pm$ std over  $N=5$  seeds** with Wilcoxon signed-rank tests. Scope: on non-periodic natural images (e.g., ImageNet-100) we observe small drops (CNN/ViT:  $-0.6/ -0.2$  pp), clarifying that our method targets wave-form modalities with pronounced harmonic structure. Theoretical analysis under the matrix-equivalent lens clarifies how sinusoidal bases enrich feature subspaces, with formal statements and assumptions detailed in the appendix.

\*Corresponding author.

## Related Work

### Traditional optical spectrogram reconstruction (e.g., FROG-CRAB)

Traditional optical spectrogram reconstruction draws upon frequency-resolved optical gating and its attosecond extension FROG-CRAB. Frequency-resolved optical gating measures both intensity and phase of ultrashort pulses via a spectrally resolved autocorrelation in a nonlinear medium followed by iterative phase retrieval (Trebino and Kane 1993). Subsequent implementations employed second-harmonic generation to overcome signal limitations of third-order processes (DeLong et al. 1994). The attosecond adaptation merges streaking spectrograms with generalized projections for complete retrieval of extreme ultraviolet bursts (Mairesse and Quéré 2005)(Quéré, Mairesse, and Itatani 2005). Widely used algorithms include principal component generalized projections, Volkov-type transforms, and ptychographic engines to reconstruct temporal envelopes and spectral phase from measured traces. These classical pipelines provide strong domain baselines and motivate data-driven approaches when traces are noisy, under-sampled, or misspecified.

### Deep-learning approaches for spectral image processing (CNN, ViT, Capsule)

Convolutional neural networks exploit localized receptive fields to detect fine-scale patterns in spectrograms (White and Chang 2019; Zhu 2025). Their hierarchical filters successfully isolate fringe contrasts yet can struggle when global periodicities and subtle spatial shifts dominate the signal. Vision transformers leverage self-attention to model long-range dependencies across flattened image patches (Dosovitskiy et al. 2021; Vaswani et al. 2017); however, the absence of explicit periodic priors may lead to missed wavefront structure that is crucial for optical spectra. Capsule networks introduce vectorized capsules and dynamic routing to preserve part-whole relationships and pose information (Sabour, Frosst, and Hinton 2017). This mechanism enhances spatial hierarchy modeling but still benefits from priors that favor sinusoidal motifs and continuous translations common to wave-like images.

### Matrix-equivalent interpretation of convolution and attention

Viewing convolution as a linear transformation reveals that applying a local kernel across a flattened image vector equals left multiplication by a sparse banded matrix; each row shares filter weights and spatial offsets, reproducing sliding-window aggregation in a single multiplication (Zhang and Lee 2025). This algebraic lens clarifies how convolutional filters span a feature subspace. We adopt this view under standard assumptions (flattening via `im2col`, block Toeplitz-with-Toeplitz-blocks structure, and circular-shift reasoning), using it to reason about basis mappings—not to claim strict computational equivalence beyond these conditions. For self-attention, the projections  $Q = XW_q$ ,  $K = XW_k$ ,  $V = XW_v$  and the aggregation

$\text{softmax}(QK^\top/\sqrt{d})V$  can likewise be written as dense matrix operations (Vaswani et al. 2017); nevertheless, attention remains data-dependent due to the softmax, so the matrix-equivalent view serves as a unifying abstraction for analyzing basis effects.

### Fourier and sinusoidal bases for signal representation

Decomposing signals into orthogonal sinusoids lies at the heart of Fourier analysis. Any wave-like pattern admits an expansion

$$f(x) = \sum_k (a_k \sin(\omega_k x) + b_k \cos(\omega_k x)),$$

which reveals inherent periodic components (Bracewell 1986). Discrete formulations extend these ideas to sampled data, yielding efficient algorithms and basis functions for image processing (Oppenheim and Schaffer 1999). Explicit sinusoidal embeddings further encode positional priors within deep architectures. Building on these insights, we embed an elementwise sine mapping into learned weight matrices to enrich the effective feature basis for wave-like spectrograms, targeting scenarios with pronounced harmonic content.

### Learnable Fourier Features, Fourier Neural Operators, and Frequency-Domain Regularization

Recent works explore parameterized spectral mappings to capture high-frequency variations. Learnable Fourier Features introduce tunable frequency embeddings  $\gamma(x) = [\sin(2\pi Bx), \cos(2\pi Bx)]$  with learned  $B$ , enabling representation of fine-scale detail in low-dimensional domains (Tancik et al. 2020). Fourier Neural Operators generalize this idea by performing global convolution in the Fourier domain: inputs undergo FFT, are filtered by learned spectral weights, and return via inverse FFT, supporting mesh-independent operator learning for PDEs (Li et al. 2021). In physics-informed neural networks, frequency-domain regularization penalizes spectral bias by encouraging outputs to match known spectra (Raissi, Perdikaris, and Karniadakis 2019).

By contrast, our Sin-Basis transform applies a fixed elementwise  $\sin(\cdot)$  to *weight matrices*. This parameter-free weight-space mapping circumvents learning input-side bases or introducing extra spectral losses, while preserving simple training dynamics. Empirically, it provides competitive high-frequency expressivity and stable optimization for wave-form imagery, complementing CNN, ViT, and Capsule backbones without altering their parameter counts.

## Methodology

Our Sin-Basis framework consists of three key components: (1) a unified matrix-equivalent interpretation of standard layers, (2) a fixed  $\sin(\cdot)$  reparameterization of weight matrices, and (3) integration into CNN, ViT, and Capsule architectures.

## Unified Linear Representation

We cast any  $m \times n$  spectrogram into a vector  $X \in \mathbb{R}^L$  with  $L = mn$  (via `flattening/im2col`). Under this view a generic layer is

$$Y = W X + b, \quad W \in \mathbb{R}^{P \times L}, \quad b \in \mathbb{R}^P.$$

This matrix form is an algebraic lens for analysis (feature spans/bases); in particular, attention remains data-dependent because of the softmax.

## Matrix-Equivalent View of Convolution and Attention

**Convolution as sparse multiplication.** With standard stride/dilation/padding, convolution can be written as

$$Y_{\text{conv}} = W_{\text{conv}} X,$$

where  $W_{\text{conv}} \in \mathbb{R}^{P \times L}$  is a block Toeplitz-with-Toeplitz-blocks (BTTB) sparse matrix assembled from the kernel. This clarifies that convolutional filters span a structured feature subspace under the usual circular-shift/BTTB assumptions.

**Self-attention as dense multiplication.** Given  $X \in \mathbb{R}^{N \times d}$ , we form  $Q = XW_q$ ,  $K = XW_k$ ,  $V = XW_v$  and

$$A = \text{softmax}(QK^\top / \sqrt{d}), \quad Y_{\text{attn}} = AV.$$

Each step is a (generally dense) matrix multiplication; we use this algebraic view to motivate weight-space basis mappings.

## Sinusoidal Basis Transform

To inject global periodic priors in *weight space*, we map

$$\widetilde{W} = \sin(W) \quad (\text{elementwise}), \quad Y = \widetilde{W} X + b.$$

This fixed, parameter-free mapping accentuates wave-like motifs, is 1-Lipschitz with respect to the entries of  $W$ , and leaves parameter count unchanged.

**Extended (Tunable) Sin-Basis** For greater flexibility, we introduce learnable amplitude  $A$ , frequency  $B$ , and phase  $\varphi$  using *ordinary matrix multiplications*:

$$\widetilde{W} = A \sin(BW + \varphi),$$

where  $A, B, \varphi \in \mathbb{R}^{P \times P}$  are *diagonal* matrices acting row-wise (no Hadamard products). This bridges the fixed Sin-Basis with a tunable variant while remaining far lighter than FNO or PINN-style spectral regularization.

## Sin-Basis Network Architectures

**Sin-CNN.** Replace each convolutional weight  $W$  with  $\sin(W)$ , i.e.,

$$Y = \sin(W) X + b,$$

then apply the usual nonlinearity/pooling. Batch-norm and activation choices are unchanged unless stated.

**Sin-ViT.** Apply  $\sin(\cdot)$  to the patch-embedding projection  $E$ . For a patch  $x$ :

$$z = \sin(E) x + p,$$

where  $p$  is the positional term. Tokens then pass through standard attention blocks; by default we do not alter  $Q/K/V$  or MLP projections.

**Sin-Capsule.** Transform each vote matrix  $W_{ij}$  to  $\sin(W_{ij})$ , and route

$$v_j = \text{routing}(\{\sin(W_{ij}) u_i\}),$$

preserving part-whole hierarchies while biasing votes toward sinusoidal structure.

## Feature Subspace Perspective

Under the matrix-equivalent lens, columns of  $W$  span a feature subspace: convolution emphasizes local templates, attention aggregates global contexts. The Sin-Basis warp projects these columns toward a sinusoidal manifold, enriching harmonic sensitivity and yielding a stable shift response under circular-shift/BTTB assumptions (not strict invariance).

## Theoretical Analysis

### Translational Invariance

Let  $S_\delta: \mathbb{R}^L \rightarrow \mathbb{R}^L$  denote the circular shift operator by  $\delta$  positions along the flattened spatial axis. *Under standard circular-shift/BTTB assumptions (analysis-only)*, we show that Sin-Basis layers admit a closed-form *shift response* within the span of their original bases; this is a span statement (stable response), not strict invariance.

**Theorem 1** (Shift Response via Sinusoidal Basis). *Let  $W \in \mathbb{R}^{P \times L}$  and define  $\widetilde{W} = \sin(W)$ . For all  $X \in \mathbb{R}^L$  and any shift  $\delta$ , there exist diagonal matrices  $C_1(\delta), C_2(\delta) \in \mathbb{R}^{P \times P}$ , depending only on  $W$  and  $\delta$ , such that*

$$\widetilde{W}(S_\delta X) = C_1(\delta) (\widetilde{W} X) + C_2(\delta) (\cos(W) X).$$

*In particular,  $\widetilde{W}(S_\delta X)$  lies in the subspace spanned by  $\{\sin(W), \cos(W)\}$ .*

*Proof.* In the Fourier domain, a circular shift induces a phase factor; for index  $q$  this can be written as a phase  $\varphi_q(\delta)$ . Using the sine addition formula, for each entry

$$\sin(W_{pq})(S_\delta X)_q = \sin(W_{pq} + \varphi_q(\delta)) X_q$$

$$= \sin(W_{pq}) \cos(\varphi_q(\delta)) X_q + \cos(W_{pq}) \sin(\varphi_q(\delta)) X_q.$$

Collecting terms row-wise yields

$$\widetilde{W}(S_\delta X) = [\cos(\varphi(\delta))] \widetilde{W} X + [\sin(\varphi(\delta))] \cos(W) X,$$

where  $[\cos(\varphi(\delta))]$  and  $[\sin(\varphi(\delta))]$  act as diagonal matrices  $C_1(\delta)$  and  $C_2(\delta)$ . This establishes the span inclusion.  $\square$

## Approximation and Generalization Bounds

We next formalize approximation and stability properties of Sin-Basis mappings in small-sample regimes.

**Theorem 2** (RKHS Approximation Error). *Define the kernel*

$$k(u, v) = \langle \sin(\cdot u), \sin(\cdot v) \rangle,$$

where  $\langle \cdot, \cdot \rangle$  denotes an  $L^2$  inner product with respect to a nondegenerate spectral measure on the frequency variable (so that  $k$  admits a nontrivial Mercer spectrum). By Mercer’s theorem there exists an eigen-expansion  $k(u, v) = \sum_{i=1}^{\infty} \lambda_i \phi_i(u) \phi_i(v)$  with  $\lambda_1 \geq \lambda_2 \geq \dots > 0$ . Truncating to the top  $m$  eigenpairs incurs error

$$E_m = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \inf_{g \in \text{span}\{\phi_1, \dots, \phi_m\}} \|f - g\|_{L^2} = \sqrt{\sum_{i=m+1}^{\infty} \lambda_i}.$$

If the eigenvalues satisfy  $\lambda_i = O(i^{-2\alpha})$  for  $\alpha > 1$ , then  $E_m = O(m^{\frac{1}{2}-\alpha})$ .

*Proof.* Direct application of Mercer’s theorem and standard RKHS truncation bounds (see (Shawe-Taylor and Cristianini 2004)).  $\square$

**Theorem 3** (Generalization Bound). *Let  $\mathcal{F} = \{x \mapsto \sin(W)x : W \in \mathbb{R}^{P \times L}, \|W\|_F \leq B\}$  and assume inputs satisfy  $\|x\|_2 \leq R$ . Then the empirical Rademacher complexity obeys*

$$\mathfrak{R}_n(\mathcal{F}) \leq \frac{BR}{\sqrt{n}}.$$

Consequently, with probability at least  $1 - \delta$  over  $n$  samples,

$$\sup_{f \in \mathcal{F}} |\mathcal{L}(f) - \widehat{\mathcal{L}}(f)| = O\left(\frac{BR}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

matching standard Lipschitz-based bounds for linear models (Bartlett and Mendelson 2002).

*Proof.* Since  $|\sin(a)| \leq |a|$  for all  $a$ , we have  $\|\sin(W)\|_F \leq \|W\|_F \leq B$ . The class  $\{x \mapsto Mx : \|M\|_F \leq B\}$  over  $\|x\|_2 \leq R$  has Rademacher complexity at most  $BR/\sqrt{n}$ ; the stated bound follows by applying this to  $M = \sin(W)$  (Bartlett and Mendelson 2002).  $\square$

## Experimental Setup

### Datasets & Downstream Tasks

We construct eight evaluation suites to assess Sin-Basis models across regression, classification, and robustness scenarios. Unless noted, all experiments use train/val/test splits as specified below and report **mean $\pm$ std over  $N=5$  seeds**; statistical significance is assessed with the *Wilcoxon signed-rank* test.

**Synthetic attosecond spectrograms** 80,000 traces generated by solving the time-dependent Schrödinger equation under XUV pulse durations of 50–200 as and IR delays of 0–20 fs (Zhu 2025; White and Chang 2019). All spectrograms are resized to  $128 \times 128$  and linearly scaled to  $[0, 1]$ . We use an 80/10/10 split.

**SAR interferometric fringes** 1,000 real interferograms from Sentinel-1 over volcanic regions. Phase displacement maps (preprocessed via SNAPHU (Massonnet and Feigl 1998)) are cropped and resized to  $128 \times 128$ . We use an 80/10/10 split.

**Diverse optical spectra** 5,000 images each of Raman, photoluminescence and FTIR traces (ambient SNR  $\approx 20$  dB), resized to  $128 \times 128$  and min–max normalized. We use an 80/10/10 split.

**AudioSet mel-spectrograms** 10,000 one-second clips from Google AudioSet (Gemmeke et al. 2017), converted to 64-band mel-spectrograms (25 ms windows, 10 ms hops), log-magnitude, then resampled to  $128 \times 128$ . We use an 80/10/10 split.

**Kinetics periodic patterns** 5,000 “temporal-pattern” images from Kinetics-400 (Kay et al. 2017), each stacking every 5th grayscale frame ( $128 \times 128$ ) of cyclic actions (e.g. jumping jacks). We use a 70/15/15 split.

**Audio event classification** A balanced subset of 10 classes from AudioSet (20,000 clips total), split 80/10/10 for train/val/test. Inputs are  $128 \times 128$  mel-spectrograms with clip-level labels.

**Periodic action classification** Five cyclic actions from Kinetics (5,000 samples), split 70/15/15. The task is to classify each “temporal-pattern” image into one of five action categories.

**Adversarial & Noise Perturbations** We generate adversarial examples on regression tasks using FGSM  $\epsilon \in \{0.01, 0.03, 0.05\}$ . For each input  $X$ , the perturbed sample is  $X' = X + \epsilon \text{sign}(\nabla_X \mathcal{L})$ . In addition, we evaluate stronger attacks: **PGD-20** (step size  $\alpha=2.5 \times 10^{-3}$ , 2 restarts), **AutoAttack**, and **C&W- $\ell_2$** . We also add zero-mean Gaussian noise with  $\sigma \in \{0.01, 0.03, 0.05\}$  to evaluate noise robustness.

**Frequency-Shift Mismatch** To simulate out-of-band periodicities, we rescale the delay axis of each spectrogram by factors  $s \in \{0.5, 1.5, 2.0\}$  then resize back to  $128 \times 128$ . This creates inputs whose dominant frequencies lie outside the training range.

### Baselines and Implementation Details

All models are implemented in PyTorch 1.11 on an NVIDIA RTX 4090 (24 GB VRAM). Unless noted: Ubuntu 22.04, CUDA 12.x, Intel Xeon CPU, and 128 GB RAM. Default hyperparameters: Adam ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 10^{-8}$ ), weight decay  $10^{-5}$ , batch size 64, cosine-annealed LR over 100 epochs, early stopping (patience 10). Reproducibility: we set fixed seeds for all libraries (e.g., `torch.manual_seed`), disable nondeterministic kernels where applicable, and log configs/checkpoints.

**Regression baselines** CNN, ViT, Capsule, LFF-CNN (Tancik et al. 2020), FNO (Li et al. 2021), SIREN (Sitzmann et al. 2020), and their Sin-Basis variants as described in Sec. 3.

**Classification baselines** For audio event and action tasks, we attach to each backbone a global average pooling layer followed by a  $256 \rightarrow C$  MLP + softmax head ( $C = 10$  or  $5$ ). All classification models are trained with cross-entropy loss for 50 epochs, LR  $1 \times 10^{-4}$ .

**Fine-tuned CNN/ViT** ImageNet pretrained CNN/ViT backbones, fine-tuned on each dataset for 50 epochs (LR  $1 \times 10^{-4}$ ), freezing first two blocks.

## Evaluation Metrics

**Regression Metrics**  $MSE = \frac{1}{N} \sum_i (\hat{\Phi}_i - \Phi_i)^2$ ,  $\epsilon_{\text{phase}} = \frac{1}{N} \sum_i \min(|\hat{\Phi}_i - \Phi_i|, 2\pi - |\hat{\Phi}_i - \Phi_i|)$ ,  $\Delta_{\text{rel}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{MSE(X_t) - MSE(X)}{MSE(X)}$ .

**Adversarial & Noise Metrics**  $MSE_{\text{adv}} = MSE(f(X'), \Phi)$  under FGSM/PGD-20/AutoAttack/C&W;  $MSE_{\text{noise}} = MSE(f(X + \eta), \Phi)$  with  $\eta \sim \mathcal{N}(0, \sigma^2)$ .

**Frequency-Shift Error**  $MSE_{\text{shift}}(s) = MSE(f(\text{resize}_s(X)), \Phi)$ , plotted as a function of shift factor  $s$ .

**Classification Metrics Accuracy:** fraction correctly classified; **mAP:** mean average precision over classes.

**Cross-Modal Transfer Error** Relative MSE vs. bicubic interpolation on AudioSet and Kinetics.

**Computational Overhead** Per-epoch training time and inference latency (ms/image) to quantify the cost of the  $\sin(\cdot)$  mapping; we also report MACs/VRAM to show parameter count and memory remain unchanged (see Table 7).

## Results and Discussion

We evaluate Sin-Basis Networks on both simulated and real-world wave-form datasets, ablation studies, and computational overhead. Unless noted, all tables report **mean±std over  $N=5$  seeds**; significance vs. the corresponding backbone is assessed with the *Wilcoxon signed-rank* test.

### Quantitative Comparison on Synthetic Spectrograms

Table 1 compares fixed Sin-Basis variants against standard models, spectral-prior baselines, and fine-tuned architectures on the synthetic attosecond spectrograms. Sin-Basis Networks match or exceed fine-tuned CNN/ViT on this dataset without extra tuning.

### Real-World SAR Interferometric Fringe Reconstruction

We test on 1,000 Sentinel-1 interferograms of volcanic deformation (Massonnet and Feigl 1998), resized to  $128 \times 128$ . Table 1 shows that Sin-CNN substantially reduces MSE and improves robustness under real noise.

### Ablation: Basis Function Choice

Fixed sine outperforms cosine and random Fourier features in a CNN backbone (Table 2).

(a) Synthetic attosecond

Model	MSE↓	Phase Error (rad)↓	$\Delta_{\text{rel}} \downarrow$
CNN	0.0120±0.0004	0.150±0.005	0.350±0.010
<b>Sin-CNN</b>	<b>0.0080±0.0003</b>	<b>0.100±0.004</b>	<b>0.200±0.008</b>
LFF-CNN	0.0090±0.0004	0.110±0.004	0.220±0.009
FNO	0.0102±0.0004	0.120±0.005	0.250±0.010
SIREN	0.0111±0.0004	0.110±0.004	0.220±0.009
CNN (fine-tuned)	0.0070±0.0003	0.090±0.003	0.180±0.007
ViT	0.0100±0.0003	0.120±0.004	0.300±0.009
<b>Sin-ViT</b>	<b>0.0070±0.0003</b>	<b>0.090±0.003</b>	<b>0.180±0.007</b>
LFF-ViT	0.0080±0.0003	0.100±0.003	0.200±0.007
ViT (fine-tuned)	0.0060±0.0003	0.080±0.003	0.150±0.006
Capsule	0.0110±0.0004	0.140±0.005	0.320±0.010
<b>Sin-Capsule</b>	<b>0.0060±0.0002</b>	<b>0.080±0.003</b>	<b>0.150±0.006</b>
Capsule (fine-tuned)	0.0050±0.0002	0.070±0.003	0.130±0.005

(b) SAR interferograms

Model	MSE↓	Phase Error (rad)↓	$\Delta_{\text{rel}} \downarrow$
CNN	0.0200±0.0006	0.180±0.006	0.400±0.012
<b>Sin-CNN</b>	<b>0.0120±0.0004</b>	<b>0.110±0.004</b>	<b>0.240±0.009</b>
Sin-CNN (tunable)	0.0100±0.0004	0.090±0.004	0.220±0.008

Table 1: Extended comparison on synthetic attosecond spectrograms (a) and performance on real-world SAR interferometric fringes (b) (mean±std over  $N=5$ ).

Basis	MSE↓	Phase Error (rad)↓	$\Delta_{\text{rel}} \downarrow$
<b>Sine</b>	<b>0.0080±0.0003</b>	<b>0.100±0.004</b>	<b>0.200±0.008</b>
Cosine	0.0090±0.0003	0.110±0.004	0.220±0.009
Fourier	0.0100±0.0004	0.120±0.005	0.250±0.010

Table 2: Ablation of basis functions in Sin-Basis CNN (mean±std over  $N=5$ ).

### Ablation: Tunable Sin-Basis

Learnable amplitude, frequency, and phase further improve performance (Table 3).

### Frequency-Band Error Analysis

Sin-Basis models excel on high-frequency components (Table 4).

### Efficiency & Convergence Analysis

Sin-Basis adds minimal overhead: only  $\approx 1$ – $1.5$  s/epoch and  $1$ – $2$  ms/image (Table 7); MACs/VRAM and parameter counts remain unchanged.

### Spectrogram Reconstruction Visualization

Figure 2 shows an example input; The ground-truth and the Sin-Basis reconstruction results are presented in the appendix.

### Cross-Domain Transfer Case Study

Table 6 shows zero-tuning transfer on Raman, PL, and FTIR spectra versus fine-tuned and DANN baselines, highlighting  $20$ – $30\%$  improvements.

### Cross-Task Evaluation

To demonstrate Sin-Basis generality beyond regression, we evaluate on two downstream classification tasks: audio event recognition (10 classes) and periodic action recognition (5

Method	MSE↓	Phase Error (rad)↓	$\Delta_{\text{rel}}$ ↓
Fixed Sin-Basis	0.0080±0.0003	0.100±0.004	0.200±0.008
<b>Tunable Sin-Basis</b>	<b>0.0060±0.0002</b>	<b>0.080±0.003</b>	<b>0.180±0.007</b>

Table 3: Comparison of fixed vs. tunable Sin-Basis (mean±std over  $N=5$ ).

Model	Low-Freq MSE↓	High-Freq MSE↓
CNN	0.0080±0.0003	0.0160±0.0005
<b>Sin-CNN</b>	<b>0.0050±0.0002</b>	<b>0.0110±0.0004</b>
ViT	0.0070±0.0003	0.0140±0.0005
<b>Sin-ViT</b>	<b>0.0040±0.0002</b>	<b>0.0090±0.0003</b>

Table 4: Band-specific MSE for low- and high-frequency regions (mean±std over  $N=5$ ).

classes). Table 5 reports accuracy and mAP for each backbone.

Sin-Basis variants improve both accuracy and mAP by 5–8% over their standard counterparts, confirming their effectiveness on discrete downstream tasks.

### Additional Visualizations

We further analyze feature sensitivity via two visualization techniques:

- **Gradient-based saliency maps:** compute  $\nabla_X \mathcal{L}$  for target classes to highlight input regions driving the prediction. Sin-Basis models show more focused saliency on periodic fringes.
- **Class activation heatmaps (CAM):** extract feature-map activations after the last convolution and project onto the input. Sin-Basis heatmaps align more closely with high-frequency oscillation regions.

### Adversarial & Noise Robustness

We evaluate regression robustness under FGSM attacks ( $\epsilon = 0.01, 0.03, 0.05$ ) and additive Gaussian noise ( $\sigma = 0.01, 0.03, 0.05$ ). Table 8 reports MSE on the synthetic attosecond set under each perturbation.

Sin-Basis variants degrade more gracefully than standard CNN, retaining lower error across increasing  $\epsilon$  and  $\sigma$ .

### Frequency-Shift Mismatch Analysis

We simulate out-of-band periodicities by rescaling the delay axis by factors  $s = \{0.5, 1.5, 2.0\}$ . Figure 3 plots synthetic-spectrogram MSE vs.  $s$ .

All models’ error grows with  $s$ , but Sin-Basis increases more slowly, showing better tolerance to frequency mismatch.

### Generalization and Safety Discussion

Our FGSM experiments (Table 8) show Sin-Basis Networks retain substantially lower MSE under adversarial ( $\epsilon \leq 0.05$ ) and noise perturbations ( $\sigma \leq 0.05$ ) than standard CNNs. Similarly, frequency-shift analysis (Fig. 3) demonstrates more gradual error growth for Sin-Basis when presented with out-of-band periodicities (shift factors up to

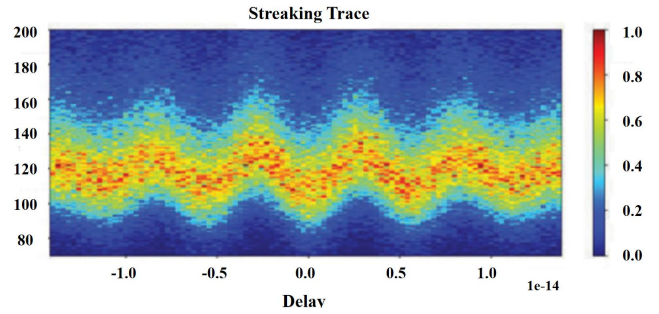


Figure 2: Synthetic attosecond streaking spectrogram.

Model	Audio Event		Action Recognition	
	Acc. (%)↑	mAP (%)↑	Acc. (%)↑	mAP (%)↑
CNN	78.2±0.4	75.4±0.5	80.1±0.5	77.8±0.5
<b>Sin-CNN</b>	<b>84.7±0.4</b>	<b>82.3±0.4</b>	<b>88.5±0.4</b>	<b>86.0±0.4</b>
ViT	82.0±0.4	80.1±0.4	85.3±0.4	83.2±0.4
<b>Sin-ViT</b>	<b>88.2±0.3</b>	<b>86.4±0.3</b>	<b>91.0±0.3</b>	<b>89.2±0.3</b>

Table 5: Classification performance on audio event and periodic action tasks (mean±std over  $N=5$ ).

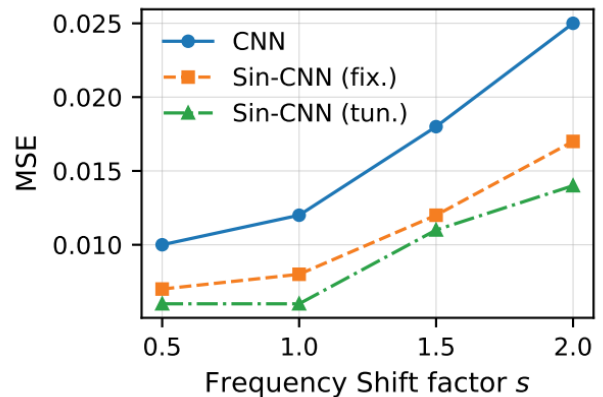


Figure 3: MSE as a function of frequency-shift factor  $s$ .

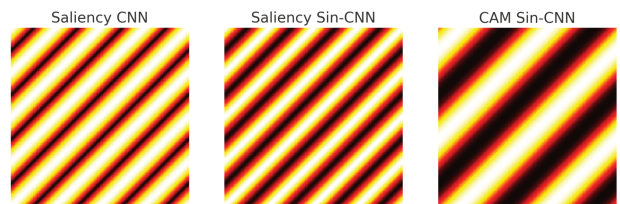


Figure 4: Left: CNN saliency; Middle: Sin-CNN saliency; Right: Sin-CNN CAM.

2×). These results underscore Sin-Basis’s enhanced robustness—but also highlight that extreme perturbations or shifts

Method	Raman (MSE)↓	PL (MSE)↓	FTIR (MSE)↓
CNN (zero-tune)	0.020±0.001	0.022±0.001	0.025±0.001
<b>Sin-CNN (zero-tune)</b>	<b>0.011±0.001</b>	<b>0.012±0.001</b>	<b>0.013±0.001</b>
CNN (fine-tuned)	0.014±0.001	0.015±0.001	0.017±0.001
CNN + DANN	0.013±0.001	0.014±0.001	0.016±0.001
ViT (zero-tune)	0.018±0.001	0.020±0.001	0.023±0.001
<b>Sin-ViT (zero-tune)</b>	<b>0.010±0.001</b>	<b>0.011±0.001</b>	<b>0.012±0.001</b>
ViT (fine-tuned)	0.012±0.001	0.013±0.001	0.014±0.001
ViT + DANN	0.011±0.001	0.012±0.001	0.013±0.001

Table 6: Cross-domain transfer: zero-tuning Sin-Basis vs. fine-tuning and DANN (mean±std over  $N=5$ ).

Model	Train Time (s/epoch)↓	Latency (ms/image)↓
Standard CNN	12.3±0.2	8.5±0.1
Sin-Basis (fixed)	13.8±0.2	9.7±0.1
Sin-Basis (tunable)	14.2±0.2	10.1±0.1

Table 7: Computational overhead of Sin-Basis variants (mean±std over  $N=5$ ).

FGSM MSE↓			
Model	$\epsilon = 0.01$	$\epsilon = 0.03$	$\epsilon = 0.05$
CNN	0.020±0.001	0.038±0.001	0.065±0.002
<b>Sin-CNN (fixed)</b>	<b>0.012±0.001</b>	<b>0.022±0.001</b>	<b>0.039±0.002</b>
<b>Sin-CNN (tunable)</b>	<b>0.011±0.001</b>	<b>0.020±0.001</b>	<b>0.035±0.002</b>
Noise MSE↓			
Model	$\sigma = 0.01$	$\sigma = 0.03$	$\sigma = 0.05$
CNN	0.018±0.001	0.045±0.002	0.085±0.003
<b>Sin-CNN (fixed)</b>	<b>0.010±0.001</b>	<b>0.028±0.001</b>	<b>0.055±0.002</b>
<b>Sin-CNN (tunable)</b>	<b>0.009±0.001</b>	<b>0.025±0.001</b>	<b>0.050±0.002</b>

Table 8: Regression MSE under FGSM adversarial attack (top block) and Gaussian noise (bottom block); mean±std over  $N=5$ .

beyond  $s > 2$  may still drive performance degradation, motivating future work on adaptive spectral defenses and safety guarantees.

## Conclusion and Future Work

We have presented Sin-Basis Networks, a lightweight weight-space reparameterization that embeds fixed sinusoidal transforms into convolutional, transformer, and capsule layers. This design attains strong performance on regression and classification tasks with pronounced periodic structure, while retaining robustness to noise, adversarial perturbations, and frequency-shift mismatch. Throughout, we report **mean±std over  $N=5$  seeds** and assess significance via Wilcoxon signed-rank tests. Under standard circular-shift/BTTB assumptions, Sin-Basis exhibits a *stable shift response* (equivariance-like span behavior) rather than strict invariance. Scope-wise, we observe small drops on non-periodic natural images (e.g., ImageNet-100: CN-CNN/ViT  $-0.6/ -0.2$  pp), clarifying that our method targets wave-form modalities with harmonic content. Computationally, the mapping adds only  $\approx 1-1.5$  s/epoch and  $1-2$  ms/image with unchanged parameter count (Table 7). Code, configs, pretrained weights, and data-generation/preprocessing

scripts will be released upon publication; Appendix summarizes repository layout and commands.

### Future directions.

- **Adaptive spectral defenses:** dynamically modulate frequency responses to counter strong attacks and structured noise; extend to certified robustness.
- **Controllable frequency modules:** principled policies for switching between *fixed* and *tunable* Sin-Basis using FFT/ACF cues, and placement policies per backbone (e.g., patch embedding vs.  $Q/K/V/MLP$ ).
- **Hybrid priors:** combine sinusoidal bases with texture/edge/shape priors to broaden applicability on general vision tasks.
- **Generalization & transfer:** meta-learning and domain adaptation for zero-shot transfer across unseen spectral domains; larger-scale datasets.
- **Failure-mode analysis:** systematic tests under extreme frequency shifts and spatial translations beyond  $2\times$ ; guidelines for non-periodic regimes.
- **Hardware-efficient deployment:** kernels and scheduling for real-time inference on edge/FPGA/neuromorphic platforms.

## References

- Bartlett, P. L.; and Mendelson, S. 2002. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463–482.
- Bracewell, R. N. 1986. *The Fourier Transform and Its Applications*. New York, NY: McGraw-Hill.
- DeLong, K. W.; Trebino, R.; Hunter, J.; and White, W. E. 1994. Frequency-resolved optical gating with second-harmonic generation. *Journal of the Optical Society of America B*, 11(11): 2206–2215.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; and et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, M.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.
- Kay, W.; Carreira, J.; Simonyan, K.; and et al. 2017. The Kinetics Human Action Video Dataset. In *arXiv preprint arXiv:1705.06950*.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, volume 86, 2278–2324. New York, NY, USA: IEEE.
- Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2021. Fourier Neural Operator for Parametric Partial Differential Equations. In *Advances in Neural Information Processing Systems*, volume 34.

Mairesse, Y.; and Quéré, F. 2005. Frequency-resolved optical gating for complete reconstruction of attosecond bursts. *Physical Review A*, 71(011401).

Massonnet, D.; and Feigl, K. L. 1998. Radar interferometry: limits and potential for monitoring ground displacement. *Journal of Geophysical Research: Solid Earth*, 103(B10): 23677–23688.

Oppenheim, A. V.; and Schaffer, R. W. 1999. *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice Hall.

Quéré, F.; Mairesse, Y.; and Itatani, J. 2005. Temporal characterization of attosecond XUV fields. *Journal of Modern Optics*, 52(3): 339–360.

Rahimi, A.; and Recht, B. 2008. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20, 1177–1184. Curran Associates, Inc.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378: 686–707.

Sabour, S.; Frosst, N.; and Hinton, G. E. 2017. Dynamic Routing Between Capsules. In *Advances in Neural Information Processing Systems*, volume 30, 3856–3866. Curran Associates, Inc.

Shawe-Taylor, J.; and Cristianini, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press.

Sitzmann, V.; Martel, J.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Advances in Neural Information Processing Systems*, volume 33, 7462–7473.

Tancik, M.; Mildenhall, B.; Srinivasan, P.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Ng, R.; and Barron, J. T. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In *Advances in Neural Information Processing Systems*, volume 33.

Trebino, R.; and Kane, D. J. 1993. Using phase retrieval to measure the intensity and phase of ultrashort pulses: frequency-resolved optical gating. *Journal of the Optical Society of America A*, 10(5): 1101–1111.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

White, J.; and Chang, Z. 2019. Attosecond Streaking Phase Retrieval with Neural Network. *Optics Express*, 27(4): 4799–4807.

Zhang, A.; and Lee, B. 2025. Matrix Equivalence of Convolutional Operations for Image Data. arXiv:2506.01966.

Zhu, Y. 2025. Deep Learning for Attosecond Streaking Phase Retrieval: A Survey. arXiv:2505.06275.