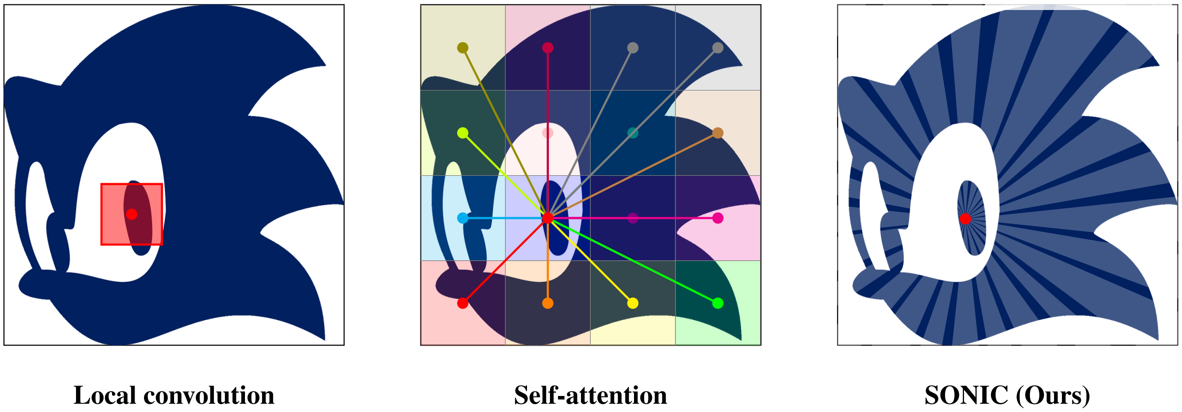
SONIC: SPECTRAL ORIENTED NEURAL INVARIANT CONVOLUTIONS

Anonymous authors
Paper under double-blind review



Local convolution **Self-attention** **SONIC (Ours)**

ABSTRACT

Convolutional Neural Networks (CNNs) rely on fixed-size kernels scanning local patches, which limits their ability to capture global context or long-range dependencies without very deep architectures. Vision Transformers (ViTs), in turn, provide global connectivity but lack spatial inductive bias, depend on explicit positional encodings, and remain tied to the initial patch size. Bridging these limitations requires a representation that is both structured and global. We introduce **SONIC (Spectral Oriented Neural Invariant Convolutions)**, a continuous spectral parameterisation that models convolutional operators using a small set of shared, orientation-selective components. These components define smooth responses across the full frequency domain, yielding global receptive fields and filters that adapt naturally across resolutions. Across synthetic benchmarks, large-scale image classification, and 3D medical datasets, SONIC shows improved robustness to geometric transformations, noise, and resolution shifts, and matches or exceeds convolutional, attention-based, and prior spectral architectures with an order of magnitude fewer parameters. These results demonstrate that continuous, orientation-aware spectral parameterisations provide a principled and scalable alternative to conventional spatial and spectral operators.

1 INTRODUCTION

Human visual processing is a remarkably complex and efficient system. It enables us to effortlessly recognise objects, detect and interpret motion, and comprehend complex scenes, adapting seamlessly across varying orientations, scales, resolutions, and even under degraded conditions, where computer vision methods often struggle. Serving as a benchmark due to its exceptional effectiveness under different circumstances, human vision highlights the areas where current artificial systems still exhibit limitations; Bridging this gap remains a central challenge in computer vision, driving the development of models that more closely approximate the versatility and robustness of human perception.

Multi-Layer Perceptrons (MLPs), as the earliest neural network models, demonstrated the feasibility of learning complex mappings but lacked the inductive biases required for large-scale vision tasks. Convolutional Neural Networks (CNNs) (LeCun et al., 2015), widely used for many vision tasks, rely on fixed-size kernels scanning local image patches. While effective for capturing local features like edges and textures, this design limits their ability to understand the overall context or capture long-range dependencies without relying on very deep architectures (as demonstrated by Luo et al. (2017). Critically, their effectiveness is limited by sensitivity to slight geometric variations, including translations (in particular out-of-frame translations), rescalings, rotations, and mild distortions (Azulay & Weiss, 2018). Vision Transformers (ViTs) (Dosovitskiy et al., 2020), inspired by advances in natural language processing, overcome this limitation by dividing images into sequences of patches and applying self-attention. This design directly models global context and alleviates the locality constraints of CNNs. Nevertheless, the self-attention mechanism is computationally demanding, as its cost grows quadratically with the number of image patches, and thus with the image area, which poses significant challenges for high-resolution inputs. Furthermore, Vision Transformers lack CNN-style spatial inductive biases and therefore require explicit mechanisms (e.g. positional encodings) to model positional relationships, and their accuracy–compute trade-off is closely tied to the chosen patch size. With the proposed method, which enables global receptive fields using significantly fewer parameters, we aim to narrow this conceptual gap and move computer vision models toward resolution-invariant perception, drawing inspiration from the robustness and adaptability of human-like visual processing.

Contribution In this paper, we introduce a theoretically grounded spectral framework for multidimensional signals that naturally provides global receptive fields, full convolutional expressiveness, and inherent resolution invariance, offering a lightweight yet versatile foundation that can support progress toward more scalable and adaptable vision models. The remainder of this paper is organised as follows. Section 2 introduces the mathematical preliminaries and related works. Section 3 presents the formulation of the SONIC approach together with implementation details. Section 4 reports the experimental results. Section 5 discusses the limitations of the proposed method and outlines directions for future research.

2 Background

Modern vision tasks demand the ability to integrate information over long spatial ranges. Although natural images often exhibit long-range structure, standard convolutions remain bounded by local receptive fields, making standard architectures inefficient, as many layers are effectively used to propagate information across the image rather than to learn increasingly abstract representations. Across established methods, global context is mostly obtained only indirectly, motivating the study of operators that provide global receptive fields as an intrinsic property of a single layer. This section reviews the mathematical foundations of such operators and develops the framework of spectral operators that underpins our approach.

Spatial-domain operators. In the spatial domain, enlarging the receptive field requires expanding the support of the discrete kernel. Large-kernel convolutions increase the neighbourhood size directly (Ding et al., 2022), dilated convolutions introduce gaps to cover larger regions without increasing the number of parameters (Yu & Koltun, 2015) and non-local (Wang et al., 2018) methods target long-range interactions; however, despite their effectiveness, convolutional layers implement filtering over a fixed sampling grid, an approach that implicitly assumes locality and smooth variation in the underlying signal. These assumptions hold for small neighborhoods but break down over large spatial ranges, where long-range structure cannot be captured efficiently through local interactions alone. As receptive fields expand, spatial filters become increasingly tied to the image resolution and scale, limiting their generalization and efficiency. Another well-studied strategy is to use self-attention mechanisms (Chen et al., 2018; Dosovitskiy et al., 2021), which compute pairwise interactions across all spatial positions and therefore provide a principled way to model long-range dependencies. However, these approaches incur computational and memory costs that grow rapidly with image resolution: large kernels scale with their area $O(K^2)$, attention scales quadratically with the number of tokens, and as resolution increases, these scaling properties make such mechanisms difficult to deploy efficiently, especially in high-resolution domains where global context is important but computational budgets are constrained.

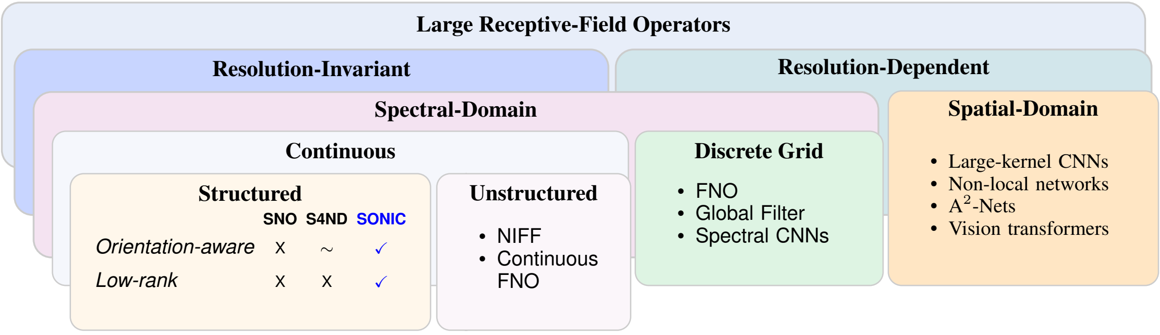


Figure 1: (a) Taxonomy of large receptive-field operators.

Spectral-domain operators. An alternative paradigm achieves global receptive fields by representing operators directly in the frequency domain (Rippel et al., 2015). This approach uses the fact that every linear, shift-invariant operator on $\mathbb{R}^D$ is fully characterised, enabling information to propagate globally through a frequency-wise multiplication.

Let $D \in \mathbb{N}$ and consider vector-valued signals

$$x : \mathbb{R}^D \to \mathbb{C}^{C_{\text{in}}}, \qquad y : \mathbb{R}^D \to \mathbb{C}^{C_{\text{out}}}.$$

For a sufficiently regular scalar function $f : \mathbb{R}^D \to \mathbb{C}$, the Fourier transform is

$$\mathcal{F}_D[f](\boldsymbol{\omega}) = \int_{\mathbb{R}^D} f(\mathbf{x})\, e^{-i\,\boldsymbol{\omega}\cdot\mathbf{x}}\, d\mathbf{x}, \qquad \boldsymbol{\omega} \in \mathbb{R}^D. \tag{1}$$

and extends component-wise to vector-valued functions. A linear, shift-invariant operator acting on x has a convolution representation:

$$y(\mathbf{x}) = \int_{\mathbb{R}^D} k(\mathbf{x} - \mathbf{z})\, x(\mathbf{z})\, d\mathbf{z}, \tag{2}$$

where $k : \mathbb{R}^D \to \mathbb{C}^{C_{\text{out}} \times C_{\text{in}}}$. The Convolution Theorem gives

$$\mathcal{F}_D[y](\boldsymbol{\omega}) = \widehat{k}(\boldsymbol{\omega})\, \mathcal{F}_D[x](\boldsymbol{\omega}), \qquad \boldsymbol{\omega} \in \mathbb{R}^D. \tag{3}$$

where $\widehat{k}(\boldsymbol{\omega})$ is the Fourier transform of a spatial kernel in the classical convolution setting. In spectral neural methods, however, we do not constrain the operator to arise from any finite-support spatial kernel. Instead, we define the spectral kernel directly by

$$\widehat{k}(\boldsymbol{\omega}) := \widehat{H}(\boldsymbol{\omega}), \tag{4}$$

where $\widehat{H}(\boldsymbol{\omega})$ is the learnable frequency response of the operator. This viewpoint treats $\widehat{H}$ as the primary parametrisation, enabling general global and resolution-invariant operators beyond those that correspond to discrete spatial kernels. In practice, images are sampled on a discrete grid. Let the spatial domain be discretised using $N_1, \dots, N_D$ samples along each axis, with pixel spacings $\Delta_1, \dots, \Delta_D$. The corresponding DFT frequency sets are given by

$$\Omega_d = 2\pi \left\{ \frac{k_d}{N_d\, \Delta_d} \;\middle|\; k_d = -\left\lfloor \tfrac{N_d}{2} \right\rfloor, \dots, \left\lceil \tfrac{N_d}{2} \right\rceil - 1 \right\}, \quad d = 1, \dots, D. \tag{5}$$

The full frequency grid is the Cartesian product $\Omega = \Omega_1 \times \cdots \times \Omega_D$, containing $N = N_1 \cdots N_D$ discrete frequencies. The DFT samples $\widehat{x}$ are defined at frequencies $\boldsymbol{\omega}_n \in \Omega$.

Resolution invariance We formalise resolution invariance by defining the operator via a continuous spectral symbol that is independent of the sampling grid. Let

$$\widehat{H}_\theta : \mathbb{R}^D \to \mathbb{C}^{C_{\text{out}} \times C_{\text{in}}} \tag{6}$$

be a continuous function parameterised by θ. Given a discretisation (N, Δ) with Fourier grid $\Omega_{N,\Delta}$, the discretised operator is obtained via sampling:

$$\widehat{y}^{(N,\Delta)}(\boldsymbol{\omega}_n) = \widehat{H}_\theta(\boldsymbol{\omega}_n)\, \widehat{x}^{(N,\Delta)}(\boldsymbol{\omega}_n), \qquad \boldsymbol{\omega}_n \in \Omega_{N,\Delta}. \tag{7}$$

We term the operator resolution-invariant if θ depends only on the underlying physics of the layer, not on the discretisation (N, Δ). Changing resolution then simply corresponds to resampling the same continuous function $\widehat{H}_\theta$ on a new grid. GFNet (Rao et al., 2021) and FNO (Li et al., 2021) parameterise $\widehat{H}$ directly on the discrete FFT grid: GFNet learns a complex mask of size N, and FNO learns a fixed number of low-frequency coefficients. Since these coefficients correspond to specific frequency indices, changing resolution alters the operator itself. Thus, such models do not define a true resolution-invariant convolution operator.

Continuous Spectral Operators A principled way to overcome this limitation is to define the operator directly in the continuous Fourier domain and then evaluate it on the discrete grid provided by the data. In this formulation, the spectral symbol $\widehat{H}(\boldsymbol{\omega})$ is a continuous function of frequency, independent of the sampling pattern, and the discrete operator is obtained merely by sampling $\widehat{H}$ at the DFT frequencies of the current resolution. This yields a truly resolution-invariant parameterisation with global receptive fields. Two families of such continuous spectral operators appear in the literature:

- **Unstructured continuous operators**, which learn $\widehat{H}(\boldsymbol{\omega})$ as a general continuous function of frequency, typically via a small MLP or other low-dimensional parametrisation. Such models include neural implicit spectral filters and continuous FNO variants (Grabinski et al., 2024; Kabri et al., 2023), where the network outputs a complex response for any $\boldsymbol{\omega} \in \mathbb{R}^D$. While this provides maximal flexibility and full continuity in $\boldsymbol{\omega}$, these parameterisations are usually isotropic or weakly anisotropic: the learned response depends primarily on $\|\boldsymbol{\omega}\|$ unless orientation structure is encoded explicitly. Moreover, the channel mixing remains fully dense, offering little inductive bias regarding frequency orientation or cross-channel structure. As a result, these operators can be expressive but often parameter-inefficient.
- **Structured continuous operators** imposes additional structure on $\widehat{H}(\boldsymbol{\omega})$ through basis expansions, separability assumptions, or functional templates. Examples include SNO (Fanaskov & Oseledets, 2024), which expands the symbol in a fixed orthogonal basis, and multidimensional SSM-based kernels like S4ND (Nguyen et al., 2022), which impose axis-aligned or separable constructions inherited from one-dimensional state-space models. Although these multidimensional SSMs are not typically framed as spectral operators, their learned kernels do admit a structured frequency-domain representation and can be interpreted through the same lens as spectral methods. For completeness, we provide the frequency-domain form of S4ND in the appendix. These approaches provide improved inductive bias and parameter efficiency by coupling nearby frequencies and reusing spectral modes across channels. While more efficient, most structured models remain tied to coordinate axes. Their separability limits their ability to capture oriented or anisotropic patterns that lie along general directions in frequency space.

Natural images contain oriented structures such as edges, textures, and oscillations that correspond to directional features in frequency space. Standard separable parameterisations cannot easily model such behaviour. To represent interactions along arbitrary frequency directions, one must go beyond axis-aligned or tensor-product constructions and design spectral operators whose modes explicitly encode orientation in $\mathbb{R}^D$.

In this paper, we address this limitation by introducing SONIC: Spectral Oriented Neural Invariant Convolutions. SONIC is a Structured Continuous Spectral Operator that moves beyond axis-aligned constructions by explicitly parameterizing the spectral symbol $\widehat{H}_\theta(\boldsymbol{\omega})$ as a superposition of directional modes. This allows us to learn complex, oriented features in the frequency domain that are fully resolution-invariant, parameter-efficient, and inherently adapted to capturing the anisotropic structures present in natural signals.

3 Method

Overview. Many spectral neural methods are either axis–separable (efficient but limited in orientation) or fully nonlocal (powerful but inefficient and not spectrally faithful). Starting from linear time-invariant (LTI) systems, we extend the formulation to N-dimensional signals in the frequency domain, yielding a compact spectral representation. This framework models linear, shift–invariant operators through a shared low–rank structure, where oriented spectral transfer functions are applied at each frequency and mixed across channels by learned matrices B and C.

3.1 Formulation

Our method draws inspiration from the analytic structure of linear time-invariant systems, the same foundation underlying modern state-space models. To make this connection precise, consider the continuous-time LTI state-space system:

$$\dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \tag{8}$$

with zero initial condition. Its impulse response is obtained by setting $\mathbf{u}(t) = \delta(t)$:

$$\mathbf{K}(t) \;=\; \mathbf{C}\, e^{\mathbf{A}t}\, \mathbf{B}, \qquad t \geq 0. \tag{9}$$

The output equals the convolution of the input with the impulse response:

$$(\mathbf{K} * \mathbf{u})(t) \;=\; \int_0^\infty \mathbf{C}\, e^{\mathbf{A}\tau}\, \mathbf{B}\, \mathbf{u}(t-\tau)\, d\tau. \tag{10}$$

Taking the Laplace transform of the impulse response (derivations provided in Appendix C),

$$H(s) \;=\; \mathcal{L}\{\mathbf{K}(t)\}(s) = \; \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}. \tag{11}$$

The expression $H(s)$ above is the standard resolvent form that characterises the frequency response of a stable linear time–invariant system. We use this structure only as a modelling template: by replacing the scalar Laplace variable

s with the multi-dimensional spatial frequency $\boldsymbol{\omega}$, we obtain an analytic spectral parameterisation that inherits the smooth and structured behaviour of resolvent filters in D dimensions.

Let the input be $x \in \mathbb{R}^{C\times N_1\times\cdots\times N_D}$ and the output $y \in \mathbb{R}^{K\times N_1\times\cdots\times N_D}$. We denote their D-dimensional discrete Fourier transforms by

$$\widehat{x} = \mathcal{F}_D[x], \qquad y = \mathcal{F}_D^{-1}[\widehat{y}].$$

Central to our method is the transfer function $T_m(\boldsymbol{\omega})$, which defines the frequency response of a single mode. For each mode $m = 1, \ldots, M$ we set

$$T_m(\boldsymbol{\omega}) \;=\; \frac{1}{i\, s_m\,(\boldsymbol{\omega}\cdot\boldsymbol{v_m}) \;-\; a_m \;+\; \tau_m\,\|(I - \boldsymbol{v_m}\boldsymbol{v_m}^{\top})\boldsymbol{\omega}\|_2^2}, \tag{12}$$

Where each mode is parameterised by: (1) the orientation $\boldsymbol{v_m} \in \mathbb{R}^D$ with $\|\boldsymbol{v_m}\|_2 = 1$; (2) the scale $s_m > 0$ controlling spectral selectivity; (3) the real part $\mathrm{Re}(a_m)$ introducing damping; (4) the imaginary part $\mathrm{Im}(a_m)$ governing oscillatory behaviour; and (5) the transverse penalty $\tau_m \geq 0$ controlling decay orthogonal to $\boldsymbol{v_m}$. Together, these parameters shape the amplitude, orientation, and oscillatory nature of each spectral mode. The denominator replicates the resolvent structure of an LTI system. SONIC adopts this template by substituting the Laplace variable s with the oriented frequency component $i\, s_m(\boldsymbol{\omega}\;\cdot \boldsymbol{v}_m)$ and by adding a transverse decay term that enforces anisotropic filtering.

Rather than learning an unconstrained response $\widehat{\mathbf{H}}(\boldsymbol{\omega})$ for every frequency, SONIC factorises the spectral operator through M shared modes with entrywise form:

$$\widehat{H}_{k,c}(\boldsymbol{\omega}) = \sum_{m=1}^{M} C_{km}\, T_m(\boldsymbol{\omega})\, B_{mc}. \tag{13}$$

where $B \in \mathbb{C}^{M\times C}$ and $C \in \mathbb{C}^{K\times M}$. Given this factorised spectral response, the frequency-wise filtering applied to the input DFT is

$$\widehat{y}_k(\boldsymbol{\omega}) = \sum_{c=1}^{C} \widehat{H}_{k,c}(\boldsymbol{\omega})\, \widehat{x}_c(\boldsymbol{\omega}), \qquad k = 1, \ldots, K,\ \boldsymbol{\omega} \in \Omega, \tag{14}$$

where $\widehat{H}_{k,c}(\boldsymbol{\omega})$ is the frequency response of the $(c\!\to\!k)$ channel filter. This decomposition yields a compact, low-rank representation of the spectral operator, enabling expressive but parameter-efficient filtering. Following the frequency-domain filtering, the spatial output is added to a learnable skip projection and then passed through a pointwise nonlinearity, yielding the next-layer activation $x^{(\ell+1)}$:

$$x^{(\ell+1)} = \sigma\big(y^{(\ell)} + W_s x^{(\ell)}\big). \tag{15}$$

This nonlinear recursion allows multiple SONIC blocks to be stacked, providing depth-wise expressivity in the same manner as conventional convolutional architectures. Although SONIC is not a state-space model, its mode parameterisation is inspired by resolvent structures of linear time-invariant systems. Appendix C shows that restricting SONIC's orientations to the coordinate axes yields the Multidimensional SSM form.

3.2 Intuition

We use a compact collection of oriented modes that are shared across channels. Instead of learning an unconstrained spectrum for every input–output pair, each mode has a learnable analytic shape with a few learnable knobs, yielding interpretable, spatially localised filters. We also illustrate the effect of each parameter in Figure 2.

Each mode learns a preferred direction via a unit vector v_m, a compass needle in frequency space. Any frequency vector $\boldsymbol{\omega}$ decomposes uniquely into components along and across to this needle:

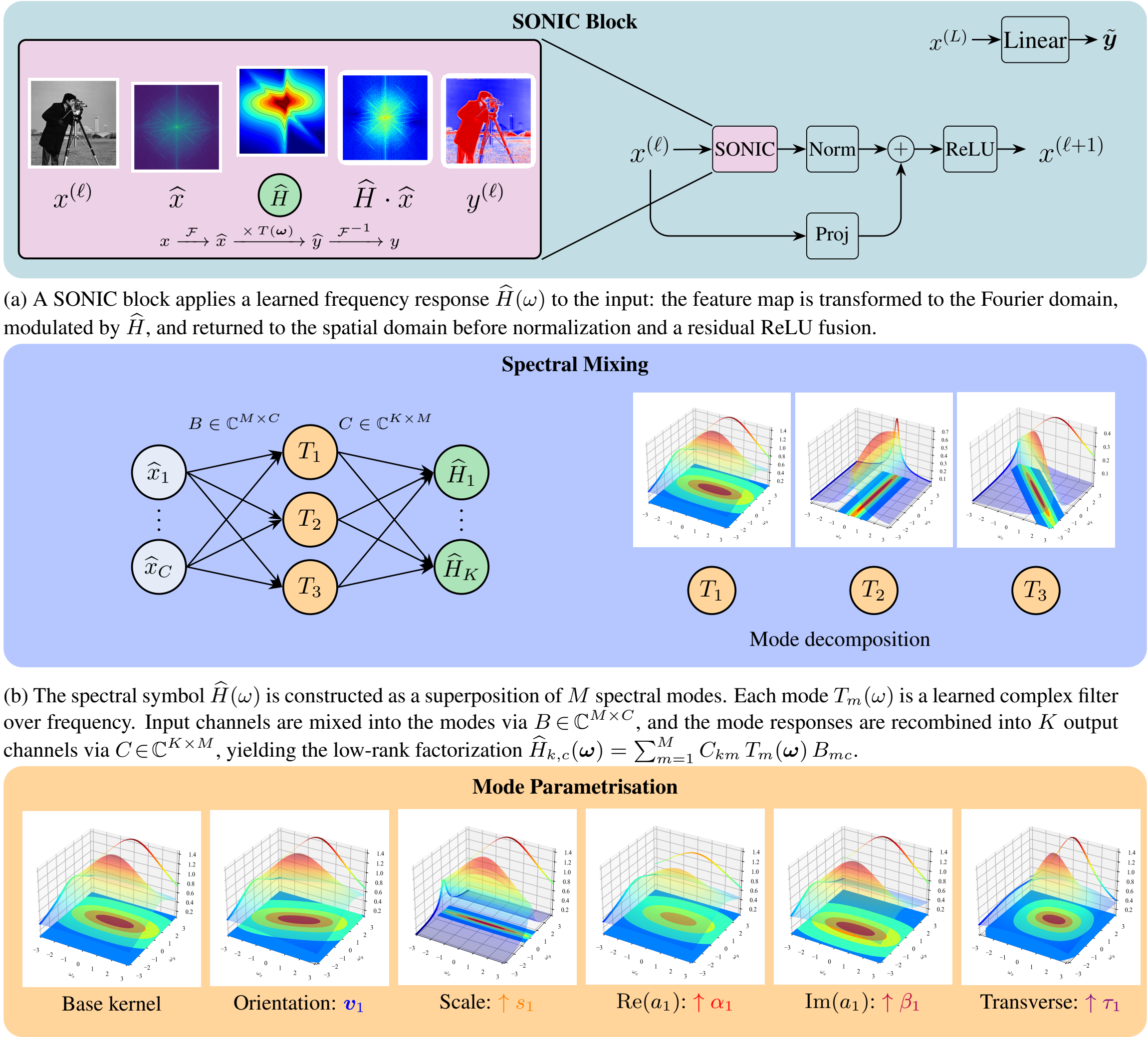
$$\omega_{||m} := \boldsymbol{\omega}\cdot\boldsymbol{v_m} \quad , \qquad \boldsymbol{\omega}_{\perp m} := (I - \boldsymbol{v_m}\boldsymbol{v_m}^{\top})\boldsymbol{\omega} \quad .$$

The mode passes slow variation along its needle and increasingly damps faster oscillations in that direction, so gently varying, needle-aligned content is emphasized while rapidly oscillating content along the axis is attenuated. It also suppresses energy that lies across the needle, so components that are not aligned with the needle's orientation contribute less. In spatial terms, the resulting kernel is stretched along v_m (making it sensitive to lines, flows, or ridges in that direction) and compressed across it.

The scale parameter s_m regulates the mode's spectral selectivity. Small values produce a broad response that pools over a wide band of along-axis frequencies, acting as an orientation-aware smoother that preserves coarse structure while suppressing fine fluctuations. Large values narrow the passband and sharpen selectivity, emphasizing only a thin slice of along-axis variation; in the spatial domain, this corresponds to a longer, more finely structured kernel along v_m. During learning, s_m adapts locally to the content of the signal: scenes dominated by broad shapes tend to drive s_m down, while scenes rich in fine oriented detail push it up.

By contrast, the complex coefficient a_m governs the global dynamics of each mode. Its real part controls damping, ensuring stability, while its imaginary part introduces oscillations that can be amplified or suppressed. These oscillations enrich the representation, allowing the mode to capture structured patterns in the plane. Unlike s_m, which tunes frequency selectivity along the axis, a_m balances between smoothness and oscillatory structure: smoother, slowly varying signals encourage stronger damping and broader low-pass behavior, whereas signals with repetitive, oriented fine-scale structure favor a smaller imaginary component that preserves such fine patterns.

Finally, the transverse penalty $\tau_m \geq 0$ pushes down frequencies that point away from v_m. This sharpens directional selectivity by suppressing leakage into neighboring directions and, in higher dimensions, prevents degenerate, plane-like responses. Intuitively, larger τ_m clamps the response tightly around the chosen axis, whereas smaller τ_m allows more lateral spread.



(a) A SONIC block applies a learned frequency response $\widehat{H}(\omega)$ to the input: the feature map is transformed to the Fourier domain, modulated by $\widehat{H}$, and returned to the spatial domain before normalization and a residual ReLU fusion.

(b) The spectral symbol $\widehat{H}(\omega)$ is constructed as a superposition of M spectral modes. Each mode $T_m(\omega)$ is a learned complex filter over frequency. Input channels are mixed into the modes via $B \in \mathbb{C}^{M \times C}$, and the mode responses are recombined into K output channels via $C \in \mathbb{C}^{K \times M}$, yielding the low-rank factorization $\widehat{H}_{k,c}(\boldsymbol{\omega}) = \sum_{m=1}^{M} C_{km}\, T_m(\boldsymbol{\omega})\, B_{mc}$.

(c) Each transfer function $T_m(\omega)$ is parameterized by interpretable geometric factors—orientation, scale, complex coefficients, and transverse decay—producing a structured family of spectral filters. Shown: parameter sweep for mode T_1, visualized as $Z = \log(1 + |T(\omega_x, \omega_y)|)$.

Figure 2: SONIC overview: (a) Residual Block, (b) Spectral Mixing, and (c) Learnable Spectral Modes.

Conceptually, the modes, after the spectral transfer, form a small dictionary of directional behaviors, while separate learned mixing weights decide how each input channel contributes to, and each output channel draws from, the same dictionary. This keeps parameters modest and encourages reuse of structure across channels. After building the modes, we let the model mix the different modes by C and B, this ensures that each channel c mapping to output k can be a unique superposition of all constructed modes. The key distinction compared to other spectral methods is the parameterisation of the spectral domain. Conventional spectral neural operators employ an unconstrained, discrete representation, assigning independent complex coefficients to each sampled frequency ω_k. In contrast, SONIC utilises a structured low-rank factorisation built from a small set of shared spectral modes. Each mode is governed by a smooth, orientation-sensitive transfer function $T_m(\omega)$, yielding a continuous and anisotropic dependence on the frequency variable. This induces substantial parameter sharing across both frequencies and channels, in contrast to traditional spectral approaches, whose representations are frequency-wise independent and lack functional coherence in ω.

Resolution Invariance Crucially, all of these filters are parameterized directly in the continuous spectral domain. This means their definition does not depend on the size or sampling rate of the image: defining filters as continuous functions of $\boldsymbol{\omega}$ decouples them from any particular grid size or sampling rate; the same response formula is evaluated on whatever DFT grid the data induces, yielding a resolution-invariant filter. This distinguishes our approach from spatial-domain kernels, whose size and shape are tied to a fixed grid. We made some minor adjustments to ensure resolution invariance: To make the directional parameters resolution invariant, we express directions in physical units and normalise:

$$D_\Delta \;=\; \operatorname{diag}(\Delta_1,\ldots,\Delta_D), \qquad \tilde{\boldsymbol{v}}_{\boldsymbol{m}} \;=\; D_\Delta^{-1}\boldsymbol{v}_{\boldsymbol{m}}, \qquad \hat{v}_m \;=\; \frac{\tilde{\boldsymbol{v}}_{\boldsymbol{m}}}{\|\tilde{\boldsymbol{v}}_{\boldsymbol{m}}\|_2}\,. \tag{16}$$

This resolution-aware formulation can be exploited during training, as also proposed in Nguyen et al. (2022). Beyond efficiency, it is particularly relevant in domains where resolution dependence is intrinsic, such as medical imaging, remote sensing, and microscopy.

Computation The number of learnable real scalars is:

$$\underbrace{2KM}_{C^{\text{re}},C^{im}} \;+\; \underbrace{2MC}_{B^{re},B^{im}} \;+\; \underbrace{(4+D)M+1}_{a^{\text{re}},\,a^{\text{im}},\,s,\,v,\tau\;\in\mathbb{R}^2},$$

For the FFT transformation we used the highly optimized VkFFT library (Tolmachev, 2023), with per-transform cost $O(N\log N)$ for a single (complex) channel. The spectral forward pass performs one DFT per input channel and one inverse DFT per output channel, plus $O(M(C{+}K))$ complex multiplications per frequency. The forward pass consists of one DFT per input channel and one inverse DFT per output channel, s with cost

$$O(CN\log N) \quad \text{and} \quad O(KN\log N),$$

where $N=\prod_{d=1}^{D}N_d$ is the total number of spatial points. In addition, frequency-wise multiplications incur a cost of

$$O\big(M(C{+}K)N\big),$$

since each of the M modes couples inputs and outputs across all frequencies. The total complexity is therefore

$$O\big((C{+}K)N\log N \;+\; M(C{+}K)N\big).$$

SONIC is thus particularly attractive for large receptive fields (where d is large or even global), since the cost remains manageable and the parameter count remains compact.

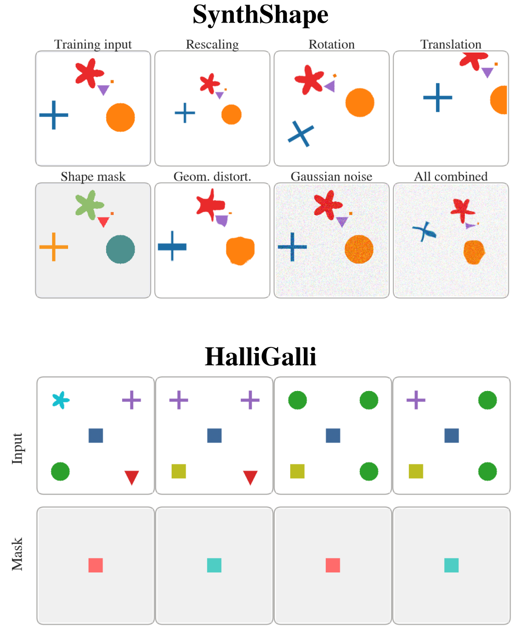
4 Empirical validation

SynthShape To evaluate the sensitivity of models to geometric variations, we introduce SynthShape (Synthetic Shape Dataset), a simple 64x64 synthetic geometric shape–based segmentation benchmark. We assess model generalization by measuring performance under controlled perturbations, Rescaling (including interpolation artefacts), in-plane rotation, out-of-frame translation, geometric distortion, and additive Gaussian noise. All experiments are conducted using 5-fold cross-validation. Furthermore, we introduce **HalliGalli**, a controlled spatial-reasoning task modeled after the well-known game, designed to test effective long-range dependency modelling rather than theoretical receptive field size. The task is to classify a central patch according to whether exactly two matching shapes appear in the four distant corners; the centre itself carries no class signal. Since the task depends on structure that cannot be captured within any local receptive field, purely local models fail. Architectures with untargeted global filters either fail due to missing orientation or degrade under Gaussian noise, as their large receptive fields accumulate noise over a broad spatial region. SONIC successfully solves the HalliGalli task and remains robust under inference-time noise, demonstrating the effectiveness of its globally oriented structured receptive field.Further implementation details are provided in Appendix 7.

Table 1: Comparison of ConvNet, ViT, S4ND, NIFF, GFNet, and SonicNet performance on SynthShape under geometric variations (left), and qualitative examples from SynthShape and HalliGalli-SRT (right).

Experiment	Value	ConvNet	ViT	S4ND	NIFF	GFNet	SonicNet
Parameter count (M)		0.153	0.468	0.186	0.042	0.415	0.072
GMACs		0.156	0.012	0.023	0.041	0.139	0.006
Distortion	2.0	0.97	0.88	0.84	0.94	0.66	0.97
	4.0	0.96	0.88	0.83	0.94	0.62	0.96
	6.0	0.94	0.87	0.82	0.92	0.60	0.96
Gaussian Noise (σ)	0.1	0.98	0.78	0.89	0.99	0.73	0.98
	0.2	0.85	0.44	0.67	0.92	0.31	0.79
	0.3	0.58	0.32	0.43	0.71	0.19	0.60
Rescaling	0.75	0.84	0.73	0.49	0.78	0.44	0.86
	1.00*	0.99	0.94	0.93	1.00	0.92	1.00
	1.50	0.62	0.68	0.32	0.59	0.37	0.74
Rotation (°)	15	0.69	0.66	0.68	0.70	0.44	0.75
	30	0.28	0.32	0.50	0.30	0.30	0.23
	45	0.28	0.30	0.44	0.29	0.28	0.24
Translation (%)	10	0.92	0.87	0.76	0.89	0.53	0.95
	20	0.96	0.90	0.76	0.96	0.74	0.97
	30	0.91	0.88	0.77	0.88	0.56	0.93
Combined	10	0.85	0.76	0.47	0.71	0.36	0.90
	20	0.62	0.59	0.30	0.50	0.28	0.76
	30	0.41	0.37	0.24	0.40	0.25	0.48
HalliGalli		0.42	0.33	0.62	1.00	0.71	1.00
HalliGalli ($\sigma = 0.1$)		0.33	0.33	0.49	0.56	0.37	0.86

* Validation accuracy on the training task.



3D Medical Image Segmentation To evaluate performance on real-world high-dimensional data requiring long-range spatial understanding, we apply our method to 3D medical image segmentation. Following the evaluation protocol of Isensee et al. (2024), we benchmark on the two datasets identified as most reliable for fair comparisons, namely Kidney and Kidney Tumour Segmentation (*KiTS*), and Automated Cardiac Diagnosis Challenge (*ACDC*). All models are trained and evaluated with identical 5-fold splits, preprocessing/target spacing, augmentations, and postprocessing. Training is conducted under identical conditions, including the same preprocessing and postprocessing steps, allowing observed differences to be attributed solely to the proposed method.

Table 2: **3D Medical Image segmentation results** (5-fold CV; mean across 5 folds). Columns report **DSC** and **NSD** (at 2 mm). "RT" is runtime (hours) and "VRAM" is peak memory (GB). Literature results are shown in gray as reported by Isensee et al. (2024).

	KiTS		**ACDC**				
Method	DSC	NSD	DSC	NSD	**Params (M)**	**RT (h)**	**VRAM (GB)**
nnU-Net ResEnc L	88.98	85.74	91.40	96.21	31.12	34	23
SonicNet (Ours)	88.55	81.19	92.02	96.07	2.59	67	61.37
nnU-Net ResEnc L	88.17	85.93	91.69	95.11	31.12	36	36.60
MedNeXt L k5	87.74	85.67	92.62	96.09	55.00	233	18.00
STU-Net L	85.84	83.02	89.34	95.12	440.30	51	26.50
SwinUNETRV2	84.14	80.11	86.24	95.15	72.80	15	13.40
nnFormer	75.85	69.43	81.55	95.83	150.0	8	5.70
CoTr	84.59	80.92	88.02	93.74	41.27	18	8.20
U-Mamba Bot	86.22	83.27	89.13	95.40	64.00	24	12.40

External validation External validation is critical in medical imaging because models frequently degrade when exposed to new scanners or protocols. Using identical training conditions and evaluating on heterogeneous external datasets provides a clinically meaningful measure of generalisability and highlights whether an architecture is suited for deployment beyond the development cohort. For this generalisability experiment, we evaluate SONIC on the

PI-CAI challenge (clinically significant prostate cancer segmentation) data (Saha et al., 2022) and compare it to the top-performing baseline, nnU-Net, on their performance on two external datasets, Prostate158 (Adams et al., 2022) and PROMIS (Ahmed et al., 2017). Qualitive comparison can be found in the appendix.

Table 3: **External validation performance on Prostate158 and PROMIS.** SonicNet achieves improved detection performance with substantially fewer parameters.

	Metric	**nnU-Net**	**SonicNet**
	TRAINABLE PARAMETERS (M/MB)	31.20/342.0	2.59/28.4
Prostate158	AUROC	0.814	**0.841**
	AP	0.533	**0.548**
	F1 Score	0.632	**0.649**
	Sensitivity	0.475	**0.495**
	Precision	0.941	**0.943**
	TP/FP/FN (%)	0.30/0.02/0.34	0.32/0.02/0.32
PROMIS	AUROC	0.646	**0.687**
	AP	0.195	**0.258**
	F1 Score	0.185	**0.223**
	Sensitivity	0.103	**0.127**
	Precision	**0.912**	0.907
	TP/FP/FN (%)	0.05/0.01/0.47	0.07/0.01/0.47

ImageNet-50M To evaluate SONIC on natural images, under highly anisotropic visual conditions, we conduct experiments on ImageNet-1K, the standard benchmark for large-scale image classification. Due to computational constraints, we adopt a reduced training setting that we denote corresponding to 200k optimization steps with a batch size of 256. We evaluate ResNet-50 variants augmented with different spectral operators and compare them against strong baselines, including a Vision Transformer. Beyond reporting standard classification accuracy, we also assess robustness under controlled resolution shifts, which serve as a proxy for the anisotropic distortions common in practical deployment scenarios. By systematically varying input resolution, we quantify how well SONIC maintains accuracy relative to competing methods, thereby characterizing its robustness to scale changes and sampling artifacts.

Table 4: **Comparison of ResNet-50 variants** and related architectures on ImageNet under 224 × 224 evaluation.

	Model Complexity			**Accuracy**	
Model	Params (M)	GFLOPs (G)	Img/s	**Top-1**	**Top-5**
ResNet-50	25.60	8.26	4758	58.47	82.68
ViT-S/16	48.60	35.21	1136	62.23	83.91
ResNet-50 NIFF	18.61	14.89	862	57.52	82.24
ResNet-50 S4ND	16.67	4.57	1421	64.38	86.44
ResNet-50 GFNet	15.72	4.57	4504	61.43	84.47
ResNet-50 RepLK	19.23	7.71	1884	65.17	86.34
ResNet-50 Dilated	25.55	38.36	2130	61.52	84.73
ResNet-50 Sonic	1.34	0.81	831	60.01	82.28

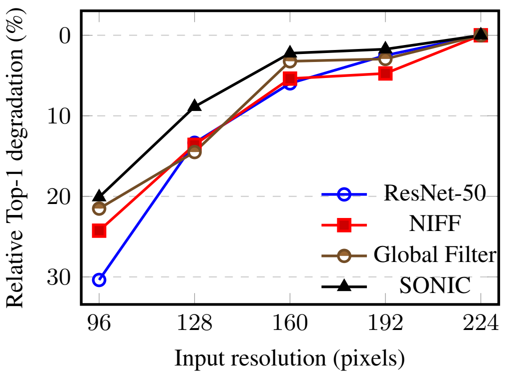
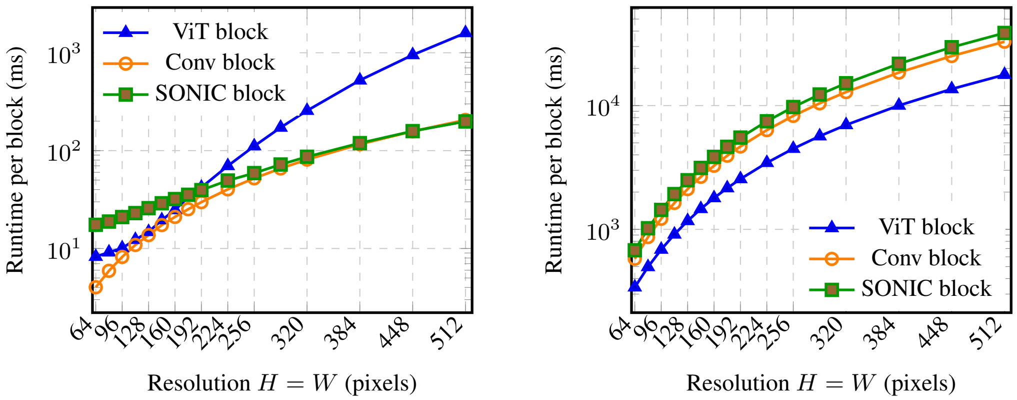


Figure 3: Relative performance degradation under resolution changes on ImageNet.

Compute and memory overhead. Figure 4 illustrates the compute and memory profile of SONIC in comparison to a standard 3×3 convolution and a ViT block with 4×4 patches. At scale (224 × 224), SONIC is only 1.23× slower and uses 1.18× more memory, representing a modest overhead for obtaining global receptive fields. At higher resolutions, the runtime gap narrows and SONIC becomes effectively on par with convolution, reflecting the favourable scaling of FFT-based filtering. SONIC's peak memory is dominated by the FFT stage, which requires storing the full spectral grid at each layer. In contrast, full-resolution self-attention grows quadratically with spatial size, becoming substantially more expensive even at moderate resolutions. Furthermore, as shown in Fig 8 (appendix), runtime and memory grow approximately linearly in both C and M in the practically relevant regime, with no unexpected spikes. This confirms that SONIC can be tuned in the number of channels and number of modes without additional overheads.

Overall, SONIC provides global spatial mixing at a fraction of the cost of global attention, while remaining close to convolution in both compute and memory.



(a) Runtime per block (log scale)

(b) Peak GPU memory (log scale)

Figure 4: Runtime and memory characteristics of ViT, convolutional, and SONIC blocks across spatial resolutions.

5 Discussion

We introduced a spectral factorisation framework, where SONIC serves as a theoretically grounded alternative to spatial convolution blocks. Unlike conventional spatial kernels, SONIC employs low-rank, orientation-aware operators in the frequency domain. This design provides a principled inductive bias for modelling long-range, structured interactions while remaining highly parameter-efficient. Our empirical evaluation demonstrates SONIC's properties. On SynthShape, the model exhibited superior robustness to image distortions compared to conventional CNN and ViT baselines and previous spectral-domain architectures. In the HalliGalli spatial reasoning task, SONIC was the only architecture capable of solving strict long-range dependencies within a single block, highlighting the effectiveness of its global receptive field. Furthermore, these theoretical advantages translated into real-world performance in 3D medical segmentation benchmarks (KiTS and ACDC), where SONIC matched or exceeded state-of-the-art performance while requiring significantly fewer parameters ($< 10\%$) than established heavyweights such as nnU-Net and MedNeXt.

At the same time, important limitations remain. Nonlinearities must be applied in the spatial domain. This prevents us from stacking multiple SONIC layers purely in the frequency domain and forces repeated FFT/IFFT operations, which introduce additional overhead. Although this limitation is shared by most spectral neural architectures, it does constrain how fully the model can operate within the spectral domain. Furthermore, we observed occasional instabilities during SONIC block initialisation, stemming from the same property that defines the operator: in imaging tasks, identical spatial dimensions may correspond to very different physical scales across datasets. Developing a more general and robust initialisation scheme for SONIC, therefore, remains an important direction for future work. Moreover, the global nature of the frequency-domain representation can limit the capture of very fine local structure, which suggests that hybrid architectures may ultimately be needed to combine the strengths of both domains. Our goal here is to provide SONIC as a general and simple operator that can be integrated in the same way as other well-known alternatives. Further work should explore how to incorporate SONIC thoughtfully into new or existing architectures. In summary, spectral factorisation offers a new building block for neural architectures that complements existing paradigms. Its strengths lie in long-range receptive field, parameter efficiency, orientation-awareness, and robustness, while future work should focus on improving efficiency, mitigating memory demands, and exploring hybrid spectral-spatial architectures.

6 Acknowledgement

In this paper, we used large language models to refine wording and improve the clarity of information transfer. All conceptual ideas, discussion of related work, and factual content were developed manually by the authors; the models were employed solely for assistance with presentation.

References

Lisa C. Adams, Marcus R. Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M. Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K. Bressem. Prostate158 - an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148:105817, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105817. URL https://www.sciencedirect.com/science/article/pii/S0010482522005789.

Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, Alex P Kirkham, Robert Oldroyd, Chris Parker, and Mark Emberton. Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822, 2017. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(16)32401-1. URL https://www.sciencedirect.com/science/article/pii/S0140673616324011.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *CoRR*, abs/1805.12177, 2018. URL http://arxiv.org/abs/1805.12177.

Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. a^2-nets: Double attention networks, 2018. URL https://arxiv.org/abs/1810.11579.

Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns, 2022. URL https://arxiv.org/abs/2203.06717.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

V. Fanaskov and I. Oseledets. Spectral neural operators, 2024. URL https://arxiv.org/abs/2205.10573.

Julia Grabinski, Janis Keuper, and Margret Keuper. As large as it gets: Learning infinitely large filters via neural implicit functions in the fourier domain, 2024. URL https://arxiv.org/abs/2307.10001.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation, 2024. URL https://arxiv.org/abs/2404.09556.

Samira Kabri, Tim Roith, Daniel Tenbrinck, and Martin Burger. Resolution-invariant image classification based on fourier neural operators, 2023. URL https://arxiv.org/abs/2304.01227.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2021. URL https://arxiv.org/abs/2010.08895.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. *CoRR*, abs/1701.04128, 2017. URL http://arxiv.org/abs/1701.04128.

Eric Nguyen, Karan Goel, Albert Gu, Gordon W. Downs, Preey Shah, Tri Dao, Stephen A. Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals using state spaces, 2022. URL https://arxiv.org/abs/2210.06583.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification, 2021. URL https://arxiv.org/abs/2107.00645.

Oren Rippel, Jasper Snoek, and Ryan P. Adams. Spectral representations for convolutional neural networks, 2015. URL https://arxiv.org/abs/1506.03767.

Anindo Saha, Jasper Jonathan Twilt, Joeran Sander Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, and Henkjan Huisman. The pi-cai challenge: Public training and development dataset, June 2022. URL `https://doi.org/10.5281/zenodo.6624726`.

Dmitrii Tolmachev. Vkfft-a performant, cross-platform and open-source gpu fft library. *IEEE Access*, 11:12039–12058, 2023. doi: 10.1109/ACCESS.2023.3242240.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2018. URL `https://arxiv.org/abs/1711.07971`.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

7 APPENDIX

APPENDIX A: IMPLEMENTATION DETAILS

We constrain $s_m > 0$ and typically enforce $\mathrm{Re}(\mathrm{a_m}) < 0$) so that the spatial response function decays rather than grows. The imaginary part $\mathrm{Im}(\mathrm{a_m})$ can be bounded in magnitude (e.g., $|a_m^{\mathrm{im}}| \leq \rho$). We initialize $v_m \sim U(0, \pi)$.

All parameters are learned end-to-end by backpropagation. A convenient reparameterisation that enforces the constraints is:

$$s_m = \mathrm{softplus}(\sigma_m) + \varepsilon, \qquad a_m^{\mathrm{re}} = -\,\mathrm{softplus}(\alpha_m), \qquad a_m^{\mathrm{im}} = \rho\,\tanh(\beta_m), \qquad v_m = \frac{u_m}{\|u_m\|_2},$$

with free variables $\sigma_m, \alpha_m, \beta_m, \rho \in \mathbb{R}$ and $u_m \in \mathbb{R}^2$, small $\varepsilon > 0$. The mixing matrices B and C are complex-valued and learned without constraints.

Implementation notes. (i) We standardize each input channel to zero mean and unit variance, with a small noise for numerical stability. (ii) We apply an RMS transfer gain normalisation over the (half-)spectrum to keep the overall response well-scaled across resolutions.(iii) We use real–FFT (rFFT/irFFT) along the last two spatial dimensions; consequently we enforce Hermitian consistency by zeroing the imaginary part at DC.(iii) For memory efficiency the computation is performed in frequency *slabs* (blocks over rows of Ω) without altering the continuous formulation above. (iv) Direction vectors are rescaled by D_{Δ}^{-1} and renormalized (unit length) before use, ensuring invariance to pixel spacing. (v) Optional mode dropout is applied to V_m as a regularizer.

SYNTHSHAPE

The dataset consist of a random number of geometric primitives (circle, square, triangle, cross, star) at random positions and scales within the image, while preventing overlaps through collision checks. Each object is assigned a randomly perturbed base colour, ensuring that models cannot exploit a trivial mapping between RGB values and semantic classes. The ground-truth segmentation mask assigns a unique class label to each shape type, with background indexed as class 0.

Models. All models use an embedding width of c=128

- **ConvNet:** A lightweight stack of L convolutional layers (default $L = 4$), each followed by group normalisation and GELU activations. A 1×1 convolution projects the final feature map to the number of classes. The patch size is set to 16 to give the model a fair opportunity to capture broader context, rather than learning solely from small local receptive fields.
- **ViT:** A Vision Transformer consisting of a patch embedding layer, sinusoidal positional encodings (interpolated if image resolution differs), and a stack of transformer blocks with multi-head self-attention and MLP layers. The output features are reshaped and upsampled to the original spatial resolution, followed by a 1×1 convolution for classification.
- **SonicNet:** For SonicNet we use a depth of 4 stacked SonicBlocks, each consisting of GroupNorm, GELU, and a residual spectral convolutional mapping. The final stage applies GroupNorm, GELU, and a 3×3 convolution to project features to class logits.
- **GFNet:** Each block replaces local convolutions by a learned complex-valued mask applied in the Fourier domain. Features are normalised and transformed by the global filter, followed by a pointwise MLP and residual connections, while the overall encoder–head structure is kept identical to the ConvNet.
- **NIFF:** Rach block learns a continuous frequency response via a small MLP that maps frequency coordinates to complex filter values. These filters are applied depthwise in the Fourier domain and wrapped in the same normalisation, residual, and head structure as the ConvNet.
- **S4ND:** A state-space baseline where the convolutional backbone is replaced by stacked S4ND layers operating directly on the $H \times W$ grid. Each block applies a 2D structured state-space update to the feature map and is embedded in the same residual and segmentation head pattern as the other models.

Training. All models were trained using the AdamW optimizer with learning rate 10^{-2} and weight decay 10^{-4}, for 1000 epochs and batch size 32. A one-cycle learning rate schedule was applied. To account for class imbalance, inverse-frequency class weights were computed dynamically from a large synthetic batch and used in the cross-entropy loss. The final training objective combined cross-entropy with the multi-class Dice loss in equal weighting.

Evaluation. Model robustness was assessed by applying five geometric transformations at inference: rescaling, rotation, translation, distortion, and Gaussian noise. Each transformation was applied with three levels of severity. Rescaling resized the full image before resampling it back to 64×64, introducing interpolation artefacts. Translation shifted the input by a fixed percentage of image width/height, potentially moving parts of objects out of frame. Distortion was implemented via bicubic upsampling of a low-resolution displacement field. Rotation was performed around the image centre, and Gaussian noise was added per pixel channel.

Metrics. The primary evaluation metric was the multi-class Dice score (excluding background), averaged across folds. All experiments were repeated 5 times with different seeds to estimate variance. 1.

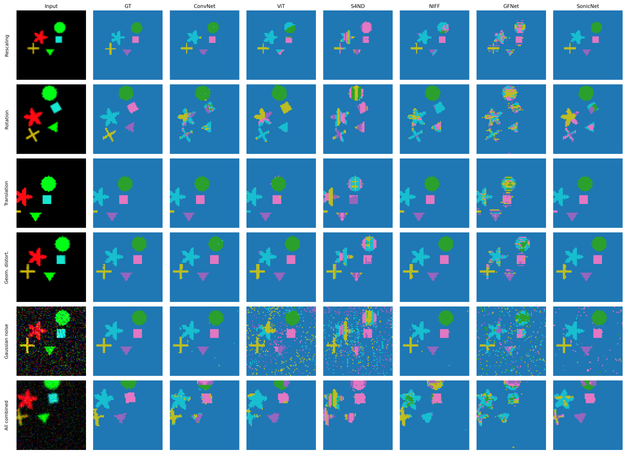


Figure 5: Qualitative visual examples on the SynthShape benchmark

ImageNet

For ImageNet we follow the public "academic default" implementation[1] and keep all training hyperparameters and optimisation settings unchanged. We replace the standard ResNet-50 bottleneck blocks by SonicBlocks, where each 3×3 convolution in the main path is substituted with a Sonic layer, while the 1×1 convolutions in the skip path and classifier head remain unchanged. Full architectural details and the exact PyTorch implementation of `resnet50_sonic` are provided in the supplementary material.

Medical Imaging Benchmark

Setup. Following the recommendations of Isensee et al. (2024), we minimize confounding factors and keep the experiment as plain as possible. We retain the baseline nnU-Net preprocessing and postprocessing and change only the network backbone: the original U-Net is replaced by a stack of SONIC Blocks ("SonicNet"). The first block lifts the input from C to K channels; the remaining $D-1$ blocks keep K channels. We apply GroupNorm and GELU before a final 3×3 convolution to produce n_{classes} output channels. For this experiment, we used four stacked SonicBlocks (i.e., a depth of 4). We employed stochastic gradient descent with an initial learning rate of 10^{-2} and a weight decay of 10^{-5}. Training was performed with a mini-batch size of two for a total of 1000 epochs, each consisting of 250 iterations. For inference, we used the checkpoint corresponding to the highest validation performance during training.

[1] `https://github.com/landskape-ai/imagenet`

7.1 Qualitative comparison of the external validation

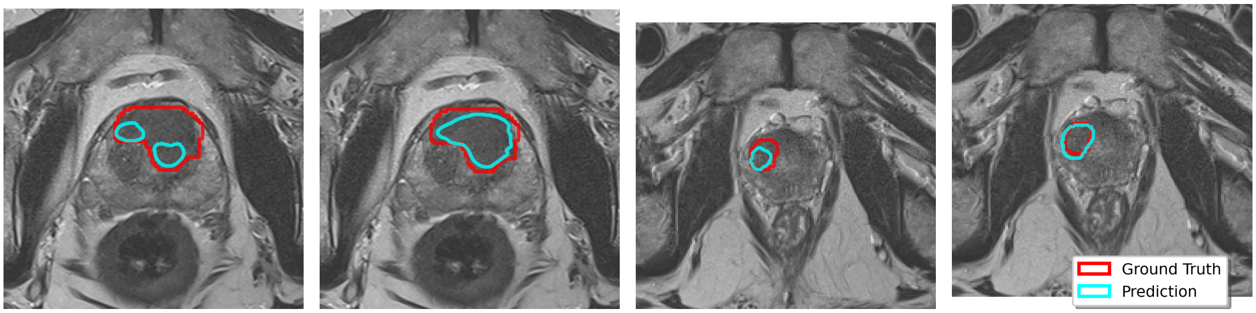


Figure 6: **Qualitative comparison of prostate cancer detection methods.** The figure shows representative cases from the Prostate158 (left) and PROMIS (right) datasets, with ground truth lesions (red) and model predictions (cyan) overlaid on T2-weighted MRI slices (confidence ≥ 0.5).

8 Appendix B: Practical implementation

Role of K and M The parameters K and M play complementary roles in shaping the behaviour of a SONIC block. The number of modes M determines the spectral diversity of the operator; in contrast, the channel width K controls the capacity with which these shared modes are mixed across feature channels. The ratio between K and M therefore reflects the balance between channel-mixing capacity and spectral richness. Understanding this trade-off helps guide architectural choices across different model sizes.

Qualitative analysis of receptive fields To better understand how SONIC behaves in practice, we visualize the normalized spectral energy of the learned filters across the four stages of the network. Each plot shows the log-scaled energy distribution over the spatial frequency plane, giving an intuitive sense of the effective receptive field and directional structure captured at different depths.

As the network progresses through stages, the spectral responses become increasingly smooth, structured, and oriented—indicating that early stages capture broad, irregular frequency content, while deeper stages refine this into cleaner, more coherent spectral patterns. These visualizations highlight how SONIC gradually organizes its spectral modes and how the parameterization remains stable and well-behaved across depth.

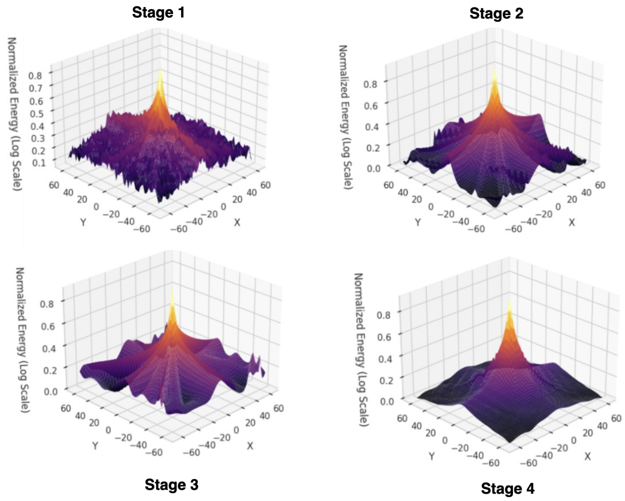


Figure 7: Visualisation of a randomly selected learned 2D convolutional kernel from our medical image segmentation model across four stages of the network.

Practical scalability of SONIC Block To validate that the SONIC block exhibits the intended linear scaling behavior, we empirically benchmark its runtime and memory usage across a range of channel dimensions and mode counts.We confirm this behavior by measuring wall-clock runtime and peak memory under controlled synthetic settings, sweeping C and M independently while keeping spatial resolution fixed. Across all tested configurations, runtime increases as a straight line with respect to both variables, and memory usage follows the same linear trend.

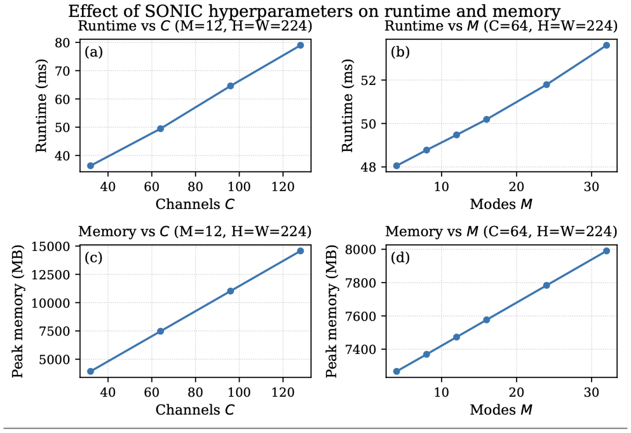


Figure 8: Runtime and memory of SONIC when varying channels C and modes M at fixed resolution.

Appendix C: Supporting Proofs

Convolution Theorem for the D-dimensional Fourier Transform

Let the convolution of two functions on $\mathbb{R}^D$ be defined by

$$(f * g)(\mathbf{x}) := \int_{\mathbb{R}^D} f(\boldsymbol{\tau})\, g(\mathbf{x} - \boldsymbol{\tau})\, d\boldsymbol{\tau}, \qquad \mathbf{x} \in \mathbb{R}^D.$$

Then, for $\boldsymbol{\omega} \in \mathbb{R}^D$, the D-dimensional Fourier transform satisfies

$$\begin{aligned}
\mathcal{F}\{f * g\}(\boldsymbol{\omega}) &= \int_{\mathbb{R}^D} \left(\int_{\mathbb{R}^D} f(\boldsymbol{\tau})\, g(\mathbf{x} - \boldsymbol{\tau})\, d\boldsymbol{\tau} \right) e^{-i\boldsymbol{\omega}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} f(\boldsymbol{\tau})\, g(\mathbf{u})\, e^{-i\boldsymbol{\omega}\cdot(\boldsymbol{\tau}+\mathbf{u})}\, d\mathbf{u}\, d\boldsymbol{\tau} \\
&= \left(\int_{\mathbb{R}^D} f(\boldsymbol{\tau})\, e^{-i\boldsymbol{\omega}\cdot\boldsymbol{\tau}}\, d\boldsymbol{\tau} \right) \left(\int_{\mathbb{R}^D} g(\mathbf{u})\, e^{-i\boldsymbol{\omega}\cdot\mathbf{u}}\, d\mathbf{u} \right).
\end{aligned}$$

Hence,

$$\mathcal{F}\{f * g\}(\boldsymbol{\omega}) = \mathcal{F}\{f\}(\boldsymbol{\omega})\, \mathcal{F}\{g\}(\boldsymbol{\omega}).$$

Connection to State-Space Kernels

Consider the linear time-invariant state-space model

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t), \tag{17}$$

with $x(t) \in \mathbb{C}^n$, $u(t) \in \mathbb{C}^m$, $y(t) \in \mathbb{C}^p$, and system matrices $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Assume a zero initial state $x(0^-) = 0$ and a strictly proper output.

The corresponding impulse response (or kernel) is

$$\mathcal{K}(t) = Ce^{At}B, \qquad t \geq 0. \tag{18}$$

By definition, the transfer function is the Laplace transform of the impulse response:

$$H(s) = \int_0^\infty e^{-st}\,\mathcal{K}(t)\,dt = \int_0^\infty e^{-st}\,Ce^{At}B\,dt. \tag{19}$$

Pulling out C and B gives

$$H(s) = C\left(\int_0^\infty e^{(A-sI)t}\,dt\right)B. \tag{20}$$

For $\mathrm{Re}(s)$ sufficiently large, the integral converges to

$$\int_0^\infty e^{(A-sI)t}\,dt = (sI - A)^{-1}. \tag{21}$$

Hence the transfer function is

$$H(s) = C(sI - A)^{-1}B. \tag{22}$$

SONIC with Restricted Modes. We show that our general Fourier domain formulation reduces to the Laplace resolvent parameterisation of S4ND when orientations are restricted to the coordinate axes.

Recall our frequency response factorisation

$$\widehat{H}_{c,k}(\boldsymbol{\omega}) = \sum_{m=1}^{M} C_{km}\,T_m(\boldsymbol{\omega})\,B_{mc}, \tag{23}$$

with mode response

$$T_m(\boldsymbol{\omega}) = \frac{1}{is_m(\boldsymbol{\omega}\cdot v_m)\ -\ a_m\ +\ \tau_m\|(I - v_m v_m^\top)\boldsymbol{\omega}\|^2}. \tag{24}$$

Suppose $v_m = e_d$, the d-th standard basis vector. Then

$$\boldsymbol{\omega}\cdot v_m = \omega_d, \qquad (I - v_m v_m^\top)\boldsymbol{\omega} = \sum_{j\neq d}\omega_j e_j,$$

so that

$$\|(I - v_m v_m^\top)\boldsymbol{\omega}\|^2 = \sum_{j\neq d}\omega_j^2.$$

In this case,

$$T_m(\boldsymbol{\omega}) = \frac{1}{is_m\,\omega_d - a_m + \tau_m\sum_{j\neq d}\omega_j^2}. \tag{25}$$

We discard the transverse penalty $\tau_m = 0$, then

$$T_m(\omega_d) = \frac{1}{is_m\omega_d - a_m} = \frac{1}{s_m}\,\frac{1}{i\omega_d - \frac{a_m}{s_m}},$$

where the absorption is into the learned parameters (a_m/s_m in A, and B or C absorb $1/s_m$). Thus

$$\widehat{H}_{c,k}(\omega_d) = \left[C\,(i\omega_d I - A)^{-1}B\right]_{kc}, \qquad H(s) = C(sI - A)^{-1}B.$$