State Space Models for Event Cameras

Nikola Zubić and Mathias Gehrig and Davide Scaramuzza
Robotics and Perception Group, University of Zurich, Switzerland
Andreasstrasse 15, 8050 Zurich
rpg.ifi.uzh.ch

Abstract

Today, state-of-the-art deep neural networks that process event-camera data first convert a temporal window of events into dense, grid-like input representations. As such, they exhibit poor generalizability when deployed at higher inference frequencies (i.e., smaller temporal windows) than the ones they were trained on. We address this challenge by introducing state-space models (SSMs) with learnable timescale parameters to event-based vision. This design adapts to varying frequencies without the need to retrain the network at different frequencies. Additionally, we investigate two strategies to counteract aliasing effects when deploying the model at higher frequencies. We comprehensively evaluate our approach against existing methods based on RNN and Transformer architectures across various benchmarks, including Gen1 and 1 Mpx event camera datasets. Our results demonstrate that SSM-based models train 33% faster and also exhibit minimal performance degradation when tested at higher frequencies than the training input. Traditional RNN and Transformer models exhibit performance drops of more than 20 mAP, with SSMs having a drop of 3.31 mAP, highlighting the effectiveness of SSMs in event-based vision tasks.

1. Introduction

Event cameras emerged as a class of sensor technologies that noticeably deviate from the operational mechanics of conventional frame-based cameras [10]. While standard frame-based cameras capture full-frame luminance levels at fixed intervals, event cameras record per-pixel relative brightness changes in a scene at the time they occur. The output is, therefore, an asynchronous stream of events in a four-dimensional spatio-temporal space. Each event is represented by its spatial coordinates on the pixel array, the
temporal instance of the occurrence, and its polarity, which denotes the direction of the brightness change and encapsulates the increase or decrease in luminance. The power of event cameras primarily lies in their ability to capture dynamic scenes with unparalleled temporal resolution (microseconds). This property becomes invaluable in rapidly changing environments or applications requiring very fast response times [9, 44]. However, the richness of the spatiotemporal data they generate introduces complexities in data interpretation and processing. Sophisticated algorithms are required to efficiently parse and make sense of the high-dimensional data space. As such, while event cameras represent a promising frontier in visual sensor technologies, their pervasive utilization depends upon solving these inherent computational challenges.

Current methodologies for addressing problems with event cameras fall predominantly into two categories. The first involves converting the raw stream of spatio-temporal events into dense representations akin to multi-channel images [13, 27, 45, 47]. This transformation allows leveraging conventional computer vision techniques designed for frame-based data. The second category employs sparse computational paradigms, often utilizing spiking neural networks or graph-based architectures [5, 12]. While promising, these methods are not without limitations; they frequently encounter hardware incompatibility issues and compromised accuracy. In this work, we utilize dense representations for their advantages in computational efficiency.

Despite the advances in both paradigms, models trained on event representations at a specific frequency exhibit poor generalizability when deployed in settings with higher frequencies which is crucial for high-speed, dynamical visual scenarios. Additionally, to achieve high performance, it is necessary to include recurrent memory, thereby sacrificing computational efficiency during the training phase. An ideal model would seamlessly merge the training speed of convolutional architectures with the temporal sensitivity and memory benefits inherent to recurrent models.

While recent advancements have introduced efficient recurrent vision transformer architectures [13, 29] to achieve better performance, they face several limitations. Specifically, these architectures suffer from longer training cycles due to their reliance on conventional recurrent mechanisms.

The issue of slow training and generalization at higher event representation frequency than the lower one we trained on remains unresolved as conventional recurrent training methodologies are predominantly utilized in event-based vision. These methods do not incorporate learnable timescale parameters, thus inhibiting the model’s capacity to generalize across varying inference frequencies.

In this work, we address this limitation by introducing structured variations of state-space models [19, 41] as layers within our neural network framework.

State-space models [16] function as CNN during training and are converted to an efficient RNN at test time. To achieve this, S4 [19] employs a technique where the SSM, which is not practical for training on modern hardware due to its sequential nature, is transformed into a CNN by unrolling. S5 [41] uses parallel scans during training, and is employed as RNN during inference. State-space models [16] can be deployed at different frequencies at inference time because they are Linear Time-Invariant (LTI) continuous-time systems that can be transformed into a discrete-time system with an arbitrary step size. This feature permits inference at arbitrary frequencies by globally adjusting the timescale parameter based on the ratio between the new and old frequency sampling rates. Consequently, we tackle the longstanding issue in event-based vision, which requires multiple training cycles with different frequencies to adapt the neural network for various frequencies during inference.

For the task of object detection, we find that incorporating state-space layers accelerates the training by up to 33% relative to existing recurrent vision transformer approaches [13, 29]. This is achieved while maintaining performance comparable to existing methods. Notably, our approach demonstrates superior generalization to higher temporal frequencies with a drop of only 3.31 mAP, while previous methods experience a drop of 21.25 mAP or more. Also, we achieve comparable performance to RVT although we use a linear state-space model rather than a non-linear LSTM model. This also shows that the complexity of the LSTM might not be needed. For this to work, we introduce two strategies (frequency-selective masking and $H_2$ norm) to counteract the aliasing effect encountered with increased temporal frequencies in event cameras. First one is a low-pass bandlimiting modification to SSM that encourages the learned convolutional kernels to be smooth. Second one mitigates the aliasing problem by attenuating the frequency response after a chosen frequency. We argue that state-space modeling offers a significant new direction for research in event-based vision, offering promising solutions to the challenges inherent in processing event-based data effectively and efficiently.

Our contributions are concisely outlined as follows:

- We introduce state-space models for event cameras to address two key challenges in event-based vision: (i) model performance degradation when operating event cameras at temporal frequencies different from their training conditions and (ii) training efficiency.
- Our experimental results outperform existing methods at higher frequencies by 20 mAP on average and show 33% increase in the training speed.
- We introduce two strategies (bandlimiting & $H_2$ norm) designed to alleviate aliasing issues.
2. Related Work

2.1. Object detection with Event Cameras

Approaches in event camera literature, thus in object detection, can be broadly classified into two branches.

The first research branch investigates dense feed-forward methods. Early attempts in this direction relied on a constrained temporal window for generating event representations [4, 22, 24]. The resultant models were deficient in tracking slow-moving or stationary objects, as they failed to incorporate data beyond this limited time-frame. To mitigate these drawbacks, later studies introduced recurrent neural network (RNN) layers [34] into the architecture. The RNN component enhanced the model’s capacity for temporal understanding, thereby improving its object detection capabilities. The work of Zubic et al. [47] takes this a step further by optimizing event representations and incorporating cutting-edge spinn transformer architecture. Nonetheless, their approach was limited in its ability to re-detect slowly moving objects following extended periods of absence. Subsequent research [13, 29], merged the transformer and RNN architectures, pushing the performance further. However, this significantly increased computational demands during the training phase. Importantly, all methodologies examined to date suffer from an inability to adapt when deployed at variable frequencies.

The second research branch investigates the use of Graph Neural Networks (GNNs) or Spiking Neural Networks (SNNs). GNNs dynamically construct a spatiotemporal graph where new nodes and edges are instantiated by selectively sub-sampling events and identifying pre-existing nodes that are proximate in the space-time continuum [12, 23, 43]. A pivotal challenge lies in architecting the network such that information can disseminate effectively across extensive regions of the space-time volume. This becomes particularly important when dealing with large objects that exhibit slow relative motion to the camera. Moreover, while aggressive sub-sampling is often necessary to achieve low-latency inference, it introduces the risk of omitting important information from the event stream. On the other hand, SNNs [2, 7, 39] transmit information sparsely within the system. Unlike RNNs, SNNs emit spikes only when a threshold is met, making them hard to optimize due to the non-differentiable spike mechanism. Some solutions [32] bypass the threshold, but this sacrifices sparsity in deeper layers. Overall, SNNs remain a challenging area needing more foundational research for optimized performance.

2.2. Continuous-time Models

Gu et al. [19] introduced the S4 model as an alternative to CNNs and Transformers for capturing long-range dependencies through LTI systems. This was followed by the S4D model [18], designed for easier understanding and implementation, offering similar performance to S4. The S5 model [41] improved efficiency by avoiding frequency domain computations and utilizing time-domain parallel scans. However, these models have not been thoroughly evaluated on complex, high-dimensional visual data with significant temporal resolution and we are the first ones to do it.

In our study, we empirically show that S4, S4D and S5 models achieve state-of-the-art results when combined with attention mechanisms on complex data. We also identify and address aliasing issues in these models, proposing two corrective strategies. Our work extends the range and robustness of continuous-time models for complex sequence modeling tasks such as object detection for event cameras.

3. Method

In this section, we firstly formalize the operating mechanism of event cameras and provide a notation used for describing state-space models in the preliminaries (Sec. 3.1). Secondly, we describe our approach of incorporating variants of state-space models as layers within our block (Sec. 3.2). This innovative design solves the problems associated with slow training and the variable frequency inference for event cameras.

3.1. Preliminaries

**Event cameras.** Event cameras are bio-inspired vision sensors that capture changes in log intensity per pixel asynchronously, rather than capturing entire frames at fixed intervals. Formally, let $I(x, y, t)$ denote the log intensity at pixel coordinates $(x, y)$ and time $t$. An event $e$ is generated at $(x, y, t)$ whenever the change in log intensity $\Delta I$ exceeds a certain threshold $C$:

$$\Delta I(x, y, t) = I(x, y, t) - I(x, y, t-\Delta t) \geq C$$  

Each event $e$ is a tuple $(x, y, t, p)$, where $(x, y)$ are the pixel coordinates, $t$ is the timestamp, and $p = \{-1, 1\}$ is the polarity of the event, indicating the direction of the intensity change.

**State-Space Models (SSMs).** Linear State-Space Models (SSMs) form the crucial part of the backbone in our architecture, where we compare S4 [19], S4D [18] and S5 [41] layer variants. Given an input vector $u(t) \in \mathbb{R}^U$, a latent state vector $x(t) \in \mathbb{R}^P$, and an output vector $y(t) \in \mathbb{R}^M$, the governing equations of a continuous-time linear SSM can be mathematically represented as:

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t).$$  

The model is parameterized by a state transition matrix $A \in \mathbb{R}^{P \times P}$, an input matrix $B \in \mathbb{R}^{P \times U}$, an output matrix $C \in \mathbb{R}^{M \times P}$, and a direct transmission matrix $D \in \mathbb{R}^{M \times U}$. 
Given a fixed step size $\Delta$, this continuous-time model can be discretized into a linear recurrence using various methods such as Euler, bilinear, or zero-order hold (ZOH):

$$x_k = \overline{A}x_{k-1} + \overline{B}u_k, \quad y_k = \overline{C}x_k + \overline{D}u_k, \quad (3)$$

The parameters in the discrete-time model are functions of the continuous-time parameters, defined by the chosen discretization method. Details on SSMs are available in the appendix.

![SSM-ViT block structure](image)

**3.2. SSM-ViT block**

In this section, we introduce the SSM-ViT block, a novel block depicted in Figure 2, which showcases the structured flow of the proposed block structure, designed for efficient event-based information processing.

We build a 4-stage hierarchical backbone like in [13] where in each stage we utilize our proposed SSM-ViT block. Events are processed into a tensor representation like in [13] before they are used as input to the first stage. Each stage takes the previous features as input and reuses the SSM state from the last timestep to compute features for the next stage. By saving SSM states, each stage retains temporal information for the whole feature map, while also being able to generalize to different frequencies since we use SSM instead of RNN that is used in [13].

Regarding the block structure, initially, the input undergoes convolution with a defined kernel of size $k \times k$ with a stride $s$. This operation effectively captures the essential spatial features of the input. Following the convolution operation, the structure introduces a 'Block-SA' module. This module is pivotal in implementing self-attention mechanisms, but it does so specifically within local windows. The localized nature of the attention in this block ensures a focus on immediate spatial relations, allowing for a detailed representation of close-proximity features.

Subsequent to 'Block-SA', the 'Grid-SA' module comes into play. In contrast to the localized approach of the previous block, 'Grid-SA' employs dilated attention, ensuring a global perspective. This module, by considering a broader scope, encapsulates a more comprehensive understanding of the spatial relations and the overall structure of the input.

The final and crucial component of the block structure is the State-Space Model (SSM) block, that is designed to compute both the output and the state in parallel, either by using S4 [19], S4D [18] or S5 [41] model variant. This ensures temporal consistency and a seamless transition of information between consecutive time steps. Efficient computation is crucial since it allows for the faster training than RNN, and timescale parameter on temporal aggregation is important since we can rescale it during inference and deploy at any frequency we want.

**3.3. Low-pass bandlimiting**

However, the capability to deploy state space models at resolutions higher than those encountered during training is not without its drawbacks. It gives rise to the well-documented issue of aliasing [37, 38], which occurs when the kernel bandwidth surpasses the Nyquist frequency. We adress this challenge in following two subsections - 3.3.1 & 3.3.2.

**3.3.1 Output Masking**

In the realm of signal processing, frequency content beyond the Nyquist rate can lead to aliasing, adversely affecting model performance. To address this, we integrate a frequency-selective masking strategy into the training and inference processes. This bandlimiting method has been empirically validated to be crucial for generalizing across different frequencies [37, 38], with ablation studies indicating a decrease in accuracy by as much as 20% in its absence.

Let the SSM be governed by a matrix $A$, with its diagonal elements $a_n$ influencing the temporal evolution of the system states. The kernel’s basis function for the $n$-th state is given by $K_n(t) = e^{i\alpha_n B_n}$, where the frequency characteristics are primarily dictated by the imaginary part of $\alpha_n$, $\Im(\alpha_n)$ and $B_n$ represents the $B$ matrix.

To modulate the frequency spectrum of the model, we define a hyperparameter $\alpha$, which is used for masking of the computed effective frequency $f_n$ for each state:

$$f_n = \frac{\Delta t}{r} \cdot \left| \Im(\alpha_n) \right| \frac{2\pi}{2}, \quad (4)$$

where $\Delta t$ denotes the discrete time-step, and $r$ is the rate at which we train the model, by default it is 1, rate gets halved when deploying at twice the trained frequency etc.

The bandlimiting mask is then applied as follows:

$$C_n = \begin{cases} 
C_n & \text{if } f_n \leq \frac{\pi}{2} \\
0 & \text{otherwise},
\end{cases} \quad (5)$$

where $C_n$ represents the coefficients in the state representation. The selection of $\alpha$ is critical, with $\alpha = 1.0$ representing the Nyquist limit under idealized conditions. However, practical considerations and model constraints typically necessitate a lower empirical threshold. Our findings suggest that setting $\alpha = 0.5$ yields optimal outcomes for systems
with diagonal state matrices, such as in S4D and S5 configurations.

This frequency-selective masking is proven to be a cornerstone for the adaptability of our SSMs, significantly contributing to their generalization across differing frequencies.

3.3.2 H₂ Norm

This section introduces another strategy to mitigate the aliasing problem by suppressing the frequency response of the continuous-time state-space model beyond a selected frequency \( \omega_{\text{min}} \).

This approach makes use of the \( H_2 \) norm of a continuous-time linear system which measures the power (or steady-state covariance) of the output response to unit white-noise input.

Given a continuous-time system described by matrices \( A, B, \) and \( C \), the transfer function \( G(s) \) from the Laplace transform can be defined in the frequency domain as:

\[
G(j\omega) = C(j\omega I - A)^{-1}B,
\]

where \( \omega \) denotes the frequency, and \( I \) is the identity matrix. The \( H_2 \) norm is computed as the integral of the squared magnitude of the frequency response over the range of interest, typically the entire frequency spectrum. However, in our case, we would like to suppress the frequency response of the system to frequencies beyond \( \omega_{\text{min}} \) which can be done by minimizing the following:

\[
\|G\|_{H_2(\omega_{\text{min}}, \infty)} = \sqrt{\frac{1}{\pi} \int_{\omega_{\text{min}}}^{\infty} \|G(j\omega)\|_F^2 \ d\omega},
\]

as part of the loss function, where \( \|\cdot\|_F \) is the Frobenius norm. In practice, to numerically estimate this integral, we choose a maximum frequency \( \omega_{\text{max}} \) and discretize the frequency range \( [\omega_{\text{min}}, \omega_{\text{max}}] \) into \( N \) points and apply numerical integration methods to the squared Frobenius norm of \( G(j\omega) \). This yields an approximate \( H_2 \) norm of the system in the desired frequency range.

4. Experiments

We perform ablation studies and systematic evaluations of our proposed models utilizing both the Gen1 [8] and 1 Mpx [34] event-based camera datasets. We assess the model’s proficiency in adapting to unseen inference frequencies during training time on both datasets. We employed two variants for training across both datasets: the base model ViT-SSM-B, alongside its scaled-down derivative ViT-SSM-S (small). Additionally, to study the robustness and generalization capabilities of our architecture, we subject it to empirical testing on the DSEC dataset [14], which was not part of the original training corpus and the visual results are provided in the appendix.

4.1. Setup

Implementation details. Our models are trained using 32-bit precision arithmetic across 400k iterations, making use of the ADAM optimizer [25] and a OneCycle learning rate schedule [42] that decays linearly from its highest value. We adopt a mixed batching technique that applies standard Backpropagation Through Time (BPTT) to half of the batch samples and Truncated BPTT (TBPTT) to the other half. This technique was introduced in the RVT [13] and ablated in their supplementary material. We found it to be equally effective also for SSMs. For the S4(D) method, details can be found in [18, 19]. As for S5 [41], we are the first to develop an efficient implementation from scratch in PyTorch; existing public versions of S5 are only available in JAX. Data augmentation is carried out through random horizontal flips and zooming operations, both inward and outward. Event representations are formed based on 50 ms time windows that correspond to 20 Hz sampling frequency, divided into \( T = 10 \) discrete bins. We incorporate a YOLOX detection head [11] that integrates IOU, class, and regression losses, averaged across both batch and sequence lengths during each optimization phase. Bandlimiting is implemented during training and inference of the SSM model by masking the output matrix \( C \) as explained in 3.3.1. In case of \( H_2 \) norm, we add it as term to the loss function of the model. We conduct training on the GEN1 dataset using A100 GPU, employing a batch size of 8 and a sequence length of 21. We use a global learning rate of \( 2e^{-4} \), which is propagated to the SSM components. For the 1 Mpx dataset, we train with a batch size of 24 and a sequence length of 5, using a learning rate of \( 3.5e^{-4} \) across two A100 GPUs.

Datasets. The Gen1 Automotive Detection dataset [8] comprises 39 hours of event camera footage with a resolution of 304 × 240 pixels. It includes 228k bounding boxes for cars and 28k for pedestrians. Using original evaluation criteria [8], we discard bounding boxes having a side length shorter than \( 10 \) pixels or a diagonal less than \( 30 \) pixels.

The 1 Mpx dataset [34] similarly focuses on driving environments but offers recordings at a higher resolution of 720 × 1280 pixels, captured over multiple months during both day-time and night-time. It contains roughly 15 hours of event data and accumulates around 25 million bounding box labels, spread across three classes: cars, pedestrians, and two-wheelers. Following original evaluation criteria, we eliminate bounding boxes with a side length under 20 pixels and a diagonal less than 60 pixels, while also reducing the input resolution to 640 × 360 pixels.

4.2. Benchmark comparisons

In this section, we present a detailed comparison of our proposed methods, S4D-ViT-B and S5-ViT-B, with the SotA approaches in the domain of event-based vision for object detection, which can be seen in Table 1. The evalua-
tion is conducted on two distinct datasets: Gen1 and 1 Mpx, to assess the performance and robustness of these methods under varying conditions.

The comparative analysis, as summarized in Table 1, encompasses different backbones and detection heads. These include Sparse CNNs, GNNs, SNNs, RNNs, and various implementations of Transformers.

Our S4D-ViT-B and S5-ViT-B models, which integrate SSMs with Attention (3.2), demonstrate competitive performance across both datasets. Specifically, on the 1 Mpx dataset, our models achieve mAP scores of 46.8 and 47.2, respectively, which are competitive with the leading scores in this benchmark. While our models do not outperform the top-performing ERGO-12 [47] and GET-T [33] in terms of mAP, they exhibit a notable balance between accuracy and model complexity, as indicated by their parameter counts of 16.5M and 17.5M, respectively.

Notably, our models perform consistently well across different frequencies as it can be seen in Table 2, highlighting their robustness and generalizability in comparison to RVT and GET-T, which tend to exhibit large performance drops. We achieve an average performance drop between training and testing frequencies of 3.31 mAP averaged on both Gen1 [8] and 1 Mpx [34] datasets, while RVT [13] and GET [33] have a drop of 21.25 and 24.53 mAP, respectively.

In summary, our proposed S4D-ViT-B and S5-ViT-B models establish themselves as effective and efficient contenders in the field of event-based vision for object detection. Their competitive performance metrics, coupled with their generalization capabilities across various inference frequencies, make them valuable contributions to event-based vision community.

### Table 1. Comparisons on test sets of Gen1 and 1 Mpx datasets (20 Hz).

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Detection Head</th>
<th>Gen1</th>
<th>1 Mpx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mAP</td>
<td>Time (ms)</td>
<td>mAP</td>
</tr>
<tr>
<td>Anynet [32]</td>
<td>Sparse CNN</td>
<td>YOLOv1 [36]</td>
<td>14.5</td>
<td>-</td>
</tr>
<tr>
<td>AEGNN [40]</td>
<td>GNN</td>
<td>YOLOv1</td>
<td>16.3</td>
<td>-</td>
</tr>
<tr>
<td>Inception + SSD [22]</td>
<td>CNN</td>
<td>SSD</td>
<td>30.1</td>
<td>19.4</td>
</tr>
<tr>
<td>RRC-Events [4]</td>
<td>CNN</td>
<td>YOLOv3 [35]</td>
<td>30.7</td>
<td>21.5</td>
</tr>
<tr>
<td>YOLOv3 Events [24]</td>
<td>CNN</td>
<td>YOLOv3</td>
<td>31.2</td>
<td>22.3</td>
</tr>
<tr>
<td>RED [34]</td>
<td>CNN + RNN</td>
<td>SSD</td>
<td>40.0</td>
<td>16.7</td>
</tr>
<tr>
<td>ERGO-12 [47]</td>
<td>Transformer</td>
<td>YOLOv6 [26]</td>
<td>50.4</td>
<td>69.9</td>
</tr>
<tr>
<td>Swin-T v2 [30]</td>
<td>Transformer</td>
<td>YOLOX</td>
<td>45.5</td>
<td>26.6</td>
</tr>
<tr>
<td>Nested-T [46]</td>
<td>Transformer</td>
<td>YOLOX</td>
<td>46.3</td>
<td>25.9</td>
</tr>
<tr>
<td>GET-T [33]</td>
<td>Transformer</td>
<td>YOLOX</td>
<td>47.9</td>
<td>16.8</td>
</tr>
<tr>
<td>S4D-ViT-B (ours)</td>
<td>Transformer + SSM</td>
<td>YOLOX</td>
<td>46.2</td>
<td>9.40</td>
</tr>
<tr>
<td>S5-ViT-B (ours)</td>
<td>Transformer + SSM</td>
<td>YOLOX</td>
<td>47.4</td>
<td>8.16</td>
</tr>
</tbody>
</table>

Table 2. Evaluation of RVT [13], S5 [41], and GET [33] across different frequencies on test datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Frequency Evaluation (Hz)</th>
<th>Perf. Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVT</td>
<td>Gen1</td>
<td>20 Hz, 40 Hz, 80 Hz, 100 Hz, 200 Hz</td>
<td>26.14</td>
</tr>
<tr>
<td></td>
<td>1Mpx</td>
<td>47.16, 35.13, 31.98, 18.61, 8.35</td>
<td>26.14</td>
</tr>
<tr>
<td>S5</td>
<td>Gen1</td>
<td>47.40, 42.51, 45.08, 42.49, 39.84</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>1Mpx</td>
<td>47.20, 46.49, 45.88, 44.29, 41.39</td>
<td>2.68</td>
</tr>
<tr>
<td>GET</td>
<td>Gen1</td>
<td>47.90, 34.15, 19.97, 15.13, 5.35</td>
<td>29.25</td>
</tr>
<tr>
<td></td>
<td>1Mpx</td>
<td>48.40, 40.51, 30.30, 28.11, 15.44</td>
<td>19.81</td>
</tr>
</tbody>
</table>
4.3.2 SSM Utilization Analysis

In this ablation study, we examine the impact of employing temporal recurrence exclusively in a subset of network stages, or not using it at all. Our methodology involves manipulating the SSMs within the network by resetting their states at predetermined stages during each timestep. This approach enables us to isolate and evaluate the impact of SSM layers’ presence or absence while maintaining a consistent parameter count across different model configurations.

Table 4 shows the outcomes of these manipulations on the Gen1 validation dataset. \(^{S4D_5}\) represents S4D model, and \(^{S5_{5,5}}\) is S5 model with \(\alpha = 0.5\). The data clearly indicate that the complete removal of SSMs from the network leads to a highest decrease in detection performance. This underscores the pivotal role of SSMs in enhancing the model’s capability. On the other hand, initiating the use of SSMs from the fourth stage onwards consistently enhances performance, suggesting a critical threshold for the impact of temporal information processing in the later stages of the network. Another intriguing observation is the performance boost obtained by incorporating an SSM at the very initial stage. This suggests that the early integration of temporal information is also beneficial for the overall detection performance. As a result of these findings, our preferred configuration includes the SSM component right from the initial stage, thereby leveraging the advantages of temporal information processing throughout the entire network.

Noteworthy is the observation that the performance drop when employing SSMs is less pronounced than with RVT, suggesting our approach’s superior robustness to temporal aggregation from certain stages compared to RVT.

4.3.3 Evaluation at different frequencies

In this section, we delve into the comparative performance analysis of three models: RVT \([13]\), S4D \([18]\), and S5 \([41]\), across various frequency ranges. This evaluation is crucial in determining the robustness and adaptability of these models under diverse inference frequencies, specifically at 20 Hz, 40 Hz, 80 Hz, 100 Hz, and 200 Hz. Our model is trained with event representations formed based on 50 ms time windows that correspond to 20 Hz sampling frequency. Evaluation is done at this frequency, and others above mentioned. With RVT, there are no changes when evaluating at different frequencies, while with SSMs rate \(r\) is halved for the double the inference frequency etc.

Each model in Table 5 is assessed across datasets and model sizes, including \(\text{Gen1}_{\text{Base}}, \text{Gen1}_{\text{Small}}, \text{and} \) 1 Mpix\(_{\text{Base}}\). \(\text{Base}\) and \(\text{Small}\) represent two variants of mod-
<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset/Size</th>
<th>Strategy</th>
<th>Frequency Evaluation (Hz)</th>
<th>Performance Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>20 Hz</td>
<td>40 Hz</td>
</tr>
<tr>
<td>RVT [13]</td>
<td>Gen1&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>-</td>
<td>49.52</td>
<td>37.16</td>
</tr>
<tr>
<td></td>
<td>Gen1&lt;sub&gt;Small&lt;/sub&gt;</td>
<td>-</td>
<td>48.68</td>
<td>35.28</td>
</tr>
<tr>
<td></td>
<td>1Mpx&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>-</td>
<td>45.95</td>
<td>40.93</td>
</tr>
<tr>
<td>S4D [18]</td>
<td>Gen1&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>$H_2$ norm</td>
<td>46.83</td>
<td>45.98</td>
</tr>
<tr>
<td></td>
<td>Gen1&lt;sub&gt;Small&lt;/sub&gt;</td>
<td>$H_2$ norm</td>
<td>45.88</td>
<td>45.11</td>
</tr>
<tr>
<td></td>
<td>1Mpx&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>$H_2$ norm</td>
<td>46.66</td>
<td>45.85</td>
</tr>
<tr>
<td>S5 [41]</td>
<td>Gen1&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>$H_2$ norm</td>
<td>48.60</td>
<td>47.11</td>
</tr>
<tr>
<td></td>
<td>Gen1&lt;sub&gt;Small&lt;/sub&gt;</td>
<td>$H_2$ norm</td>
<td>47.33</td>
<td>46.36</td>
</tr>
<tr>
<td></td>
<td>1Mpx&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>$H_2$ norm</td>
<td>47.93</td>
<td>46.78</td>
</tr>
</tbody>
</table>

Table 5. Evaluation of RVT, S4D, and S5 across different frequencies on validation datasets.

Tableels on the Gen1 and 1 Mpx datasets, the base one being the larger one with 16.5M parameters for the S4D model and 17.5M parameters for the S5 model (as presented in Table 1 parameters’ column), and small one with 8.8M parameters for S4D and 9.7M parameters for S5 model. The focal point of this analysis is the performance drop, calculated as the average difference between the original performance at 20 Hz and performances at higher frequencies. A notable aspect of this study is the inherent advantage of the S4D and S5 models due to their incorporation of a learnable timescale parameter. This significantly enhances their adaptability, allowing them to dynamically adjust to varying frequencies. This feature is particularly salient in the S4D and S5 models, which are further analyzed based on different operational strategies: the $H_2$ norm and bandlimiting with $\alpha = 0.5$. The inclusion of the learnable timescale parameter in these models underscores their capability to maintain performance across a wide range of frequencies.

The analysis reveals that both the $H_2$ norm and bandlimiting strategies with $\alpha = 0.5$ offer comparable performance across the assessed frequency ranges. However, a slight edge is observed with the $H_2$ norm, particularly at very high frequencies. This marginal superiority can be attributed to the fact that $H_2$ norm approach does not explicitly mask output’s ($C$) matrix columns.

In the Appendix, we study pure-SSM and SSM models in combination with ConvNext [31].

5. Conclusion

In this paper, we presented a novel approach for enhancing the adaptability and training efficiency of models designed for event-based vision, particularly in object detection tasks. Our methodology leverages the integration of SSMs with a Vision Transformer (ViT) architecture, creating a hybrid SSM-ViT model. This integration not only addresses the long-standing challenge of performance degradation at varying temporal frequencies but also significantly accelerates the training process.

The key innovation of our work lies in the use of learnable timescale parameters within the SSMs, enabling the model to adapt dynamically to different inference frequencies without necessitating multiple training cycles. This feature represents a substantial advancement over existing methods, which require retraining for different frequencies.

The SSM-ViT model outperforms existing methods by 20 mAP at higher frequencies and exhibits a 33% increase in training speed. Furthermore, our introduction of two novel strategies to counteract the aliasing effect (a crucial consideration in high-frequency deployment) further reinforces the model’s suitability for real-world applications. These strategies, involving frequency-selective masking and $H_2$ norm adjustments, effectively mitigate the adverse effects of aliasing, ensuring the model’s reliability across a spectrum of temporal resolutions. We believe that our approach opens new avenues for research and application in high-speed, dynamic visual environments, setting a new benchmark in the domain of event-based vision.

6. Acknowledgment

This work was supported by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) Robotics (grant number 51NF40_185543), and the European Research Council (ERC) under grant agreement No. 864042 (AGILE-FLIGHT).
State Space Models for Event Cameras
Supplementary Material

Contents:

- Appendix 7: Explanation of S4(D) and S5 models.
- Appendix 8: Initialization of continuous-time matrices.
- Appendix 9: Background on Parallel Scans for Linear Recurrences.
- Appendix 10: DSEC dataset evaluation.
- Appendix 11: PyTorch implementation of Parallel Scan and S5 model.
- Appendix 12: Exploring the S4 and S5 Architectural Link.
- Appendix 13: SSM-only and SSM-ConvNext [31] models.

7. Explanation of S4(D) and S5 models

The Structured State Spaces (S4 and its diagonal variant S4D [19, 21]) and Simplified State Space (S5) [41] models are advanced approaches in sequence modeling, each with its unique characteristics.

S4 introduces a sequence model based on the Structured State Space Model (SSM) [16, 17, 20]. It addresses computational bottlenecks in previous work [17] by reparameterizing structured state matrices to maintain a hidden state that encodes the long history of input. S4D [21] is a variant of the S4 model, simplifying it by using a fully diagonal state matrix. This adaptation preserves the performance of the original model but with a simpler implementation. Gupta [21] observed that removing the low-rank part of S4’s HiPPO-LegS matrix [16] results in a diagonal matrix (referred to as the normal-HiPPO matrix) with comparable performance to the original S4. S5 [41] is an evolution of S4, using a multi-input, multi-output (MIMO) single SSM instead of multiple single-input, single-output (SISO) SSMs used in S4 [19].

The HiPPO matrix [16], a non-normal matrix, is decomposed as a sum of a normal and a low-rank matrix. S4 applies new techniques to overcome computational limitations associated with this decomposition. The S5 model simplifies the S4 layer by replacing the bank of many independent SISO SSMs with one MIMO SSM and implementing efficient parallel scans [1]. This shift eliminates the need for convolutional and frequency-domain approaches, making the model purely recurrent and time-domain based.

S4 equates the diagonal matrix case to the computation of a Cauchy kernel, applying to any matrix that can be decomposed as Normal Plus Low-Rank (NPLR). S5 utilizes a diagonal state matrix for efficient computation using parallel scans. It inherits HiPPO initialization schemes from S4, using a diagonal approximation to the HiPPO matrix for comparable performance.

S5 matches the computational complexity of S4 for both online generation and offline recurrence. It handles time-varying SSMs and irregularly sampled observations, which are challenges for the convolutional implementation in S4. S5 has been shown to match or outperform S4 in various long-range sequence modeling tasks, including speech classification and 1-D image classification [41]. This design opens up new possibilities in deep sequence modeling, including handling time-varying SSMs and combining state space layers with attention mechanisms for enhanced performance, which is our contribution along with specific problem and architecture design.

In summary, while S4 introduced a novel reparameterization of state space models for efficient long sequence modeling, S5 builds upon this by simplifying the model structure and computation, leading to a more usable and potentially more flexible approach for various sequence modeling tasks. Both S4 and S5 models can be seen on Figure 3 and Figure 4 respectively.

8. Initialization of continuous-time matrices

8.1. State Matrix Initialization

This subsection elaborates on the initialization process of continuous-time matrices which are crucial since they allow us to discretize them with different time steps to deploy our model at higher inference frequencies. As described by [16], S4’s capacity to handle long-range dependencies stems from employing the HiPPO-LegS matrix, which decomposes the input considering an infinitely long, exponentially diminishing measure [17, 19]. The HiPPO-LegS matrix and its corresponding single-input-single-output (SISO) vector are defined as:

\[
(A_{\text{LegS}})_{nk} = \begin{cases} 
(2n + 1)^{1/2}(2k + 1)^{1/2}, & n > k \\
(n + 1), & n = k \\
0, & n < k 
\end{cases} \quad (8)
\]

Here, the input matrix \(B_{\text{LegS}} \in \mathbb{R}^{N \times H}\) is constructed by concatenating \(b_{\text{LegS}} \in \mathbb{R}^N \times H\) times.

Theorem 1 of Gu et al. [17] establishes that HiPPO matrices from [16], \(A_{\text{HiPPO}} \in \mathbb{R}^{N \times N}\), can be represented in a normal plus low-rank (NPLR) format, comprising a normal matrix, \(A_{\text{Normal}} \in \mathbb{R}^{N \times N}\), and a low-rank
Figure 3. In the S4 layer, each dimension of the input sequence $u_{1:L} \in \mathbb{R}^{L \times H}$ is processed by a separate SSM. This process involves using a Cauchy kernel to determine the coefficients for frequency domain convolutions. The convolutions, done via FFTs, generate the output $y_{1:L} \in \mathbb{R}^{L \times H}$ for each SSM. The outputs then go through a nonlinear activation function, which includes a layer that mixes them to produce the final output of the layer.

Figure 4. For the S5 layer, a parallel scan technique is employed on a diagonal linear SSM to get the SSM outputs. This approach bypasses the need for frequency domain operations and convolution kernel computations required by S4, resulting in a model that functions in a purely time-domain, recurrent manner. Recurrence is parallelized with the employment of parallel scans [1].

Component:

$$A_{\text{HiPPO}} = A_{\text{Normal}} - PQ^\top = A_{\text{HiPPO}}$$

with unitary $V \in \mathbb{C}^{N \times N}$, diagonal $\Lambda \in \mathbb{C}^{N \times N}$, and low-
rank factors $P, Q \in \mathbb{R}^{N \times r}$. This equation reveals the conjugation of HiPPO matrices into a diagonal plus low-rank (DPLR) structure. Therefore, the HiPPO-LegS matrix can be expressed through the normal HiPPO-N matrix and the low-rank term $P_{\text{LegS}} \in \mathbb{R}^N$, as suggested by Goel et al. [15]:

$$A_{\text{LegS}} = A_{\text{LegS}}^{\text{Normal}} - P_{\text{LegS}}P_{\text{LegS}}^T \quad (12)$$

where

$$A_{\text{LegS}}^{\text{Normal}} = \begin{cases} \frac{1}{2}(n + \frac{1}{2})^{1/2}(k + \frac{1}{2})^{1/2}, & n > k \\ \frac{1}{2}, & n = k \\ \frac{1}{2}(n + \frac{1}{2})^{1/2}(k + \frac{1}{2})^{1/2}, & n < k \end{cases} \quad (13)$$

$$P_{\text{LegS}} = (n + 1) \frac{1}{2} \quad (14)$$

We default to initializing the S5 layer state matrix $A$ as $A_{\text{LegS}}^{\text{Normal}} \in \mathbb{R}^{P \times P}$ and then decompose it to obtain the initial $A$. We often find benefits in initializing $\tilde{B}$ and $\tilde{C}$ using $V$ and its inverse, as detailed below.

Performance improvements on various tasks were noted with the S5 state matrix initialized as block-diagonal [41], with each diagonal block equaling $A_{\text{LegS}}^{\text{Normal}} \in \mathbb{R}^{R \times R}$ (where $R < P$). This initialization process also involves decomposing the matrix to obtain $A$, as well as $B$ and $C$. Even in this case, $B$ and $C$ remain densely initialized without constraints to keep $A$ block-diagonal during learning. The hyperparameter $J$ [41] denotes the number of HiPPO-N blocks on the diagonal for initialization, with $J = 1$ indicating the default single HiPPO-N matrix initialization. Further discussion on the rationale behind this block-diagonal approach is in Appendix 12.3.

8.2. Input, Output, and Feed-through Matrices Initialization

The input matrix $\tilde{B}$ and output matrix $\tilde{C}$ are explicitly initialized using the eigenvectors from the diagonalization of the initial state matrix. We start by sampling $B$ and $C$ and then set the complex learnable parameters $\tilde{B} = V^{-1}B$ and $\tilde{C} = CV$.

The feed-through matrix $D \in \mathbb{R}^{H}$ is initialized by sampling each element independently from a standard normal distribution.

8.3. Timescale Initialization

Prior research, notably [21] and [20], has underscored the significance of initializing timescale parameters. Explored in detail by Gu et al. [20], we align with S4 practices and initialize each element of $\log \Delta \in \mathbb{R}^P$ from a uniform distribution over the interval $[\log \delta_{\text{min}}, \log \delta_{\text{max}}]$, with default values set to $\delta_{\text{min}} = 0.001$ and $\delta_{\text{max}} = 0.1$.

9. Background on Parallel Scans for Linear Recurrences

This section provides a concise introduction to parallel scans, which are central element for the parallelization of the S5 method implemented in 11.1. For an in-depth exploration, refer to [1].

In essence, a scan operation takes a binary associative operator $\bullet$, adhering to the property $(ab)c = (a\bullet b)c$, and a sequence of $L$ elements $[a_1, a_2, ..., a_L]$. The output is the cumulative sequence: $[a_1, (a_1 \bullet a_2), ..., (a_1 \bullet a_2 \bullet ... \bullet a_L)]$.

The principle behind parallel scans is leveraging the flexible computation order of associative operators. For linear recurrences, parallel scans are applicable on the following state update equations:

$$x_k = \Lambda x_{k-1} + \tilde{B} u_k, \quad y_k = \tilde{C} x_k + \tilde{D} u_k \quad (15)$$

We initiate the process with scan tuples $c_k = (c_{k,a}, c_{k,b}) := (\Lambda, \tilde{B} u_k)$. The binary associative operator, applied to tuples $q_i, q_j$ (initial or intermediate), generates a new tuple $q_i \bullet q_j := (q_i \circ q_j, q_i \circ q_j \circ q_i \circ q_j + q_j)$, where $\circ$ and $\otimes$ represent matrix-matrix and matrix-vector multiplication, respectively. With adequate processing power, this parallel scan approach can compute the linear recurrence of (15) in $O(\log L)$ sequential steps, thereby significantly enhancing computational efficiency [1].

10. DSEC dataset evaluation

To test the generalizability of our model, we perform inference on the DSEC dataset [14] with a model trained on 1 Mpx dataset [34]. Some results are shown below:

![Detections on DSEC dataset](image1)

Figure 5. Detections on DSEC dataset [14] with model trained on 1 Mpx dataset [34]. Names of the specific DSEC scenes are in the subcaptions.
11. PyTorch implementation of Parallel Scan and S5 model

In the following subsections, we give implementations of parallel scans and S5 layer in PyTorch.

11.1. Parallel Scan Operation

Listing 1. Implementation of efficient parallel scan used in S5 model.

```python
from typing import Callable
import torch
from torch.utils._pytree import tree_flatten, tree_unflatten
from functools import partial

def safe_map(f, *args):
    args = list(map(list, args))
    n = len(args[0])
    return list(map(f, *args))

def combine(tree, operator, a_flat, b_flat):
    a = tree_unflatten(a_flat, tree)
    b = tree_unflatten(b_flat, tree)
    c = operator(a, b)
    c_flat, _ = tree_flatten(c)
    return c_flat

def _scan(tree, operator, elems, axis:int):
    num_elems = elems[0].shape[axis]
    if num_elems < 2:
        return elems
    reduced_elements = combine(tree, operator,
                               [torch.ops.aten.slice(elem, axis, 0, -1, 2) for elem in elems],
                               [torch.ops.aten.slice(elem, axis, 1, None, 2) for elem in elems],
                               )
    # Recursively compute scan for partially reduced tensors.
    odd_elements = _scan(tree, operator, reduced_elements, axis)
    if num_elems % 2 == 0:
        even_elements = combine(tree, operator, odd_elements,
                                 [torch.ops.aten.slice(e, axis, 2, None, 2) for e in elems],
                                 )
    else:
        even_elements = combine(tree, operator, odd_elements,
                                 [torch.ops.aten.slice(e, axis, 2, None, 2) for e in elems],
                                 )
    # The first element of a scan is the same as the first element of the original 'elems'.
    even_elements = [torch.cat([torch.ops.aten.slice(elem, axis, 0, 1), result], dim=axis)
                     if result.shape.numel() > 0 and elem.shape[axis] > 0
                     else result
                     for elem, result in zip(elems, even_elements)]
    return tree_unflatten(safe_map(partial(_interleave, axis=axis), even_elements, odd_elements))

def associative_scan(operator: Callable, elems, axis:int = 0, reverse: bool = False):
    elements, tree = tree_flatten(elems)
    num_elements = int(elements[0].shape[axis])
    if reverse:
        elements = [torch.flip(elem, [axis]) for elem in elements]
    scans = _scan(tree, operator, elements, axis)
    return tree_unflatten(scans, tree)
```
11.2. S5 model

Listing 2. PyTorch implementation to apply a single S5 layer to a batch of input sequences.

```python
import torch
from typing import Tuple

def discretize_bilinear(Lambda, B_tilde, Delta):
    """ Discretize a diagonalized, continuous-time linear SSM using bilinear transform method.
    Args:
        Lambda (float32): diagonal state matrix (P, 2)
        B_tilde (complex64): input matrix (P, H)
        Delta (float32): discretization step sizes (P,)
    Returns:
        discretized Lambda_bar (float32), B_bar (float32) (P, 2), (P, H, 2)
    """
    Lambda = torch.view_as_complex(Lambda)
    Identity = torch.ones(Lambda.shape[0], device=Lambda.device)
    BL = 1 / (Identity - (Delta / 2.0) * Lambda)
    Lambda_bar = BL * (Identity + (Delta / 2.0) * Lambda)
    B_bar = (BL * Delta)[..., None] * B_tilde
    Lambda_bar = torch.view_as_real(Lambda_bar)
    B_bar = torch.view_as_real(B_bar)
    return Lambda_bar, B_bar

@torch.jit.script
    """Binary operator for parallel scan of linear recurrence. Assumes a diagonal matrix A.
    Args:
        q_i: tuple containing A_i and Bu_i at position i (P,), (P,)
        q_j: tuple containing A_j and Bu_j at position j (P,), (P,)
    Returns:
        new element ( A_out, Bu_out )
    """
    A_i, b_i = q_i
    A_j, b_j = q_j
    return A_j * A_i, torch.addcmul(b_j, A_j, b_i)

def apply_ssm(Lambda_bars: torch.Tensor, B_bars, C_tilde, D, input_sequence, prev_state, bidir: bool = False):
    """ Compute the LxH output of discretized SSM given an LxH input.
    Args:
        Lambda_bars (float32): discretized diagonal state matrix (P, 2)
        B_bars (float32): discretized input matrix (P, H, 2)
        C_tilde (float32): output matrix (H, P, 2)
        input_sequence (float32): input sequence of features (L, H)
        prev_state (complex64): hidden state (H,)
    Returns:
        ys (float32): the SSM outputs (S5 layer preactivations) (L, H)
        xs (complex64): hidden state (H,)
    """
    B_bars, C_tilde, Lambda_bars = as_complex(B_bars), as_complex(C_tilde), as_complex(Lambda_bars)
    cinput_sequence = input_sequence.type(Lambda_bars.dtype) # Cast to correct complex type
    Bu_elements = torch.vmap(lambda u: B_bars @ u)(cinput_sequence)
    if Lambda_bars.ndim == 1: # Repeat for associative_scan
        Lambda_bars = Lambda_bars.tile(input_sequence.shape[0], 1)
    _, xs = associative_scan(binary_operator, (Lambda_bars, Bu_elements))
    if bidir:
        _, xs2 = associative_scan(binary_operator, (Lambda_bars, Bu_elements), reverse=True)
        xs = torch.cat((xs, xs2), axis=-1)
    Du = torch.vmap(lambda u: D * u)(input_sequence)
    return torch.vmap(lambda x: (C_tilde @ x).real)(xs) + Du, xs[-1]
```
Listing 3. Raw S5 operator that can be instantiated and used as block in any architecture.

```python
import torch

class S5(torch.nn.Module):
    def __init__(...):
        self.seq = S5SSM(...)
        ...

    def forward(self, signal, prev_state, step_scale: float | torch.Tensor = 1.0):
        if not torch.is_tensor(step_scale):
            # Duplicate across batchdim
            step_scale = torch.ones(signal.shape[0], device=signal.device) * step_scale
        return torch.vmap(
            lambda s, ps, ss: self.seq(s, prev_state=ps, step_scale=ss))(signal, prev_state, step_scale)

class S5SSM(torch.nn.Module):
    def __init__(...):
        self.degree = degree
        self.discretize = self.discretize_bilinear
        ...

    def get_BC_tilde(self):
        match self.bcInit:
            case "dense_columns" | "dense" | "complex_normal":
                B_tilde = as_complex(self.B)
                C_tilde = self.C
            case "factorized":
                B_tilde = self.BP @ self.BH.T
                C_tilde = self.CH.T @ self.CP
        return B_tilde, C_tilde

    def forward(self, signal, prev_state, step_scale: float | torch.Tensor = 1.0):
        B_tilde, C_tilde = self.get_BC_tilde()
        if self.degree != 1:
            assert (B_bar.shape[-2] == B_bar.shape[-1]), "higher-order input operators must be full-rank"
            B_bar **= self.degree
        step = step_scale * torch.exp(self.log_step)
        Lambda_bars, B_bars = self.discretize(self.Lambda, B_tilde, step)
        return apply_ssm(Lambda_bars, B_bars, C_tilde, self.D, signal, prev_state, bidir=self.bidir)
```

12. Exploring the S4 and S5 Architectural Link

This section delves into the relationship between the S4 and S5 architectures, which is instrumental in the evolution of more efficient architectures and the expansion of theoretical insights from preceding research [16, 19, 41].

Our analysis is segmented into three distinct parts:

1. We utilize the linear nature of these systems to explain that the latent states generated by the S5 SSM are effectively a linear combination of those produced by the H SISO S4 SSMs. Moreover, the outputs of the S5 SSM represent an additional linear transformation of these states (12.2).

2. For the SISO scenario, as N becomes significantly large, the dynamics derived from a (non-diagonalizable) HiPPO-LegS matrix can be accurately approximated by the (diagonalizable) normal component of the HiPPO-LegS matrix. This is expanded to encompass the MIMO context, providing a rationale for initializing with the HiPPO-N matrix and consequently enabling efficient parallel scans (12.3).

3. We conclude with a discussion that S5, through strategic initialization of its state matrix, can emulate multiple independent S4 systems, thus easing previously held assumptions (12.3). We also discuss the set of timescale parameters, which have been observed to enhance performance (12.4).

It is important to note that many of these results are derived directly from the linearity of the recurrence.

12.1. Underlying Assumptions

The forthcoming sections are predicated on the following assumptions unless stated otherwise:

- We exclusively consider sequence maps that are H-dimensional to H-dimensional.
- The state matrix for each S4 SSM is the same, denoted as A(\(h\)) = A \in \mathbb{C}^{N \times N}.
- We assert that the timescales for each S4 SSM are consistent, represented by \(\Delta(\(h\)) = \Delta \in \mathbb{R}_+\).
- S5 employs the identical state matrix A as in S4 [19], implying that the S5’s latent size P is such that P = N. Additionally, it is presumed that the S5 input matrix is a horizontal concatenation of the S4’s column input vectors, expressed as B \(\triangleq [B(1) \mid \ldots \mid B(H)]\).

12.2. Distinct Output Projections from Equivalent Dynamics

In an S5 layer characterized by state matrix A, input matrix B, and an output matrix C, and an S4 layer, where each of the H individual S4 SSMs possesses a state matrix A and input vector B(\(h\)), the outputs of the S5 SSM, \(y_k\), are equivalent to a linear combination of the latent states from the H S4 SSMs, \(y_k = C^\text{equiv} x_k^{(1:H)}\), where C^equiv = [ C \ldots C ], provided that both layers are discretized using identical timescales.

Proof. Considering a singular S4 SSM, the discretized latent states in relation to the input sequence \(u_{i:L} \in \mathbb{R}^{L \times H}\) are described as:

\[
x_k^{(h)} = \sum_{i=1}^{k} \Delta^{k-i-1} B u_i^{(h)}.
\]

For the S5 layer, the latent states are given by:

\[
x_k = \sum_{i=1}^{k} \Delta^{k-i-1} B u_i,
\]

where B is defined as B \(\triangleq [B(1) \mid \ldots \mid B(H)]\) and u, is \([u_i(1), \ldots, u_i(H)]^\top\).

Observation leads to:

\[
x_k = \sum_{h=1}^{H} x_k^{(h)},
\]

This outcome directly stems from the linearity of Equations (16) and (17), indicating that the MIMO S5 SSM states are equivalent to the sum of the states from the H SISO S4 SSMs.

The output matrix C for S5 is a singular dense matrix:

\[
y_k = C x_k.
\]

Substituting the relationship from (18) into (19) enables expressing the outputs of the MIMO S5 SSM in terms of the states of the H SISO S4 SSMs:

\[
y_k = C \sum_{h=1}^{H} x_k^{(h)} = \sum_{h=1}^{H} C x_k^{(h)}.
\]

Defining the vertical concatenation of the H S4 SSM state vectors as \(x_k^{(1:H)} = [x_k^{(1)^\top}, \ldots, x_k^{(H)^\top}]^\top\), we ascertain that the S5 SSM outputs can be written as:

\[
y_k = C^\text{equiv} x_k^{(1:H)}, \quad \text{with } C^\text{equiv} = [ C \ldots C ],
\]

thereby confirming their equivalence to a linear combination of the HN states computed by the H S4 SSMs.

This proof demonstrates that the outputs of the S5 SSM, under the specified constraints, can be interpreted as a linear combination of the latent states generated by H similarly constrained S4 SSMs, sharing identical state matrices and timescale parameters. However, it does not imply that the outputs of the S5 SSM are directly identical to those of the effective block-diagonal S4 SSM; indeed, they differ, as will be further clarified in the analysis of the S4 layer.

Assuming the output vector for each S4 SSM corresponds to a row in the S5 output matrix, i.e., C =
\[
\begin{bmatrix}
C^{(1)^\top} & \cdots & C^{(H)^\top}
\end{bmatrix}^\top,
\] the output of each S4 SSM can be expressed as:
\[
y_k^{(h)} = C^{(h)}x_k^{(h)},
\] where \(y_k^{(h)} \in \mathbb{R}\). The effective output matrix operating on the entire latent space in S4 is:
\[
y_k^{(h)} = (C^{S4}x_k)^{(h)}
\] By comparing (21) and (23), the distinction in the equivalent output matrices employed by both layers becomes clear:
\[
C^{S4} = \begin{bmatrix}
C^{(1)} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & C^{(H)}
\end{bmatrix},
\]
\[
C^{\text{equiv}} = \begin{bmatrix}
C^{(1)} & \cdots & C^{(1)} \\
\vdots & \ddots & \vdots \\
C^{(H)} & \cdots & C^{(H)}
\end{bmatrix} = [C | \cdots | C].
\]

In S4, the effective output matrix comprises independent vectors on the diagonal, while in S5, it uniformly connects dense output matrices across the \(H\) S4 SSMs. Thus, S5 can be seen as defining a different projection of the \(H\) independent SISO SSMs than S4 does. Both projection matrices possess an identical parameter count.

Despite variations in projection, the interpretability of the latent dynamics in S5 as a linear projection from S4’s latent dynamics suggests a promising approach. This observation leads to the hypothesis that initializing the state dynamics in S5 with the HiPPO-LegS matrix, the same as in the method employed in S4 [19], may yield similarly effective results.

It remains an open question whether one method consistently surpasses the other in terms of expressiveness. It’s also important to underscore that practical implementation of S4 and S5 would not directly utilize the block diagonal matrix and repeated matrix, respectively, as described in Equation 25. These matrices serve primarily as theoretical tools to explain the conceptual equivalence between S4 and S5 models.

12.3. Relaxing the Assumptions

Consider an instance where the S5 SSM state matrix is configured in a block-diagonal form. In such a scenario, an S5 SSM with a latent size of \(JN = O(H)\) would employ a block-diagonal matrix \(A \in \mathbb{R}^{JN \times JN}\), complemented by dense matrices \(B \in \mathbb{R}^{JN \times H}\) and \(C \in \mathbb{R}^{H \times JN}\), and \(J\) distinct timescale parameters \(\Delta \in \mathbb{R}^J\). The latent state \(x_k \in \mathbb{R}^{JN}\) of this system can be divided into \(J\) distinct states \(x_k^{(j)} \in \mathbb{R}^N\). Consequently, this allows the decomposition of the system into \(J\) individual subsystems, with each subsystem discretized using a respective \(\Delta^{(j)}\). The discretization can be represented as:
\[
\begin{align*}
\bar{A} &= \begin{bmatrix}
\bar{A}^{(1)} \\
\vdots \\
\bar{A}^{(J)}
\end{bmatrix}, & \quad \bar{B} &= \begin{bmatrix}
\bar{B}^{(1)} \\
\vdots \\
\bar{B}^{(J)}
\end{bmatrix}, & \quad C &= \begin{bmatrix}
C^{(1)} & \cdots & C^{(J)}
\end{bmatrix},
\end{align*}
\]
where \(\bar{A}^{(j)} \in \mathbb{R}^{N \times N}, \bar{B}^{(j)} \in \mathbb{R}^{N \times H},\) and \(C^{(j)} \in \mathbb{R}^{H \times N}\). This division implies that the system can be interpreted as \(J\) independent \(N\)-dimensional S5 SSM subsystems, with the total output being the sum of the outputs from these \(J\) subsystems:
\[
y_k = Cx_k = \sum_{j=1}^{J} C^{(j)}x_k^{(j)}.
\]

The dynamics of each of these \(J\) S5 SSM subsystems can be correlated to the dynamics of a distinct S4 system. Each of these S4 systems possesses its unique set of tied S4 SSMs, including separate state matrices, timescale parameters, and output matrices. Hence, the outputs of a \(JN\)-dimensional S5 SSM effectively correspond to the linear combination of the latent states from \(J\) different S4 systems. This realization opens the possibility of initializing a block-diagonal S5 state matrix with multiple HiPPO-N matrices across its blocks, rather than a singular, larger HiPPO-N matrix.

12.4. Timescale Parameterization

This section delves into the parameterization nuances of the timescale parameters \(\Delta\). S4 possesses the capability to learn distinct timescale parameters for each S4 SSM [19], thereby accommodating various data timescales. Additionally, the initial setting of these timescales is crucial, as highlighted in the works of [18] and [21]. Relying solely on a single initial parameter might result in suboptimal initialization. The previous discussion in this paper proposes the learning of \(J\) separate timescale parameters, corresponding to each of the \(J\) subsystems. However, empirical evidence indicates superior performance when employing \(P\) distinct timescale parameters [19, 41], one assigned to each state.

This approach can be interpreted in two ways. Firstly, it may be seen as assigning a unique scaling factor to each eigenvalue within the diagonalized system framework. Alternatively, it could be considered a strategy to increase the diversity of timescale parameters at the initialization phase, thereby mitigating the risks associated with inadequate initialization. It is noteworthy that the system has the potential to learn to operate with a singular timescale [41], achieved by equalizing all timescale values.
13. SSM-only and SSM-ConvNext models

To study the performance of SSM-only models, we replaced the attention block with a 2D-SSM block and trained it on Gen1 and 1 Mpx datasets. We obtain 46.1 mAP and 45.74 mAP which is lower than the original S5-ViT-B model’s performance, as it can be seen in Table 6. To evaluate a CNN-based backbone, we replace the attention block with the ConvNext block [31] showing that ViT achieves better performance than the CNN structure.

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP_{Gen1}</th>
<th>mAP_{1Mpx}</th>
</tr>
</thead>
<tbody>
<tr>
<td>S5-ViT-B</td>
<td>47.40</td>
<td>47.20</td>
</tr>
<tr>
<td>S5-ConvNext-B</td>
<td>45.92</td>
<td>45.66</td>
</tr>
<tr>
<td>S5-SSM2D-B</td>
<td>46.10</td>
<td>45.74</td>
</tr>
</tbody>
</table>

Table 6. Comparison of mAP scores for Gen1 and 1 Mpx datasets across different base models.

References

[26] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolín Wei. Yolov6: A single-stage object detection framework for industrial applications, 2022. 6


