S4ND: Modeling Images and Videos as Multidimensional Signals Using State Spaces

Anonymous Author(s)
Affiliation
Address
email

Abstract

Visual data such as images and videos are typically modeled as discretizations of inherently continuous, multidimensional signals. Existing continuous-signal models attempt to exploit this fact by modeling the underlying signals of visual (e.g., image) data directly. However, they have not yet been able to achieve competitive performance on practical vision tasks such as large-scale image and video classification. Building on a recent line of work on deep state space models (SSMs), we propose S4ND, a new multidimensional SSM layer that extends SSMs' continuous-signal modeling ability to multidimensional data including images and videos. We show that S4ND can model large-scale visual data in 1D, 2D, and 3D as continuous multidimensional signals and demonstrate strong performance by simply swapping Conv2D and self-attention layers with S4ND layers in existing state-of-the-art models. On ImageNet-1k, S4ND exceeds the performance of a ViT baseline by 1.5% accuracy when training with a 1D sequence of patches, and matches ConvNeXt when modeling images in 2D. For videos, S4ND improves on an inflated 3D ConvNeXt in activity classification on HMDB-51 by 4% accuracy. S4ND implicitly learns global, continuous convolutional kernels that are resolution invariant by construction, providing an inductive bias that enables generalization across multiple resolutions. By developing a simple bandlimiting modification to S4 to overcome aliasing, S4ND achieves strong zero-shot (unseen at test time) resolution performance, e.g. achieving 88.7% accuracy on CIFAR-10 when trained on 16×16 and tested on 32×32 images. When trained with progressive resizing, S4ND comes within $\sim 1\%$ of a high-resolution model while training 22% faster.

1 Introduction

Modeling visual data such as images and videos is a canonical problem in deep learning. In the last few years, many modern deep learning backbones that achieve strong performance on benchmarks like ImageNet [46] have been proposed. These backbones are diverse, and include 1D sequence models such as the Vision Transformer (ViT) [10], which treats images as sequences of patches, and 2D and 3D models that use local convolutions over images and videos (ConvNets) [31, 20, 49, 51, 52, 36, 19, 26, 42, 55, 12].

A commonality among modern vision models capable of achieving state-of-the-art (SotA) performance is that they treat visual data as discrete pixels rather than continuous-signals. However, images and videos are discretizations of multidimensional and naturally *continuous* signals, sampled at a fixed rate in the spatial and temporal dimensions. Ideally, we would want approaches that are capable of recognizing this distinction between data and signal, and directly model the underlying continuous-signals. This would give them capabilities like the ability to adapt the model to data sampled at different resolutions.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

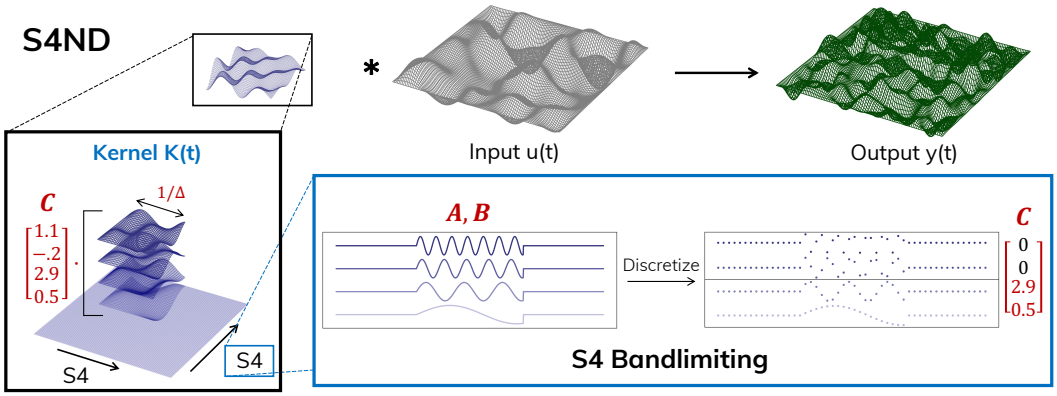


Figure 1: (**S4ND**.) (Parameters in *red*.) (*Top*) S4ND can be viewed as a depthwise convolution that maps a multidimensional input (*black*) to output (*green*) through a continuous convolution kernel (*blue*). (*Bottom Left*) The kernel can be interpreted as a linear combination (controlled by $\boldsymbol{C}$) of basis functions (controlled by $\boldsymbol{A}, \boldsymbol{B}$) with flexible width (controlled by Δ). For structured $\boldsymbol{C}$, the kernel can further factor as a low-rank tensor product of 1D kernels, and can be interpreted as independent S4 transformations on each dimension. (*Bottom Right*) Choosing $\boldsymbol{A}, \boldsymbol{B}$ appropriately yields Fourier basis functions with controllable frequencies. To avoid aliasing in the final discrete kernels, the coefficients in $\boldsymbol{C}$ corresponding to high frequencies can simply be masked out.

A natural approach to building such models is to parameterize and learn continuous convolutional kernels, which can then be sampled differently for data at different resolutions [13, 43, 47, 16, 17]. Among these, deep state space models (SSM) [16], in particular S4 [17], have achieved SotA results in modeling sequence data derived from continuous-signals, such as audio [14]. However, a key limitation of SSMs is that they were developed for 1D signals, and cannot directly be applied to visual data derived from multidimensional "ND" signals. Given that 1D SSMs outperform other continuous modeling solutions for sequence data [17], and have had preliminary success on image [17] and video classification [25], we hypothesize that they may be well suited to modeling visual data when appropriately generalized to the setting of multidimensional signals.

Our main contribution is S4ND, a new deep learning layer that extends S4 to multidimensional signals. The key idea is to turn the standard SSM (a 1D ODE) into a multidimensional PDE governed by an independent SSM per dimension. By adding additional structure to this ND SSM, we show that it is equivalent to an ND continuous convolution that can be factored into a separate 1D SSM convolution per dimension. This results in a model that is efficient and easy to implement, using the standard 1D S4 layer as a black box. Furthermore, it can be controlled by S4's parameterization, allowing it to model both long-range dependencies, or finite windows with a learnable window size that generalize conventional local convolutions [0].[1]

We show that S4ND can be used as a drop-in replacement in strong modern vision architectures while matching or improving performance in 1D, 2D, and 3D. With minimal change to the training procedure, replacing the self-attention in ViT with S4-1D improves top-1 accuracy by 1.5%, and replacing the convolution layers in a 2D ConvNeXt backbone [36] preserves its performance on ImageNet-1k [7]. Simply inflating this pretrained S4-2D-ConvNeXt backbone to 3D improves video activity classification results on HMDB-51 [32] by 4 points over the pretrained ConvNeXt baseline.

Additionally, we propose a low-pass bandlimiting modification to S4 that encourages the learned convolutional kernels to be smooth. While S4ND can be used at any resolution, performance suffers when moving between resolutions due to aliasing artifacts in the kernel, an issue also noted by prior work on continuous models [43]. While S4 was capable of transferring between different resolutions on audio data [17], visual data presents a greater challenge due to the scale-invariant properties of images in space and time [45], as sampled images with more distant objects are more likely to contain power at frequencies above the Nyquist cutoff frequency. Motivated by this, we propose a simple criteria that masks out frequencies in the S4ND kernel that lie above the Nyquist cutoff frequency.

The continuous-signal modeling capabilities of S4ND open the door to new training recipes, such as the ability to train and test at different resolutions. On the standard CIFAR-10 [30] and Celeb-A

[1]The citation [0] refers to a concurrent submission which is included in the supplementary.

[37] datasets, S4ND degrades by as little as 1.3% when upsampling from low- to high-resolution data (e.g. $128 \times 128 \rightarrow 160 \times 160$), and can be used to facilitate progressive resizing to speed up training by 22% with $\sim 1\%$ drop in final accuracy compared to training at the high resolution alone. We also validate that our new bandlimiting method is critical to these capabilities, with ablations showing absolute performance degradation of up to $20\%+$ without it.

2 Related Work

Image Classification. There is a long line of work in image classification, with much of the 2010s dominated by ConvNet backbones [31, 20, 49, 51, 52]. Recently, Transformer backbones, such as ViT [10], have achieved SotA performance on images using self-attention over a sequence of 1D patches [34, 33, 54, 61, 9]. Their scaling behavior in both model and dataset training size is believed to give them an inherent advantage over ConvNets [10], even with minimal inductive bias. Liu et al. [36] introduce ConvNeXt, which modernizes the standard ResNet architecture [20] using modern training techniques, matching the performance of Transformers on image classification. We select a backbone in the 1D and 2D settings, ViT and ConvNeXt, to convert into continuous-signal models by replacing the multi-headed self-attention layers in ViT and the standard Conv2D layers in ConvNeXt with S4ND layers, maintaining their top-1 accuracy on large-scale image classification.

S4 & Video Classification. To handle the long-range dependancies inherent in videos, [25] used 1D S4 for video classification on the Long-form Video Understanding dataset [58]. They first applied a Transformer to each frame to obtain a sequence of patch embeddings for each video frame independently, followed by a standard 1D S4 to model across the concatenated sequence of patches. This is akin to previous methods that learned spatial and temporal information separately [27], for example using ConvNets on single frames, followed by an LSTM [22] to aggregate temporal information. In contrast, modern video architectures such as 3D ConvNets and Transformers [19, 26, 42, 55, 12, 29, 35, 58, 2, 1] show stronger results when learning spatiotemporal features simultaneously, which the generalization of S4ND into multidimensions now enables us to do.

Continuous-signal Models. Visual data are discretizations of naturally continuous signals that possess extensive structure in the joint distribution of spatial frequencies, including the properties of scale and translation invariance. For example, an object in an image generates correlations between lower and higher frequencies that arises in part from phase alignment at edges [41]. As an object changes distances in the image, these correlations remain the same but the frequencies shift. This relationship can potentially be learned from a coarsely sampled image and then applied at higher frequency at higher resolution.

A number of continuous-signal models have been proposed for the visual domain to learn these inductive biases, and have led to additional desirable properties and capabilities. A classic example of continuous-signal driven processing is the fast Fourier transform, which is routinely used for filtering and data consistency in computational and medical imaging [8]. Using a continuous function, NeRF represents a static scene as a continuous function, allowing them to render scenes smoothly from multiple viewpoints [39]. CKConv [44] learns a continuous representation to create kernels of arbitrary size for several data types including images, with additional benefits such as the ability to handle irregularly sampled data. FlexConv [43] extends this work with a learned kernel size, and show that images can be trained at low resolution and tested at high resolution if the aliasing problem is addressed. S4 [17] increased abilities to model long-range dependancies using continuous kernels, allowing SSMs to achieve SotA on sequential CIFAR [30]. However, these methods including 1D S4 have been applied to relatively low dimensional data, e.g., time series, and small image datasets. S4ND is the first continuous-signal model applied to high dimensional visual data with the ability to maintain SotA performance on large-scale image and video classification.

Progressive Resizing. Training times for large-scale image classification can be quite long, a trend that is exacerbated by the emergence of foundation models [3]. A number of strategies have emerged for reducing overall training time. Fix-Res [53] trains entirely at a lower resolution, and then fine-tunes at the higher test resolution to speed up training in a two-stage process. Mix-and-Match [23] randomly samples low and high resolutions during training in an interleaved manner. An effective method to reduce training time on images is to utilize progressive resizing. This involves training at a lower resolution and gradually upsampling in stages. For instance, fastai [11] utilized progressive resizing to train an ImageNet in under 4 hours. EfficientNetV2 [52] coupled resizing with

a progressively regularization schedule, increasing the regularization as well to maintain accuracy. In EfficientNetV2 and other described approaches, the models eventually train on the final test resolution. As a continuous-signal model, we demonstrate that S4ND is naturally suited to progressive resizing, while being able to generalize to unseen resolutions at test time.

3 Preliminaries

State space models. S4 investigated state space models, which are linear time-invariant systems that map signals $u(t) \mapsto y(t)$ and can be represented either as a linear ODE (equation (1)) or convolution (equation (2)). Its parameters are $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ and $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^{N}$ for a state size N.

$$\begin{aligned} x'(t) &= \boldsymbol{A}x(t) + \boldsymbol{B}u(t) \\ y(t) &= \boldsymbol{C}x(t) \end{aligned} \tag{1} \qquad \begin{aligned} K(t) &= \boldsymbol{C}e^{t\boldsymbol{A}}\boldsymbol{B} \\ y(t) &= (K * u)(t) \end{aligned} \tag{2}$$

Basis functions. For the clearest intuition, we think of the convolution kernel as a linear combination (controlled by $\boldsymbol{C}$) of **basis kernels** $K_n(t)$ (controlled by $\boldsymbol{A}, \boldsymbol{B}$)

$$K(t) = \sum_{k=0}^{N-1} \boldsymbol{C}_k K_k(t) \qquad K_n(t) = (e^{t\boldsymbol{A}}\boldsymbol{B})_n \tag{3}$$

Discretization. The SSM (1) is defined over a continuous-time axis and produces continuous-time convolution kernels (2)(3). Given a discrete input sequence $u_0, u_1, \ldots$ sampled uniformly from an underlying signal $u(t)$ at a step size Δ (i.e. $u_k = u(k\Delta)$), the kernel can be sampled to match the rate of the input. Note that instead of directly sampling the kernel, standard discretization rules should be applied to minimize the error from the discrete to the continuous-time kernel [17]. For inputs given at different resolutions, the model can then simply change its Δ value to compute the kernel at different resolutions.

We note that the step size Δ does not have to be exactly equal to a "true sampling rate" of the underlying signal, but only the relative rate matters. Concretely, the discrete-time kernel depends only on the *product* $\Delta\boldsymbol{A}$ and $\Delta\boldsymbol{B}$, and S4 learns separate parameters $\Delta, \boldsymbol{A}, \boldsymbol{B}$.

S4. S4 is a special SSM with prescribed $(\boldsymbol{A}, \boldsymbol{B})$ matrices that define well-behaved basis functions, and an algorithm that allows the convolution kernel to be computed efficiently. Variants of S4 exist that define different basis functions, such as simple diagonal SSMs [18], or one that defines **truncated Fourier functions** $K_n(t) = \sin(2\pi nt)\mathbb{I}[0, 1]$ [0] (Fig. 1). These versions of S4 have easy-to-interpret basis functions that will allow us to control the frequencies in the kernel (Section 4.2).

4 Method

We describe the proposed S4ND model for the 2D case only, for ease of notation and presentation. The results extend readily to general dimensions; full statements and proofs for the general case are in Appendix B. Section 4.1 describes the multidimensional S4ND layer, and Section 4.2 describes our simple modification to restrict frequencies in the kernels. Fig. 1 illustrates the complete S4ND layer.

4.1 S4ND

We begin by generalizing the (linear time-invariant) SSM (1) to higher dimensions. Notationally, we denote the individual time axes with superscripts in parentheses. Let $u = u(t^{(1)}, t^{(2)})$ and $y = y(t^{(1)}, t^{(2)})$ be the input and output which are signals $\mathbb{R}^2 \to \mathbb{C}$, and $x = (x^{(1)}(t^{(1)}, t^{(2)}), x^{(2)}(t^{(1)}, t^{(2)})) \in \mathbb{C}^{N^{(1)} \times N^{(2)}}$ be the SSM state of dimension $N^{(1)} \times N^{(2)}$, where $x^{(\tau)} : \mathbb{R}^2 \to \mathbb{C}^{N^{(\tau)}}$.

Definition 1 (Multidimensional SSM)**.** *Given parameters $\boldsymbol{A}^{(\tau)} \in \mathbb{C}^{N^{(\tau)} \times N^{(\tau)}}$, $\boldsymbol{B}^{(\tau)} \in \mathbb{C}^{N^{(\tau)} \times 1}$, $\boldsymbol{C} \in \mathbb{C}^{N^{(1)} \times N^{(2)}}$, the 2D SSM is the map $u \mapsto y$ defined by the linear PDE with initial condition*

$x(0,0) = 0$:

$$\begin{aligned}
\frac{\partial}{\partial t^{(1)}} x(t^{(1)}, t^{(2)}) &= (\boldsymbol{A}^{(1)} x^{(1)}(t^{(1)}, t^{(2)}), x^{(2)}(t^{(1)}, t^{(2)})) + \boldsymbol{B}^{(1)} u \\
\frac{\partial}{\partial t^{(2)}} x(t^{(1)}, t^{(2)}) &= (x^{(1)}(t^{(1)}, t^{(2)}), \boldsymbol{A}^{(2)} x^{(2)}(t^{(1)}, t^{(2)})) + \boldsymbol{B}^{(1)} u \\
y(t^{(1)}, t^{(2)}) &= \langle \boldsymbol{C}, x(t^{(1)}, t^{(2)}) \rangle
\end{aligned} \tag{4}$$

Note that Definition 1 differs from the usual notion of multidimensional SSM, which is simply a map from $u(t) \in \mathbb{C}^n \mapsto y(t) \in \mathbb{C}^m$ for higher-dimensional $n, m > 1$ but still with 1 time axis. However, Definition 1 is a map from $u(t_1, t_2) \in \mathbb{C}^1 \mapsto y(t_1, t_2) \in \mathbb{C}^1$ for *scalar* input/outputs but over *multiple* time axes. When thinking of the input $u(t^{(1)}, t^{(2)})$ as a function over a 2D grid, Definition 1 can be thought of as a simple linear PDE that just runs a standard 1D SSM over each axis independently.

Analogous to equation (2), the 2D SSM can also be viewed as a multidimensional convolution.

Theorem 1. (4) *is a time-invariant system that is equivalent to a 2D convolution $y = K * u$ by the kernel*

$$K(t^{(1)}, t^{(2)}) = \langle \boldsymbol{C}, (e^{t\boldsymbol{A}^{(1)}} \boldsymbol{B}^{(1)}) \otimes (e^{t\boldsymbol{A}^{(2)}} \boldsymbol{B}^{(2)}) \rangle \tag{5}$$

This kernel is a linear combination of the $N^{(1)} \times N^{(2)}$ basis kernels $\{K^{(1)}_{n^{(1)}}(t^{(1)}) \otimes K^{(1)}_{n^{(2)}}(t^{(2)}) : n^{(1)} \in [N^{(1)}], n^{(2)} \in [N^{(2)}]\}$ where $K^{(\tau)}$ are the standard 1D SSM kernels (3) *for each axis .*

However, a limitation of this general form is that the number of basis functions $N^{(1)} \times N^{(2)} \times \ldots$ grows exponentially in the dimension, increasing the parameter count (of $\boldsymbol{C}$) and overall computation dramatically. This can be mitigated by factoring $\boldsymbol{C}$ as a low-rank tensor.

Theorem 2. *Suppose that $\boldsymbol{C} \in \mathbb{C}^{N^{(1)} \times N^{(2)}}$ is a low-rank tensor $\boldsymbol{C} = \sum_{i=1}^{r} \boldsymbol{C}_{\boldsymbol{i}}^{(1)} \otimes \boldsymbol{C}_{\boldsymbol{i}}^{(2)}$ where each $\boldsymbol{C}_i^{(\boldsymbol{\tau})} \in \mathbb{C}^{N^{(\tau)}}$. Then the kernel* (5) *also factors as a tensor product of 1D kernels*

$$K(t^{(1)}, t^{(2)}) = \sum_{i=1}^{r} K_i^{(1)}(t^{(1)}) \otimes K_i^{(2)}(t^{(2)}) := \sum_{i=1}^{r} (\boldsymbol{C}_{\boldsymbol{i}}^{(\boldsymbol{1})} e^{t^{(2)}\boldsymbol{A}^{(1)}} \boldsymbol{B}^{(\boldsymbol{1})}) \otimes (\boldsymbol{C}_{\boldsymbol{i}}^{(\boldsymbol{2})} e^{t^{(2)}\boldsymbol{A}^{(2)}} \boldsymbol{B}^{(\boldsymbol{2})})$$

In our experiments, we choose $\boldsymbol{C}$ as a rank-1 tensor, but the rank can be freely adjusted to tradeoff parameters and computation for expressivity. Using the equivalence between (1) and (2), Theorem 2 also has the simple interpretion as defining an independent 1D SSM along each axis of the multidimensional input.

4.2 Resolution Change and Bandlimiting

SSMs in 1D have shown strong performance in the audio domain, and can nearly preserve full accuracy when tested zero-shot on inputs sampled at very different frequencies [17]. This capability relies simply on scaling Δ by the relative change in frequencies (i.e., if the input resolution is doubled, halve the SSM's Δ parameter). However, sampling rates in the spatial domain are often much lower than temporally, leading to potential aliasing when changing resolutions. A standard technique to avoid aliasing is to apply a low-pass filter to remove frequencies above the Nyquist cutoff frequency.

For example, when $\boldsymbol{A}$ is diagonal with n-th element $\boldsymbol{a}_n$, each basis function has simple form $K_n(t) = e^{t\boldsymbol{a}_n} \boldsymbol{B}_n$. Note that the frequencies are mainly controlled by the imaginary part of $\boldsymbol{a}_n$. We propose the following simple method: for any n such that $\boldsymbol{a}_n \cdot \Delta < \frac{1}{2}\alpha$, mask out the corresponding coefficient of the linear combination $\boldsymbol{C}_n$ (equation (3)). Here α is a hyperparameter that controls the cutoff; theoretically, $\alpha = 1.0$ corresponds to the Nyquist cutoff if the basis functions are pure sinusoids. However, due to the decay $e^{\Re(\boldsymbol{a}_n)}$ arising from the real part as well as approximations arising from using finite-state SSMs, α often has to be set lower empirically.

5 Experiments

We evaluate S4ND on large-scale image classification in Section 5.1 in the 1D and 2D settings, followed by activity classification in videos in Section 5.2 in the 3D setting, where using S4ND as

a drop-in replacement for standard deep learning layers matches or improves performance in all settings. In Section 5.3, we performed controlled ablations to highlight the benefits of S4ND as a continuous-signal model in images.

5.1 S4ND in 1D & 2D: Large-scale Image Classification

First, we show that S4ND is a drop-in replacement for existing visual modeling layers such as 1D self-attention and 2D local convolutions, with no degradation in top-1 performance when used in modern backbones such as ViT [10] and ConvNeXt [36] on ImageNet-1k [7].

Baselines and Methodology. We consider large-scale image classification on the ImageNet-1k dataset, which consists of 1000 classes and 1.3M images. We start with two strong baselines: ViT-B (base, 88M parameters) for processing images in the 1D setting and ConvNeXt-T (tiny, 28.4M) in the 2D setting. (We omit the postfix "B" and "T" for brevity). More recent works using Transformers on images have surpassed ViT, but we focus on the original ViT model to highlight specifically the drop-in capability and performance difference in self-attention vs. S4ND layers. We first swap the self-attention layers in ViT with S4ND layers, and call this model S4ND-ViT. Notably, we simplify ViT by removing the positional encodings, as S4ND does not require injecting this inductive bias. Similarly, we swap the depthwise Conv2D layers in the ConvNeXt blocks with S4ND layers, which we call S4ND-ConvNeXt. These result in similar parameter counts compared to their baseline models, with S4ND-ViT at 88.8M and S4ND-ConvNeXt at 30M.

Training. For all ImageNet models, we train from scratch with no outside data and adopt the training procedure from [54, 60], which uses the AdamW optimizer [38] for 300 epochs, cosine decay learning rate, weight decay 0.05, and aggressive data augmentations including RandAugment [5], Mixup [62], and AugMix [21]. We add RepeatAug [24] for ConvNeXt and S4ND-ConvNeXt. The initial learning rate for ViT (and S4ND-ViT) is 0.001, while for ConvNeXt (and S4ND-ConvNeXt) it is 0.0001. See Appendix C.1 for additional training procedure details.

Table 1: (**Performance on image classification.**) Top-1 test accuracy benchmarks for images in the 1D and 2D settings. ConvNeXt-M, for "micro", is a reduced model size for Celeb-A, while "-ISO" is an isotropic S4ND backbone [17].

Model	Dataset	Acc
ViT-B	ImageNet	78.9
S4ND-ViT-B	ImageNet	**80.4**
ConvNeXt-T	ImageNet	82.1
S4ND-ConvNeXt-T	ImageNet	**82.2**
Conv2D-ISO	CIFAR-10	93.7
S4ND-ISO	CIFAR-10	**94.1**
ConvNeXt-M	Celeb-A	91.0
S4ND-ConvNeXt-M	Celeb-A	**91.3**

Results. Table 1 shows top-1 accuracy results for each model on ImageNet. After reproducing the baselines, S4ND-ViT was able to moderately boost performance by $+1.5\%$ over ViT, while S4ND-ConvNeXt matched the original ConvNeXt's performance. This indicates that S4ND is a strong primitive that can replace self-attention and standard 2D convolutions in practical image settings with large-scale data.

5.2 S4ND in 3D: Video Classification

Next, we demonstrate the flexible capabilities of S4ND in settings involving pretraining and even higher-dimensional signals. We use the activity recognition dataset HMDB-51 [32] which involves classifying videos in 51 activity classes.

Baselines and Methodology. Prior work demonstrated that 2D CNNs (e.g. pretrained on ImageNet) can be adapted to 3D models by 2D to 3D kernel *inflation* (I3D [4]), in which the 2D kernels are repeated temporally N times and normalized by $1/N$. Our baseline, which we call ConvNeXt-I3D, uses the 2D ConvNeXt pretrained on ImageNet (Section 5.1) with I3D inflation. We additionally test the more modern spatial-temporal separated 3D convolution used by S3D [59] and R(2+1)D [56], which factor the 3D convolution kernel as the outer product of a 2D (spatial) by 1D (temporal) kernel.

Because of its flexible factored form (Section 4.1), S4ND automatically has these inflation capabilities. We inflate the pretrained S4ND-ConvNeXt simply by loading the pretrained 2D model weights for the spatial dimensions, and initializing the temporal kernel parameters $\boldsymbol{A}^{(3)}, \boldsymbol{B}^{(3)}, \boldsymbol{C}^{(3)}$ from scratch. We note that this model is essentially identical to the baseline ConvNeXt-S3D except that each component of the factored kernels use standard 1D S4 layers instead of 1D local convolutions.

Table 3: (**Settings for continuous capabilities experiments.**) Datasets and resolutions used for continuous capabilities experiments, as well as the model backbones used are summarized.

Dataset	Classes	Resolution			Backbone
		Base	Mid	Low	
CIFAR-10	10	32×32	16×16 $(2\times)$	8×8 $(4\times)$	Isotropic
Celeb-A	40 multilabel	160×160	128×128 $(1.25\times)$	64×64 $(2.50\times)$	ConvNeXt

Finally, by varying the initialization of these parameters, we can investigate additional factors affecting model training; in particular, we also run an ablation on the kernel timescales Δ.

Training. Our training procedure is minimal, using only RGB frames (no optical flow). We sample clips of 2 seconds with 30 total frames at 224×224, followed by applying RandAugment; we performed a small sweep of the RandAugment magnitude for each model. All models are trained with learning rate 0.0001 and weight decay 0.05. Additional details are included in Appendix C.2.

Results. Results are presented in Table 2. Our baselines are much stronger than prior work in this setting, 8% top-1 accuracy higher than the original I3D model in the RGB frames only setting, and confirming that separable kernels (S3D) perform at least as well as standard inflation ($+0.53\%$).

S4ND-ConvNeXt-3D improves over the baseline ConvNeXt-I3D by $+4.0\%$ with no difference in models other than using a temporal S4 kernel. This even exceeds the performance of I3D when trained on optical flow.

Table 2: (**HMDB-51 Activity Recognition with ImageNet-pretrained models.**) (*Left*) Top-1 accuracy with 2D to 3D kernel inflation. (*Right*) Ablation of initial temporal kernel lengths, controlled by S4's Δ parameter.

	FLOW	RGB
Inception-I3D [4]	61.9	49.8
ConvNeXt-I3D	-	58.07
ConvNeXt-S3D	-	58.60
S4ND-ConvNeXt-3D	-	**62.07**

Init. Length	Acc.
20.0	53.74
4.0	58.33
2.0	60.30
1.0	62.07

Finally, we show how S4ND's parameters can control for factors such as the kernel length (Table 2). Note that our temporal kernels $K^3(t^3)$ are always full length (30 frames in this case), while standard convolution kernels are shorter temporally and require setting the width of each layer manually as a hyperparameter [59]. S4 layers have a parameter Δ that can be interpreted such that $\frac{1}{\Delta}$ is the expected length of the kernel. By simply adjusting this hyperparameter, S4ND can be essentially initialized with length-1 temporal kernels that can automatically learn to cover the whole temporal length if needed. We hypothesize that this contributes to S4ND's improved performance over baselines.

5.3 Continuous-signal Capabilities for Images

Images are often collected at varied resolutions, even with the same hardware, so it is desirable that models generalize to data sampled at different resolutions. We show that S4ND inherits this capability as a continuous-signal model, with strong zero-shot performance when changing resolutions, and the ability to train with progressively resized multi-resolution data. We ablate that our proposed bandlimiting modification is critical to achieving strong performance when changing resolutions.

Setup. We focus on image classification on 2D benchmark datasets: a dataset with low-resolution images (CIFAR-10) and one with higher-resolution images (Celeb-A). For each dataset, we specify a base image resolution (base), and two lower resolutions (mid, low), summarized in Table 3. To highlight that S4ND's continuous capabilities are independent of backbone, we experiment with 2 different 2D backbones: an isotropic, fixed-width model backbone on CIFAR-10 [30] and a small ConvNeXt backbone on Celeb-A [37]. For each backbone, we compare S4ND's performance to Conv2D layers as a standard, widely used baseline. Additional details can be found in Appendix C.3.

Table 4: (**Zero-Shot Resolution Change.**) Results on validation accuracy for models trained on one resolution (one of low / mid / base), and zero-shot tested on another.

Resolution		CIFAR-10			Celeb-A	
Train	Test	S4ND	Conv2D	FlexNet-16	S4ND	Conv2D
base	base	**93.4**	91.9	92.2	**91.8**	91.4
mid	mid	**88.7**	87.2	86.5	**91.6**	91.1
mid	base	**88.7**	73.1	82.7	**90.5**	81.3
low	low	**78.0**	76.0	-	**91.0**	90.4
low	mid	**79.1**	57.4	-	**84.5**	80.5
low	base	**74.2**	33.1	-	**84.3**	80.7

We first verify that S4ND achieves comparable test classification performance with baseline Conv2D models. Table 1 shows that we exceed the performance of Conv2D on both tasks, with S4ND improving over Conv2D models by 0.4% on CIFAR-10 and 0.3% on Celeb-A.

Zero-Shot Resolution Change. We train S4ND and Conv2D models at either low or mid resolution, and test them at the base resolution for each dataset. We also compare to FlexConv [43] on CIFAR-10, which is the current SotA for zero-shot resolution change. Compared to training and testing at the base resolution, we expect that Conv2D should degrade more strongly than S4ND, since it cannot adapt its kernel appropriately to the changed resolution. Table 4 shows that S4ND outperforms Conv2D on $\mathrm{mid} \to \mathrm{base}$ by $15+$ points and $\mathrm{low} \to \mathrm{base}$ by $40+$ points on CIFAR-10, and $9+$ points and $3+$ points on Celeb-A. In fact, S4ND yields better performance on $\mathrm{low} \to \mathrm{base}$ (a more difficult task) than Conv2D does from $\mathrm{mid} \to \mathrm{base}$ (an easier task), improving by 1.1% on CIFAR-10 and 3% on Celeb-A. Compared to FlexConv on CIFAR-10 $\mathrm{mid} \to \mathrm{base}$, S4ND improves zero-shot performance by $5+$ points, setting a new SotA.

Progressively Resized Training. We provide an exploration of training with progressive resizing [11, 52] i.e. training in multiple stages at different resolutions. The only change we make from standard training is to reset the learning rate scheduler at the beginning of each stage (details in Appendix C.3). We compare S4ND and the Conv2D baseline with progressive resizing in Table 5.

For CIFAR-10, we train with a $\mathrm{low} \to \mathrm{base}$, $80 - 20$ epoch schedule, and perform within $\sim 1\%$ of an S4ND model trained with base resolution data while speeding up training by 21.8%. We note that Conv2D attains much higher speedups as a consequence of highly optimized implementations, which we discuss in more detail in Section 6. For Celeb-A, we explore flexibly combining the benefits of both progressive resizing and zero-shot testing, training with a $\mathrm{low} \to \mathrm{mid}$, $16 - 4$ epoch schedule that uses no base data. We outperform Conv2D by $7.5\%+$, and attain large speedups of $50\%+$ over training at the base resolution.

Effect of Bandlimiting. Bandlimiting in S4ND is critical to generalization at different resolutions. We analyze the effect of the bandlimiting parameter α on CIFAR-10 performance when doing zero-shot resolution change. We additionally vary the choice of basis function $K_n(t)$ used in the SSM. Fig. 2 shows that zero-shot performance on base degrades for larger values of α i.e. for cutoffs that do not remove high frequencies that violate the Nyquist cutoff. As we would expect, this holds regardless of the choice of basis function. In Section 5.3, we visualize learned kernels with and without bandlimiting, showing that bandlimiting improves smoothness. Appendix C.3 includes additional experiments that analyze α on Celeb-A, and when doing progressively resized training.

6 Discussion

Summary. We introduced S4ND, a multidimensional extension of S4 that models visual data as continuous valued signals. S4ND is the first continuous model that matches SotA baselines on large-scale 1D and 2D image classification on ImageNet-1k, as well as outperforming a strong pretrained model in a 3D video classification setting. As a continuous-signal model, S4ND inherits useful properties that are absent from standard visual modeling approaches, such as zero-shot testing on unseen resolutions without a significant performance drop.

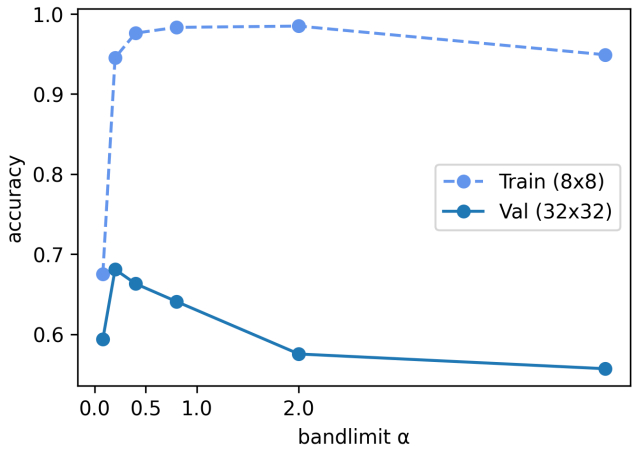
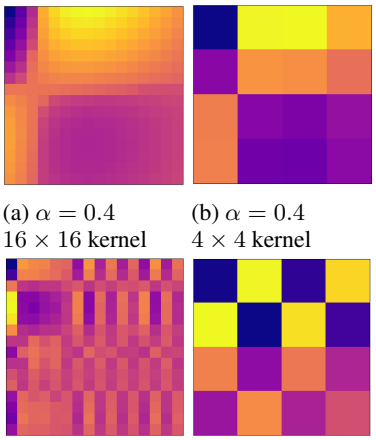


Figure 2: (**Bandlimiting ablation.**) Zero-shot performance training at 8×8 and evaluating at 32×32. Train performance is high for large enough values of $\alpha \geq 0.5$, but validation performance goes down as aliasing occurs in the kernel. Both train and validation drop for lower values of α, as it limits the expressivity of the kernel.



(a) $\alpha = 0.4$ 16×16 kernel (b) $\alpha = 0.4$ 4×4 kernel

(c) $\alpha = \infty$ 16×16 kernel (d) $\alpha = \infty$ 4×4 kernel

Figure 4: (**Effect of bandlimiting on learned kernels.**) Bandlimiting significantly increases the smoothness of the kernels when resizing resolutions.

Table 5: (**Progressive resizing results.**) Validation performance for progressively resized training at base resolution, and speedup compared to training at base resolution on CIFAR-10 and Celeb-A. We use a $80 - 20$ and $16 - 4$ schedule for CIFAR-10 and Celeb-A, and also report performance training only at base resolution.

Dataset	Model	Epoch Schedule	Train Resolution	Val @ Base Res.	Speedup (Step Time)
CIFAR-10	Conv2D	-	base	91.90%	0%
	S4ND	-	base	**93.40%**	0%
	Conv2D	$80 - 20$	low→ base	90.94%	51.7%
	S4ND	$80 - 20$	low→ base	**92.32%**	21.8%
CelebA	Conv2D	-	base	91.44%	0%
	S4ND	-	base	**91.75%**	0%
	Conv2D	$16 - 4$	low→ mid	80.89%	76.7%
	S4ND	$16 - 4$	low→ mid	**88.57%**	57.3%

Limitations. A limitation of S4ND is its training speed in high dimensions. In the 1D image setting, S4ND-ViT has similar training speed to ViT; however, in the 2D setting, the S4ND-ConvNeXt was $2\times$ slower than the baseline ConvNeXt. We remark that vanilla local convolutions have been heavily optimized for years, and we expect that layers such as S4ND can be substantially sped up with more optimized implementations. Our core computational primitives accounting for 65% of our runtime (FFT, pointwise operations, inverse FFT) are all bottlenecked by reading from and writing to GPU memory [50]. With a more optimized implementation that fuses these operations [57] (i.e., loading the input once from GPU memory, perform all operations, then write result back to GPU memory), we expect to speed these up by $3\times$. Further discussion can be found in Appendix D.

Future work. We presented a first step in using continuous-signal models in images and videos, and believe this opens the door to new capabilities and directions. For example, recent video benchmark datasets are significantly larger than the HMDB-51 dataset used in our experiments [28, 15, 6, 40, 48], and can be explored. In addition, we demonstrated capabilities in zero-shot resolution *spatially*, but even less work has been done on zero-shot testing *temporally*. This opens up an exciting new direction of work that could allow models to be agnostic to different video sampling rates as well, capable of testing on higher unseen sampling rates, or irregular (non-uniform) sampling rates. As models become larger and combine multiple modalities, S4ND has shown strong promise in being able to better model underlying continuous-signals and create new capabilities across visual data, audio, time-series (e.g., wearable sensors, health data), and beyond.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Arjun D Desai, Beliz Gunel, Batu M Ozturkler, Harris Beg, Shreyas Vasanawala, Brian A Hargreaves, Christopher Ré, John M Pauly, and Akshay S Chaudhari. Vortex: Physics-driven data augmentations for consistency training for robust accelerated mri reconstruction. *arXiv preprint arXiv:2111.02549*, 2021.

[9] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *arXiv preprint arXiv:2204.03645*, 2022.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] fastai. Training imagenet in 3 hours for 25 minutes. `https://www.fast.ai/2018/04/30/dawnbench-fastai/`, 2018.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[13] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *ICML*, 2020.

[14] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. *arXiv preprint arXiv:2202.09729*, 2022.

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[16] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with the linear state space layer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

[18] Ankit Gupta. Diagonal state spaces are as effective as structured state spaces. *arXiv preprint arXiv:2203.14343*, 2022.

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

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

[23] Elad Hoffer, Berry Weinstein, Itay Hubara, Tal Ben-Nun, Torsten Hoefler, and Daniel Soudry. Mix & match: training convnets with mixed image sizes for improved accuracy, speed and scale resiliency. *arXiv preprint arXiv:1908.08986*, 2019.

[24] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

[25] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. *arXiv preprint arXiv:2204.01692*, 2022.

[26] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231, 2012.

[27] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[29] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021.

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Citeseer*, 2009.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[32] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[33] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society, 2015. ISBN 978-1-4673-8391-2. URL `http://dblp.uni-trier.de/db/conf/iccv/iccv2015.html#LiuLWT15`.

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[40] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[41] Bruno A Olshausen and David J Field. Natural image statistics and efficient coding. *Network: computation in neural systems*, 7(2):333, 1996.

[42] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[43] David W. Romero, Robert-Jan Bruintjes, Jakub M. Tomczak, Erik J. Bekkers, Mark Hoogendoorn, and Jan C. van Gemert. Flexconv: Continuous kernel convolutions with differentiable kernel sizes. *ArXiv*, abs/2110.08059, 2021.

[44] David W Romero, Anna Kuzina, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Ckconv: Continuous kernel convolution for sequential data. *arXiv preprint arXiv:2102.02611*, 2021.

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

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[47] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NeurIPS*, 2017.

[48] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[50] David Střelák and Jiří Filipovič. Performance analysis and autotuning setup of the cufft library. In *Proceedings of the 2nd Workshop on AutotuniNg and aDaptivity AppRoaches for Energy efficient HPC Systems*, pages 1–6, 2018.

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[52] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.

[53] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019.

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[55] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[56] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[57] Guibin Wang, YiSong Lin, and Wei Yi. Kernel fusion: An effective method for better power efficiency on multithreaded gpu. In *2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, pages 344–350. IEEE, 2010.

[58] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021.

[59] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017.

[60] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[61] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *CoRR*, abs/2106.04560, 2021.

[62] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes]
(b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The main method can be implemented based on the released S4 repository, and we will publically release our code. All hyperparameters and experimental details are reported in the Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Full experimental results including standard deviations can be found in the Appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Reported in the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]