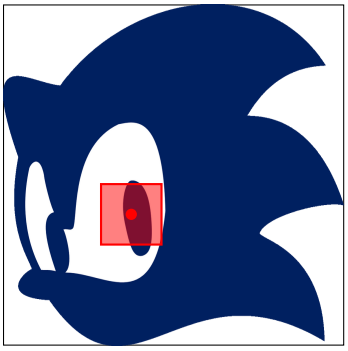
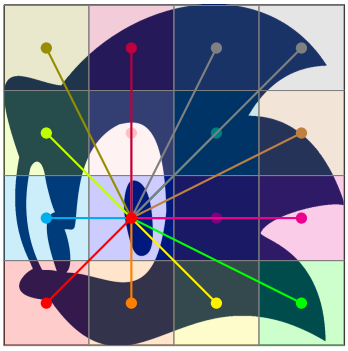
SONIC: SPECTRAL ORIENTED NEURAL INVARIANT CONVOLUTIONS

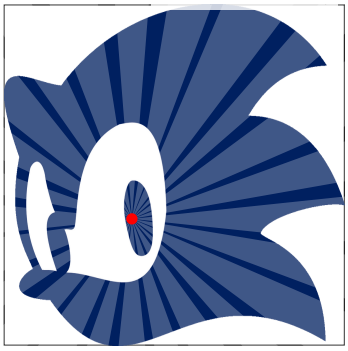
Anonymous authors
Paper under double-blind review



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 compact collection of spectral filters that learns directly in the Fourier domain. SONIC factorises multi-channel frequency responses through a small set of shared oriented components. This yields filters that are directional, interpretable, and resolution-invariant, extending globally beyond patch-size limitations. Parameters scale linearly with the number of channels, enabling efficient learning without loss of expressivity. Experiments on standard vision benchmarks show that SONIC delivers more robust performance than conventional models, while matching or exceeding their accuracy with substantially fewer parameters.

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. The Sonic model is a drop-in replacement for vision or higher-dimensional signal tasks requiring long receptive fields, robustness to quantifiable resolution variance (e.g., medical imaging) and tasks which require high directional selectivity.

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. The approach combines simplicity with generality, offering a lightweight yet versatile foundation that can support progress toward more scalable and adaptable vision models.

The remainder of this paper is organized 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

To motivate our approach, we introduce the mathematical foundations that connect convolution, linear systems, and their extensions into the spectral domain.

Linear Time-Invariant (LTI) systems Convolution is a fundamental operation in signal processing and neural networks, describing how one function modifies or filters another. A central class of convolutional systems are Linear Time-Invariant (LTI) systems, characterized by their impulse response $\mathcal{K}(t)$. For D-dimensional signals $u(t)$, the output is

$$y(\mathbf{x}) = \int_{\mathbb{R}^D} \mathcal{K}(\mathbf{x} - \boldsymbol{\tau})\, u(\boldsymbol{\tau})\, d\boldsymbol{\tau}. \tag{1}$$

In the one-dimensional temporal case, finite-dimensional LTI systems admit a state-space representation:

$$\mathcal{K}(t) = Ce^{At}B, \tag{2}$$

where $\mathcal{K}(t)$ is causal and absolutely integrable and consists of mixtures of exponentials and damped oscillations, with $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ specifying the dynamics, input, and output maps, respectively. This formulation, known as a state-space system, underlies many modern learning-based sequence models including the Kalman filter (Kalman, 1960), linear dynamical systems (Roweis & Ghahramani, 1999), and nonlinear extensions such as recurrent neural networks, LSTMs (Hochreiter & Schmidhuber, 1997), and GRUs (Chung et al., 2014). However, LTIs gained renewed attention with the introduction of the Structured State Space model (S4) by Gu et al. (2021), which emphasizes the linear ODE representation of the system:

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t). \tag{3}$$

Here $x(t) \in \mathbb{R}^N$ acts as a hidden state, encoding information from all past inputs up to time t. To make this formulation suitable for sequence-to-sequence modeling on discrete data, the continuous-time dynamics, on initialisation, are projected onto a polynomial orthogonal basis and then discretized. This yields a stable and trainable recurrence relation that can be implemented efficiently on GPUs. Mamba (Gu & Dao, 2024), the successor to S4, incorporates input-dependent selectivity for adaptive dynamics, at the cost of strict LTI structure.

While state-space models were originally designed for one-dimensional sequences, many domains, especially vision, require multidimensional inputs. Extensions have followed three main paradigms. (i) **Flatten-and-scan** (Zhu et al., 2024), (Liu et al., 2024) reshapes D-dimensional arrays into 1D sequences for standard SSMs, sometimes with multiple scan directions to reduce ordering bias, but at the cost of locality. (ii) **Tensor-product bases** (Baron et al., 2023) extend the S4 spectral parameterisation to higher dimensions by constructing kernels from tensor products of 1D basis functions. This preserves spatial alignment but remains limited to separable spectral forms. (iii) **Spatial separable multidimensional kernels**(Nguyen et al., 2022) instead imposes separability directly on the convolution kernel, modeling it as a Kronecker product of independent 1D SSMs. This enables efficient computation but constrains cross-dimensional interactions. All these methods enable efficient computation and maintain the structure of multidimensional convolutions, but strict separability limits expressiveness by constraining cross-dimensional interactions to factorized forms.

Spectral decomposition We extend multidimensional SSMs by leveraging the spectral decomposition of D-dimensional signals. The frequency analysis of such signals is formalized using the Fourier transform. We denote by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_D) \in \mathbb{R}^D$ a D-dimensional angular frequency, with each component ω_d restricted to the discrete DFT grid

$$\Omega_d \;=\; 2\pi \cdot \left\{ \frac{k_d}{N_d\,\Delta_d} \;\middle|\; k_d = -\left\lfloor \frac{N_d}{2} \right\rfloor, \ldots, \left\lceil \frac{N_d}{2} \right\rceil - 1 \right\}, \quad d = 1, \ldots, D,$$

so that the full frequency domain is $\Omega = \Omega_1 \times \cdots \times \Omega_D$. For a function $f : \mathbb{R}^D \to \mathbb{C}$, the D-dimensional Fourier transform is defined as

$$\mathcal{F}\{f\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^D} f(\mathbf{x})\, e^{-i\boldsymbol{\omega}\cdot\mathbf{x}}\, d\mathbf{x}, \quad \boldsymbol{\omega} \in \mathbb{R}^D. \tag{4}$$

A central result is the Convolution Theorem (see Appendix 7), which states that convolution in the spatial domain corresponds to multiplication in the frequency domain:

$$\mathcal{F}\{f * g\}(\boldsymbol{\omega}) = \mathcal{F}\{f\}(\boldsymbol{\omega})\, \mathcal{F}\{g\}(\boldsymbol{\omega}). \tag{5}$$

Beyond an analytical formulation, this property can be extended to kernels specified directly in the frequency domain; in particular, one may learn a spectral kernel $T(\boldsymbol{\omega})$ such that

$$\mathcal{F}\{f * g\}(\boldsymbol{\omega}) = T(\boldsymbol{\omega})\, \mathcal{F}\{g\}(\boldsymbol{\omega}), \tag{6}$$

which forms the basis of spectral neural methods.

Spectral Neural Methods A number of approaches have explored learning filters directly in the spectral domain rather than the spatial domain. While not strictly a spectral parameterisation, early work such as Rippel et al. (2015) proposed to learn the full convolution kernel in the spatial domain and then apply it efficiently in the frequency domain using the convolution theorem. More recently, Rao et al. (2021) introduced *global filter networks* (GFNet), which directly parameterize a complex-valued mask $M(\boldsymbol{\omega})$ in the Fourier domain and apply it elementwise. This provides the model with full control over the spectral response, but also introduces a limitation: the FFT grid is tied to the input resolution, and the number of learnable parameters scales with the discretisation of the frequency spectrum. An alternative line is represented by the *Fourier Neural Operator* (FNO) of Li et al. (2020), which avoids parameterizing the full frequency domain. Instead, FNO truncates to a fixed set of low-frequency fourier coefficients, typically chosen as a square block of Fourier coefficients closest to the origin, and learns weights only for those frequencies:

$$\hat{y}(\boldsymbol{\omega}) = \begin{cases} \mathcal{Q}(\boldsymbol{\omega})\, \mathcal{F}\{g\}(\boldsymbol{\omega}), & \boldsymbol{\omega} \in \Omega, \\ \mathcal{F}\{g\}(\boldsymbol{\omega}), & \boldsymbol{\omega} \notin \Omega, \end{cases} \tag{7}$$

where $\mathcal{Q}(\boldsymbol{\omega})$ are the learnable spectral coefficients. This drastically reduces the number of parameters and demonstrates improved resolution robustness of the learned operator. However, learning a separate coefficient for each retained frequency treats the spectrum as a set of disconnected points, disregarding the smooth, correlated structure across nearby frequencies and orientations. This lack of parameter sharing across ω reduces the inductive bias needed to interpolate and extrapolate, thereby limiting generalisation across resolutions and spectral regions.

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.

Formulation Consider a continuous-time LTI state-space model with zero initial condition. Its Laplace-domain transfer function is (see App. 7):

$$H(s) \;=\; \mathbf{C}\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}. \tag{8}$$

We use this resolvent form as a template for the spatial filtering; Let the input be $x \in \mathbb{R}^{n_c\times N_1\times\cdots\times N_D}$ and the output $y \in \mathbb{R}^{n_k\times N_1\times\cdots\times N_D}$, and denote their D-dimensional DFTs by $\widehat{x},\widehat{y}$:

$$\widehat{y}(\boldsymbol{\omega}) \;=\; \widehat{H}(\boldsymbol{\omega})\,\widehat{x}(\boldsymbol{\omega}), \qquad y \;=\; \mathcal{F}_D^{-1}[\widehat{y}].$$

We apply frequency-wise filtering as

$$\widehat{y}_k(\boldsymbol{\omega}) \;=\; \sum_{c=1}^{n_c} \widehat{H}_{k,c}(\boldsymbol{\omega})\;\widehat{x}_c(\boldsymbol{\omega}), \qquad k=1,\dots,n_k,\;\; \boldsymbol{\omega}\in\Omega, \tag{9}$$

where $\widehat{H}_{k,c}(\boldsymbol{\omega}) \in \mathbb{C}$ is the frequency response of the $(c\to k)$ channel filter. Rather than learning a free response for every pair (k,c) and every $\boldsymbol{\omega}$, we factorise the coupling through M shared modes reused across all channels and frequencies. In matrix form,

$$\widehat{\mathbf{H}}(\boldsymbol{\omega}) \;=\; \mathbf{C}\;\mathrm{diag}\big(T_1(\boldsymbol{\omega}),\dots,T_M(\boldsymbol{\omega})\big)\,\mathbf{B}, \tag{10}$$

such that entrywise:

$$\widehat{H}_{k,c}(\boldsymbol{\omega}) \;=\; \sum_{m=1}^{M} C_{km}\,T_m(\boldsymbol{\omega})\,B_{mc}, \qquad \mathbf{B}\in\mathbb{C}^{M\times N_c},\;\; \mathbf{C}\in\mathbb{C}^{N_k\times M}. \tag{11}$$

Central to our method is the transfer function, which defines a frequency-wise linear operator. For each mode $m=1,\dots,M$ we define:

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

Where each mode is parameterised by: (1) the orientation $\boldsymbol{v_m} \in \mathbb{R}^D$ with $\|\boldsymbol{v_m}\|_2 = 1$, which selects the direction in frequency space; (2) the scale $s_m \in \mathbb{R}_{>0}$, encoding the spectral selectivity; (3) the real part $\mathrm{Re}(a_m)$, which introduces damping; (4) the imaginary part $\mathrm{Im}(a_m)$, which governs oscillatory behaviour; and (5) the transverse penalty $\tau_m \geq 0$, which controls the decay of responses orthogonal to v_m. Together, these six parameters shape the amplitude, orientation, and oscillatory nature of the spectral transfer function.

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 1.

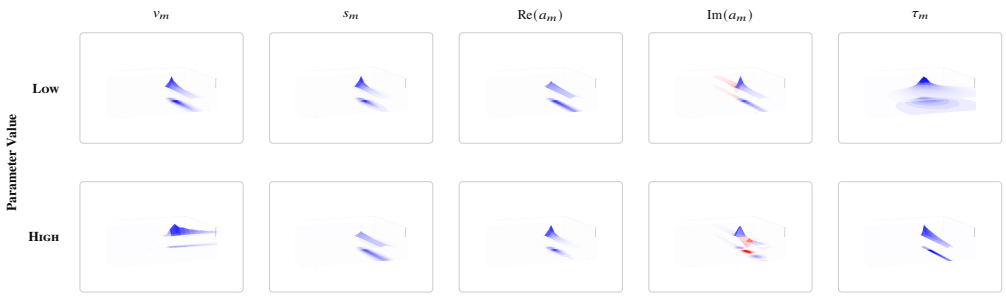


Figure 1: **Effect of parameters on the construction of the Transfer Function $T_m(\omega)$**, Columns (left→right): v, s, a_{re}, a_{im}, τ, and . Rows decodes the Parameter Value: **LOW** (top) vs. **HIGH** (bottom), with all other parameters fixed at their nominal values. Each panel shows a 3D surface of the spatial kernel $k(x, y) = \mathcal{F}^{-1}\{T_m(\omega)\}$ and a floor projection (filled contours). Heights and colors represent the kernel value (the predicted response at offset (x, y) to a unit impulse), normalized to $[-1, 1]$. For the τ column we display $|k|$; for the others we display $\mathrm{Re(k)}$. This is a qualitative illustration.

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/perpendicular 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, scaled by ρ, 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.

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. This is illustrated in figure 2.

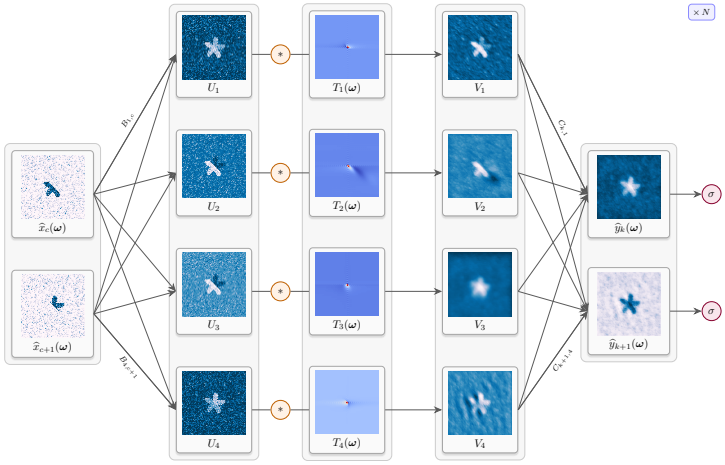


Figure 2: **Illustration of the SonicBlock($n_c = 2 \longmapsto n_k = 2,\ m = 4$)**: An overview of the frequency-wise filtering architecture, for conceptual clarity, each stage is depicted in the spatial domain, although training operates fully in the frequency domain. The block mixes the input features $\hat{x}_c(\omega)$ through coefficients B and C, producing intermediate variables U_m. These are modulated by the transfer functions $T_m(\omega)$ to yield V_m, combined to form the outputs $\hat{y}_k(\omega)$. Multiple instances of this block may be stacked, interleaved with conventional nonlinearities and normalisation operations

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 normalize:

$$D_\Delta = \mathrm{diag}(\Delta_1, \ldots, \Delta_D), \qquad \tilde{\boldsymbol{v}}_{\boldsymbol{m}} = D_\Delta^{-1} \boldsymbol{v}_{\boldsymbol{m}}, \qquad \hat{v}_m = \frac{\tilde{\boldsymbol{v}}_{\boldsymbol{m}}}{\|\tilde{\boldsymbol{v}}_{\boldsymbol{m}}\|_2}. \tag{13}$$

This formulation provides flexibility with respect to resolution, which can also be exploited during training as proposed in Nguyen et al. (2022). Beyond training efficiency, resolution-aware parameterisation is particularly relevant in domains where resolution dependence is intrinsic, such as medical imaging, remote sensing, and microscopy. An additional possibility is to relax the unit-norm constraint on the orientation modes and allow resolutions to be learned directly from data; however, this may introduce instability during optimisation.

Computation The number of learnable real scalars is:

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

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

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

For comparison, a standard spatial convolution with kernel size $d \times d$ has cost

$$O(CKNd^2).$$

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 examine the generalisation ability of these models by evaluating their performance under a range of geometric transformations, namely scaling, rotation, translation, distortion, and additive Gaussian noise. Although training involves varying object sizes and random placement on a grid, scaling and translation remain challenging. Translation can result in objects partially leaving the frame (out-of-frame translation), a scenario absent during training. Scaling is applied to the entire image, which is then resized back to the original resolution of 64×64 pixels, thereby introducing interpolation artefacts. SynthShape employs a simple network architecture with an interchangeable backbone, which can be instantiated as (i) a Convolutional Neural Network (ConvNet), (ii) a Vision Transformer (ViT), or (iii) SonicNet. All experiments are conducted using 5-fold cross-validation. Further implementation details and segmentation predictions are provided in Appendix 7.

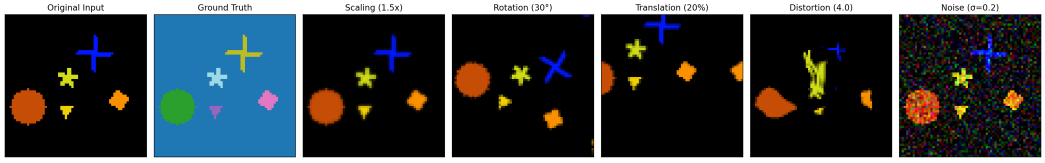


Figure 3: Representative examples of the geometric transformations applied during inference.

Table 1: Comparison of ConvNet, ViT, and SonicNet performance on SynthShape under geometric variations. Results are reported as Dice score (mean ± std) over 5-fold cross-validation.

Experiment	**Value**	**ConvNet**	**ViT**	**SonicNet**
Parameter count (M)		0.448	1.291	0.125
GMACs		1.829	0.024	0.030
Distortion	2.0	0.440 ± 0.153	0.088 ± 0.030	**0.579 ± 0.132**
	4.0	0.263 ± 0.073	0.114 ± 0.102	**0.282 ± 0.077**
	6.0	0.113 ± 0.086	0.082 ± 0.089	**0.156 ± 0.054**
Gaussian Noise (σ)	0.1	0.605 ± 0.039	0.085 ± 0.076	**0.802 ± 0.061**
	0.2	0.337 ± 0.102	0.085 ± 0.085	**0.607 ± 0.055**
	0.3	0.175 ± 0.043	0.063 ± 0.085	**0.547 ± 0.075**
Rescaling	0.75	0.688 ± 0.035	0.133 ± 0.075	**0.773 ± 0.045**
	1.00*	0.963 ± 0.073	0.131 ± 0.073	**0.991 ± 0.015**
	1.50	0.813 ± 0.044	0.136 ± 0.074	**0.815 ± 0.036**
Rotation (°)	15	0.872 ± 0.066	0.169 ± 0.111	**0.883 ± 0.026**
	30	0.797 ± 0.064	0.107 ± 0.070	**0.846 ± 0.031**
	45	0.762 ± 0.071	0.115 ± 0.067	**0.827 ± 0.064**
Translation (%)	10	0.826 ± 0.037	0.115 ± 0.080	**0.846 ± 0.029**
	20	**0.663 ± 0.113**	0.158 ± 0.066	0.651 ± 0.121
	30	0.598 ± 0.181	0.189 ± 0.115	**0.602 ± 0.159**

* Validation accuracy on the training task.

Prostate Cancer Detection For the high-stakes clinical context, we focus our study on 3D prostate imaging, given its importance in the diagnosis and management of prostate cancer. Accordingly, we train our model on the PI-CAI dataset (Saha et al., 2022), which comprises 1,500 anonymized biparametric MRI scans from 1,476 patients collected between 2012 and 2021. For external validation, we evaluate performance on the unseen Prostate158 (Adams et al., 2022) and PROMIS (Ahmed et al., 2017) datasets.The PROMIS dataset is considered challenging because its scans are older and have lower diagnostic quality, as MRI quality has markedly improved since the study Hering et al. (2024).

Table 2: Detection performance comparison on Prostate158 and PROMIS datasets (binary, threshold 0.5). Best performances are shown in **bold**.

	Metric	nnU-Net	nnDetection	SonicNet
	TRAINABLE PARAMETERS (M/MB)	31.20/342.0	31.20/188.5	0.06/1.1*
Prostate158	AUROC	0.814	0.789	**0.841**
	AP	0.533	0.442	**0.548**
	F1 Score	0.632	0.511	**0.649**
	Sensitivity	0.475	0.353	**0.495**
	Precision	0.941	0.923	**0.943**
	TP/FP/FN (%)	0.30/0.02/0.34	0.23/0.02/0.42	0.32/0.02/0.32
PROMIS	AUROC	0.646	0.593	**0.687**
	AP	0.195	0.128	**0.258**
	F1 Score	0.185	**0.288**	0.223
	Sensitivity	0.103	**0.176**	0.127
	Precision	**0.912**	0.791	0.907
	TP/FP/FN (%)	0.05/0.01/0.47	0.09/0.02/0.43	0.07/0.01/0.47

* For the full breakdown of parameters see Appendix B, Table 4.

For empirical validation, we compare our method exclusively with nnU-Net v2 and nnU-Net Detection (Isensee et al., 2021), as they currently represent the most reliable and state-of-the-art baselines for 3D medical image detection and segmentation, and were the official baselines of the PI-CAI challenge Saha et al. (2022), (Isensee et al., 2024). By restricting our comparison to nnU-Net v2 and nnU-Net Detection, we ensure that performance gains are demonstrated against the strongest and most widely recognized references in the field. Training is conducted under identical conditions, including the same preprocessing and postprocessing steps, allowing observed differences to be attributed solely to the proposed method.

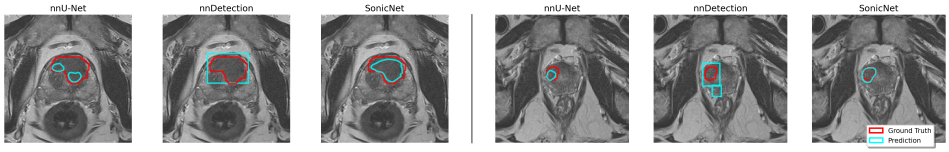


Figure 4: Qualitative comparison of prostate cancer detection methods. The figure shows representative cases from the Prostate158 (left) and PROMIS (right) datasets, with ground truth lesions (red) and model predictions (cyan) overlaid on T2-weighted MRI slices (confidence ≥ 0.5). More qualitative comparison can be found in Appendix A, Figure 7.

Table 3: Segmentation performance on overlapping true positive cases, allowing a direct comparison of lesion delineation quality.

Dataset	Method	Volumetric Dice	Hausdorff 95 (mm)	Surface Distance (mm)
Prostate158	nnU-Net	0.391 ± 0.278	14.60 ± 10.58	5.20 ± 7.17
	SonicNet	0.401 ± 0.257	12.41 ± 9.83	3.89 ± 5.74
PROMIS	nnU-Net	0.459 ± 0.222	17.69 ± 13.09	3.78 ± 3.26
	SonicNet	0.488 ± 0.213	14.37 ± 12.18	3.12 ± 2.66

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. Across synthetic and clinical benchmarks, SonicNet consistently showed advantages in robustness to geometric transformations, resilience to noise and acquisition variability, and strong generalisation despite having far fewer parameters than established baselines. These results highlight the potential of spectral operators as a compact and effective alternative to convolutional or attention-based representations. At the same time, important limitations remain. Despite the name, SONIC does not primarily excel in speed: reliance on FFTs introduces considerable computational and memory overhead, which constrains scalability on current GPU hardware compared to spatial convolutions that benefit from kernel-level optimisation. Moreover, the inherently global nature of the frequency-domain representation limits its ability to capture fine-grained local structure, suggesting that hybrid architectures may be needed to get the best of both worlds. 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.

Ethan Baron, Itamar Zimerman, and Lior Wolf. 2-d ssm: A general spatial layer for visual transformers, 2023. URL https://arxiv.org/abs/2306.06635.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL http://arxiv.org/abs/1412.3555. cite arxiv:1412.3555Comment: Presented in NIPS 2014 Deep Learning and Representation Learning Workshop.

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.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *CoRR*, abs/2111.00396, 2021. URL https://arxiv.org/abs/2111.00396.

Alessa Hering, Sarah de Boer, Anindo Saha, Jasper J Twilt, Mattias P Heinrich, Derya Yakar, Maarten de Rooij, Henkjan Huisman, and Joeran S Bosma. Deformable mri sequence registration for ai-based prostate cancer diagnosis. In *International Workshop on Biomedical Image Registration*, pp. 148–162. Springer, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Fabian Isensee, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:1–9, 02 2021. doi: 10.1038/s41592-020-01008-z.

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.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35, 1960. doi: 10.1115/1.3662552. URL http://dx.doi.org/10.1115/1.3662552.

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

Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *CoRR*, abs/2010.08895, 2020. URL https://arxiv.org/abs/2010.08895.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model, 2024. URL https://arxiv.org/abs/2401.10166.

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.

Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. 1999. URL https://cs.nyu.edu/~roweis/papers/NC110201.pdf.

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.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024. URL https://arxiv.org/abs/2401.09417.

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., $|a_m^{\mathrm{im}}| \le \rho$). 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.

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.

Results Visual results from the SynthShape experiment are shown in Figure 5 and 6, comparing convolution with Sonic (ViT omitted due to poor segmentation). The example illustrates one of the five trials summarized in Table 1.

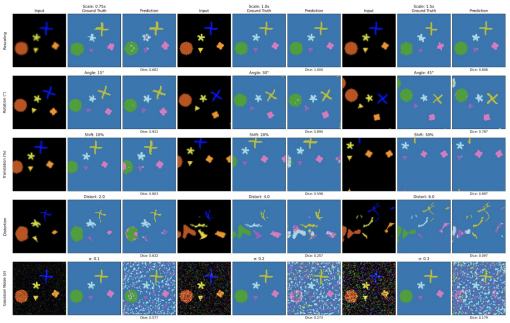


Figure 5: Segmentation result using ConvNet.

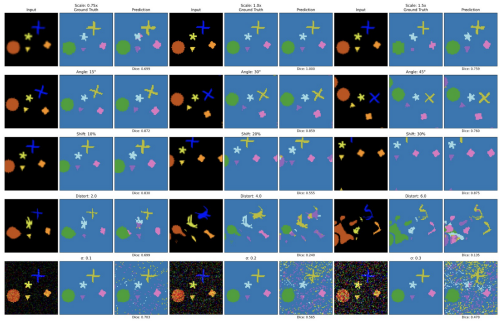


Figure 6: Segmentation result using SonicNet.

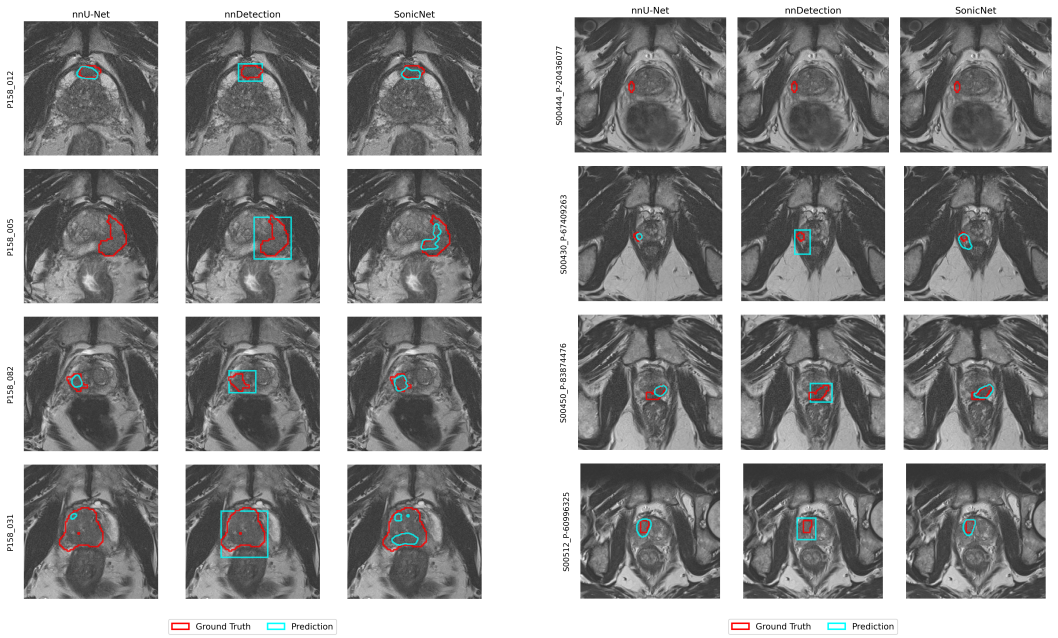
Prostate Cancer Detection

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). The decomposition of the parameters are given in Table 4:

Table 4: Parameter breakdown of `SonicNet` with C input channels, K output channels, M modes, and a depth of four SonicBlocks. "7M+1" corresponds to the transfer function parameters with $D = 3$ and 1 for ρ.

	C	K	M	$2MC$	$2KM$	$7M{+}1$	Projection	GroupNorm	Block sum
Block 0	3	64	64	8192	384	449	192	6	9223
Block 1	64	64	64	8192	8192	449		128	16961
Block 2	64	64	64	8192	8192	449		128	16961
Block 3	64	64	64	8192	8192	449		128	16961
Linear projection	64	2					3458	128	3586
Total									**63692**

All three models were optimized using the Focal Loss (Lin et al., 2018), which down-weights easy negatives and focuses the training signal on difficult examples. 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 three 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.



(a) **Prostate158** (b) **PROMIS**

Figure 7: Qualitative comparison of nnU-Net, nnDetection, and SonicNet predictions on the **Prostate158** (left) and **PROMIS** (right) datasets. Ground-truth lesions are shown in red and model predictions in cyan.

APPENDIX B: 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

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t), \tag{14}$$

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

$$\mathcal{K}(t) = Ce^{At}B, \qquad t \geq 0. \tag{15}$$

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

$$H(s) = \int_0^\infty e^{-st}\,\mathcal{K}(t)\,dt = \int_0^\infty e^{-st}\,Ce^{At}B\,dt. \tag{16}$$

Pulling out C and B gives

$$H(s) = C\left(\int_0^\infty e^{(A-sI)t}\,dt\right)B. \tag{17}$$

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

$$\int_0^\infty e^{(A-sI)t}\,dt = (sI - A)^{-1}. \tag{18}$$

Hence the transfer function is

$$H(s) = C(sI - A)^{-1}B. \tag{19}$$

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

$$\widehat{H}_{c,k}(\boldsymbol{\omega}) = \sum_{m=1}^{M} C_{km}\,T_m(\boldsymbol{\omega})\,B_{mc}, \tag{20}$$

with mode response

$$T_m(\boldsymbol{\omega}) = \frac{1}{is_m(\boldsymbol{\omega}\cdot v_m) \;-\; a_m \;+\; \tau_m\|(I - v_m v_m^\top)\boldsymbol{\omega}\|^2}. \tag{21}$$

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,

$$T_m(\boldsymbol{\omega}) = \frac{1}{is_m\,\omega_d - a_m + \tau_m\sum_{j\neq d}\omega_j^2}. \tag{22}$$

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

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