Mamba2D: A Natively Multi-Dimensional State-Space Model for Vision Tasks

Enis Baty[1*]   **Alejandro Hernández Díaz**[1*]   Chris Bridges[1]   Rebecca Davidson[2]

Steve Eckersley[2]   Simon Hadfield[1]

Abstract

State-Space Models (SSMs) have recently emerged as a powerful and efficient alternative to the long-standing transformer architecture. However, existing SSM conceptualizations retain deeply rooted biases from their roots in natural language processing. This constrains their ability to appropriately model the spatially-dependent characteristics of visual inputs. In this paper, we address these limitations by re-deriving modern selective state-space techniques, starting from a natively multidimensional formulation. Currently, prior works attempt to apply natively 1D SSMs to 2D data (i.e. images) by relying on arbitrary combinations of 1D scan directions to capture spatial dependencies. In contrast, Mamba2D improves upon this with a single 2D scan direction that factors in both dimensions of the input natively, effectively modelling spatial dependencies when constructing hidden states. Mamba2D shows comparable performance to prior adaptations of SSMs for vision tasks, on standard image classification evaluations with the ImageNet-1K dataset. Source code is available at https://github.com/cocoalex00/Mamba2D.

1. Introduction

The advent of transformer-based architectures has revolutionised the field of computer vision. Since the introduction of the Vision Transformer (ViT) [5], countless works have explored their use across a wide range of visual tasks [36] [3] [12] [15] [20] [21] [2], fundamentally reshaping how image data is processed and interpreted. This rapid success is largely credited to the powerful representational abilities of the attention mechanism, which, unlike traditional convolutional operations, enables the dynamic integration of global context within images. By focusing on relationships across non-neighbouring parts of the input, transformers have proven to excel in capturing long-range dependencies that are critical for many complex tasks.
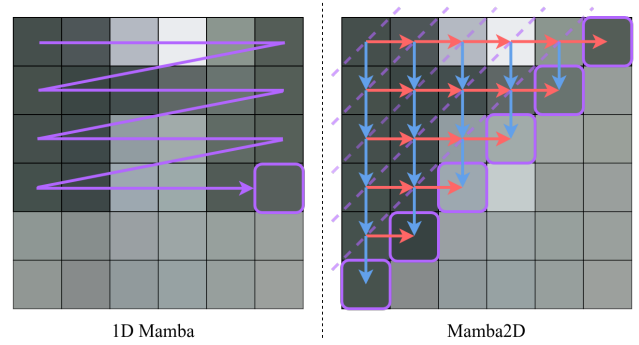

Figure 1. Illustration of a typical 1D Mamba scan (left) and our Mamba2D wavefront scan (right). Our reformulation of the S6/Mamba scan retains the spatial coherency between adjacent pixels or tokens in 2D. Dashed purple lines indicate each diagonal wavefront at which hidden states are computed in parallel.

However, transformer attention exhibits a well-documented quadratic scaling in complexity relative to image size. This behaviour leads to substantial computational overhead which becomes particularly problematic in high-resolution scenarios or dense-prediction tasks i.e. semantic segmentation, depth and optical-flow estimation *etc*.

In response to these challenges, recent work in language modelling has introduced State Space Models (SSMs) as a promising alternative. Preliminary results indicate that SSMs can provide comparable, if not superior, representational power relative to transformers, with the added advantage of linear scaling in complexity w.r.t. the number of input tokens. In particular, the recently proposed Mamba architecture [7] builds on these advantages by incorporating a novel selectivity mechanism that replicates the dynamic global context aggregation of attention, establishing a new state of the art.

These notable achievements have inspired further exploration of the Mamba paradigm within the vision community. Early studies have effectively applied this technique across a wide range of tasks, advancing the field in terms of both efficiency and scalability. However, it is known that the SSM design currently employed in these works carries inherent biases stemming from its origins in language pro-

---

* joint first authorship

1 University of Surrey, UK

2 Surrey Satellite Technology Ltd, UK

cessing. (1) SSMs were proposed for the modelling of 1-dimensional signals, and cannot be natively applied to multidimensional data without reshaping. (2) The causal structure of the architecture, while suited to language where token order is essential, does not align with the predominantly non-causal requirements of vision tasks.

These limitations have been partially addressed in previous works by applying intricate unrolling strategies on the 2-dimensional images, effectively re-shaping them into multiple 1-dimensional sequences. Embeddings are then generated by scanning each sequence individually and subsequently fusing them to capture broader spatial relationships across pixels. While this approach enables SSMs to handle multidimensional data to some extent, it introduces additional complexity and severely disrupts the spatial coherence between neighbouring pixels. In particular, it is also noteworthy that the recurrent memory of SSMs is of a fixed size. This is traditionally characterised (through the HiPPO theory) as providing exponentially decaying recall accuracy based, on the distance from the current token. This makes it impossible for any 1-dimensional flattening of an image to model the correlation patterns within a 2-dimensional neighbourhood.

In contrast, in this work we fundamentally re-design the State-Space paradigm to produce a block which is natively suited to process visual data while preserving the selectivity and global-context characteristics of Mamba. Our novel M2D-SSM block enables effective information flow across both dimensions of a spatial neighbourhood when constructing hidden states. This extension makes it possible to more effectively capture complex spatial dependencies without breaking up the connectivity between neighbouring pixels.

Unfortunately, the inclusion of 2-dimensional dependencies within the layer make it impossible to exploit the traditional convolutional, or parallel scan approaches to efficient computation. Instead, we propose a new efficient wavefront-scan computational model, designed specifically for 2-dimensional data structures.

With the aforementioned concepts, Mamba2D (M2D) posts competitive results for ImageNet-1K top-1 accuracy when compared to recent SSM based models such as VSSD [27] and MambaVision [14] and other established CNN or ViT based models, with good accuracy at low parameter counts.

In summary, to the best of our knowledge, no prior work has explored a native 2D derivation of Mamba for vision based tasks, without extensive workarounds that sacrifice the spatial relationships between pixels. Our primary contributions are as follows: (1) We introduce a novel method of natively applying SSMs to 2D inputs, for use in vision tasks where spatial structure is paramount. (2) We implement a custom M2D-SSM kernel with wavefront-scanning

to avoid intractable sequential computation of hidden states and enable parallelisable training. (3) We demonstrate the enhanced capabilities of Mamba2D in building structurally coherent long range relations within visual data, showing impressive results on the task of image classification with our hybrid Mamba2D/Transformer model.

2. Related Work

State Space Models (SSMs) are architectures recently introduced in the context of Natural Language Processing (NLP), originally inspired by continuous-signal models from traditional control theory. In the context of deep learning, LSSL [9] was one of the first SSMs to show state-of-the-art performance in multiple sequence modelling benchmarks, by employing the HiPPO theory [8] for its initialization. However, due to its state representation's high computation and memory requirements, LSSL was infeasible to use in practice. To mitigate this, Gu et al. [10] introduced the Structured State Space (S4) model, able to compute this state efficiently via a conditioned parametrization of the SSM. At this point, the promise of linearly scaling in sequence length via a recurrent formulation plus excellent parallelisation capabilities inspired further investigation in this area. For instance, Gupta et al. [13] simplified the conditional parametrization of the state matrix introduced by S4. Subsequently, Metha et al. [23] introduced the efficient Gated State Space (GSS) models, increasing speed by a factor of 3 at train time and a factor of 60 at inference time. Finally, Fu et al. [6] bridged the performance gap between Transformers and SSMs in language modelling with their novel H3 layer.

It is also worth mentioning the work of Smith et al. [28] which took a different approach by extending the original single-channel formulation of the SSM to allow for multi-input multi-output processing in its S5 model. However, the applicability of the aforementioned "traditional" state-space models was largely confined to the natural language domain. This boundary was redefined by the pioneering work of Nguyen et al. [24], which expanded their use beyond this scope. Their proposed S4ND layer generalized the convolutional formulation of traditional linear time-invariant (LTI) state-space models to process 2D and 3D signals by leveraging the outer product of multiple one-dimensional kernels.

Recently Gu et al. [7] further refined the State Space paradigm. They successfully addressed one of the key weaknesses of SSMs w.r.t. transformer architectures: their inability to perform content-based reasoning. The internal SSM parameters of previous approaches had been fixed for all the tokens in a sequence, limiting their modelling capabilities. However, the authors introduce a novel Selective SSM (Mamba/S6) which is data-dependent i.e. its parameters are re-calculated for each individual input, allowing the model to filter out irrelevant content similarly to atten-

tion mechanisms. Mamba was thus able to outperform its transformer counterparts on numerous sequence modelling tasks at various sizes. Additionally, Mamba maintained linear scaling with input sequence length (as opposed to quadratic), allowing for much longer context windows.

Inspired by these results, the vision community has been working extensively to adapt the Mamba framework to higher-dimensional visual tasks. The pioneering works of Vision Mamba [40] and VMamba [18] were the first to extend the S6 architecture to process images. They use flattened convolutional patch embeddings to encode the image into a sequence compatible with the original model. These works further extend Mamba by introducing bi-directionality in the processing of the input, with forward and backward SSMs, and learned positional encodings. VideoMamba [16] follows the same paradigm, now adapting Vision Mamba to video understanding by extending its bi-directional input scan to the temporal domain, surpassing the state of the art. Since then multiple studies have emerged applying S6 variants to many other problems such as image restoration [11], video-based object segmentation [35] and point-cloud analysis [17] among others.

Nevertheless, the underlying design of the SSM applied in the aforementioned works remained grounded in the original NLP formulation. Specifically, the model's causal nature, which aligns well with the sequential and directional flow in language, imposes an unnecessary constraint on tasks in vision, where processing is largely non-causal. This directional dependency can hinder the SSM's effectiveness in applications that do not require strict sequence ordering. MambaVision [14] aimed to address this by modifying the components around the SSM, where the causal convolution before the SSM block is replaced with a regular non-causal convolution. A non-causal convolution is also added onto the parallel "gate" branch within the S6 block to compensate for the remaining causality of the SSM branch. On the other hand, the authors of VSSD [27] mitigate this limitation through a reformulation of the model's linear recurrence. In this case, the modified SSM yields a single hidden state per input image, containing global information gathered in a direction-agnostic manner. While this reformulation is a simple and attractive proposition, and can potentially perform competitively as a vision backbone, the underlying principles of the SSM are lost and the model degenerates to a weighted pooling operation, as many elements of the original methodology are discarded.

Moreover, the inherent causal properties of SSMs, combined with their sequential nature, imposes significant constraints on the flow of information between adjacent patches/pixels due to the 1-dimensional unrolling of their inputs. The recent works of V2M [32] and 2DMamba [37] aim to address these problems by introducing state transitions that account for dependencies in both horizontal and

vertical directions, rather than adhering to a strictly linear sequence. Nevertheless, these methods still utilize independent sequential scans along each spatial axis. Similarly, Chimera [1] implements a 2D parallel scan for modelling multivariate time series data. Here, the full 2D spatial relationships are not required and a simplified transition matrix helps maintain efficiency. Although these designs enhance computational efficiency, the capacity of each method to effectively integrate information across neighbouring regions is restricted in various manners, thereby limiting the ability to fully exploit the spatial coherence inherent to image/vision data.

To the best of our knowledge, no works have yet proposed a natively 2D SSM that does not sacrifice expressibility or input-dependent modelling capabilities.

3. Method

3.1. Preliminaries

State-Space Models. State Space Models describe a 1-D mapping from the continuous signal $x(t) : \mathbb{R} \to \mathbb{R}$ to $y(t) : \mathbb{R} \to \mathbb{R}$ through an N-D hidden state $h(t) \in \mathbb{R}^N$. Such a model is parametrised by two projection matrices $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$ and an evolution matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, formulated as linear ordinary differential equations (ODEs)

$$
\begin{aligned}
h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\
y(t) &= \mathbf{C}h(t),
\end{aligned}
\tag{1}
$$

where $h'(t)$ is the derivative of the hidden state at time t.

Technically, this formulation can also include an additional parameter $\mathbf{D}$ which allows the input to directly influence the output. However, this is often regarded as a skip connection and omitted for conciseness. As mentioned above, $\mathbf{B}$ and $\mathbf{C}$ serve as projection parameters. Their role is to regulate the extent to which the input influences the model's state, and how this state in turn influences the system's output. On the other hand, the state-evolution matrix $\mathbf{A}$ captures how the previous hidden state evolves naturally over time.

Although the underlying function to be modelled is continuous, the observation of $x(t)$ and labels of $y(t)$ exist only at discrete intervals. Therefore, these dynamics need to be discretised before being incorporated into deep learning algorithms. This discretization process involves transforming the continuous-time parameters $\mathbf{A}$, $\mathbf{B}$ into the discrete-time parameters $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$. This can be achieved through various discretization rules. We use the Zero Order Hold (ZOH) formulation of the SSM's parameters as described in [10],

$$
\overline{\mathbf{A}} = \exp(\Delta \mathbf{A}) \quad \overline{\mathbf{B}} = (\Delta \mathbf{A})^{-1}(\exp(\Delta \mathbf{A}) - \mathbf{I}) \cdot \Delta \mathbf{B}, \tag{2}
$$

where Δ is introduced as an additional learnable parameter that controls the step size or sampling interval of the continuous parameters. The role of Δ can also be thought of as

an attention mechanism, controlling the degree to which a particular sample within a sequence is "remembered".

Following the discretization step, the ODE of Eq. (1) can be rewritten as a recursion,

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t,$$
$$y_t = \mathbf{C}h_t, \tag{3}$$

allowing the system's state to be determined at any given time t, by integrating the current input x_t with its previous state h_{t-1}. This recursive interpretation enables the SSM to operate as a stateful recurrent model which scales linearly in complexity with respect to its input sequence length.

Traditional conceptualizations of State Space Models yielded low task performance due to poor initialization of the state-evolution matrix. One explanation found in the literature is that the ODEs of Eq. (1) solve to an exponential function, causing its gradients to scale exponentially as the sequence length increases (similarly to traditional RNNs) [10]. This highlighted the need for an $\mathbf{A}$ matrix capable of compressing the cumulative history of the input sequence in an efficient manner, with bounded gradients. In order to address these limitations, modern SSMs utilise the HiPPO theory of continuous-time memorization [8] as the initialization mechanism for this parameter,

$$\mathbf{A}_{nk} = -\begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & \text{if } n > k \\ n+1 & \text{if } n = k \\ 0 & \text{if } n < k \end{cases} \tag{4}$$

allowing the state to integrate (remember) recent inputs with much higher fidelity than those further in the past. Theoretically, when using the HiPPO matrix, the system decomposes the input sequence into a vector of coefficients representing Lagrange polynomials, similarly to how a Fourier transform decomposes a complex signal into its simpler sinusoidal components. SSMs initialised in this manner are referred to as "structured" due to the constraints that the HiPPO theory imposes in the model's state-evolution matrix.

However, imposing structure with HiPPO is not enough to outperform the current state of the art, as traditional Structured SSMs are unable to perform context-based and selective reasoning tasks, due to their Linear Time-Invariant (LTI) nature. Specifically because the model's parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and Δ are fixed for all input elements in a single sequence.

S6/Mamba employs an input-dependent parametrization to address the context-based reasoning challenges encountered by its predecessors. To do so, the model does not learn these parameters directly, but rather a set of functions

$$\begin{aligned} \mathbf{B} &= S_{\mathbf{B}}(\mathbf{U}), \\ \mathbf{C} &= S_{\mathbf{C}}(\mathbf{U}), \\ \Delta &= \tau_{\Delta}\left(\Delta + S_{\Delta}(\mathbf{U})\right), \end{aligned} \tag{5}$$

which transform the input elements into said parameters. Now $\mathbf{B} \in \mathbb{R}^{B \times L \times N}$, $\mathbf{C} \in \mathbb{R}^{B \times L \times N}$ and $\boldsymbol{\Delta} \in \mathbb{R}^{B \times L \times D}$ are dependent on the input sequence $\mathbf{U}$ of batch size B and length L with D channels. This allows the architecture to ignore certain tokens while paying closer attention to others. $S_{\mathbf{B}}$ and $S_{\mathbf{C}}$ represent linear projections to dimension N i.e. $\text{Linear}_N(\cdot)$, whereas $S_{\Delta} = \text{Broadcast}_D(\text{Linear}_1(\cdot))$ and τ_{Δ} is softplus.

To process images with S6, the standard approach in the field is based on that proposed in [40]. In general, a given input image $I \in \mathbb{R}^{H \times W \times C}$ (where H, W represents the size of the image) is flattened to form a 1D sequence $x \in \mathbb{R}^{C \times L}$ which is now compatible with the original SSM design.

3.2. Methodology

Our proposed method to reformulate Mamba for 2D inputs closely aligns with S4ND, by initialising two fully independent sets of data-dependant SSM parameters for each axis indexed by t, z (i.e. $\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t, \mathbf{D}_t, \Delta_t$ and $\mathbf{A}_z, \mathbf{B}_z, \mathbf{C}_z, \mathbf{D}_z, \Delta_z$). However, we contrast with S4ND by avoiding taking the outer product of the two independent SSMs (which collapses the parameter space of $\mathbf{A}$ to twice the state size). This results in increased expressibility as $\mathbf{A}$ maintains a full N^2 parameter space, while also ensuring that the selectivity mechanism and non-LTI characteristics of Mamba are retained.

We begin by redefining Eq. (1) to account for both inputs and hidden states which are defined across a 2D domain, to include 2 independent partial derivatives (h_t' and h_z'):

$$h_t'(t, z) = \mathbf{A}_t h(t, z) + \mathbf{B}_t x(t, z). \tag{6}$$

$$h_z'(t, z) = \mathbf{A}_z h(t, z) + \mathbf{B}_z x(t, z). \tag{7}$$

We then form recurrence relationships similar to that of Eq. (3) by performing first-order Euler discretisations of the hidden state along 2 dimensions, and re-arranging

$$h(t + \Delta_t, z) = \Delta_t h_t'(t, z) + h(t, z), \tag{8}$$
$$h(t, z + \Delta_z) = \Delta_z h_z'(t, z) + h(t, z). \tag{9}$$

then substituting Eq. (6) and (7) into Eqns. (8) and (9) as follows:

$$h(t + \Delta_t, z) = \Delta_t(\mathbf{A}_t h(t, z) + \mathbf{B}_t x(t, z)) + h(t, z), \tag{10}$$

$$h(t, z + \Delta_z) = \Delta_z(\mathbf{A}_z h(t, z) + \mathbf{B}_z x(t, z)) + h(t, z). \tag{11}$$

We can now re-write these in the form:

$$h(t + \Delta_t, z) = \overline{\mathbf{A}}_t h(t, z) + \overline{\mathbf{B}}_t x(t, z), \tag{12}$$
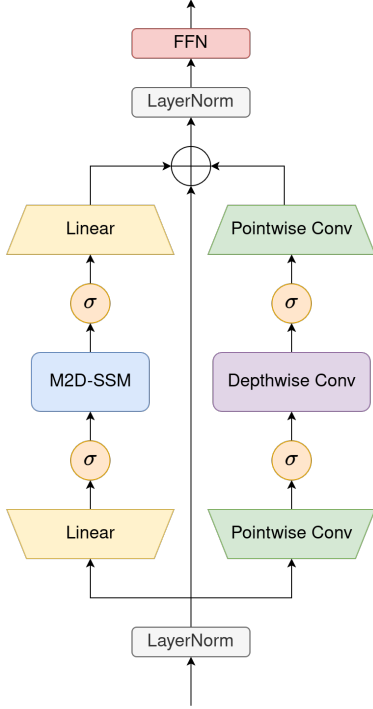$$h(t, z + \Delta_z) = \overline{\mathbf{A}}_z h(t, z) + \overline{\mathbf{B}}_z x(t, z), \tag{13}$$

4

Figure 2. Architecture of Mamba2D Block. Two parallel branches implement our Mamba2D SSM alongside the local processing path, followed by an FFN in line with traditional transformer-style blocks.

where $\overline{\mathbf{A}}_t = \Delta_t(\mathbf{A}_t + I)$ and $\overline{\mathbf{B}}_t = \Delta_t \mathbf{B}_t$ (and similarly for z).

We note that this formulation provides two independent recurrence relationships for estimating the same thing (i.e. the hidden state at a pixel can be estimated by applying Eq. (12) to the pixel above, or by applying Eq. (13) to the pixel on the left). To resolve this, we take as our final hidden state, the point which minimises the distance between the two estimates. This results in our 2D SSM equations:

$$h(t, z) = \frac{1}{2} \left(\begin{bmatrix} \overline{\mathbf{A}}_t \\ \overline{\mathbf{A}}_z \end{bmatrix}^\top \begin{bmatrix} h(t - \Delta_t, z) \\ h(t, z - \Delta_z) \end{bmatrix} + \begin{bmatrix} x(t, z) \\ x(t, z) \end{bmatrix}^\top \begin{bmatrix} \overline{\mathbf{B}}_t \\ \overline{\mathbf{B}}_z \end{bmatrix} \right),$$
(14)

$$y(t, z) = \mathbf{C}h(t, z) + \mathbf{D}x(t, z).$$
(15)

This provides a system where the influence of each input element on each output element is determined by the Manhattan distance between those points, rather than the distance along an arbitrary 1D flattening of the data.

In order to implement Mamba2D effectively, as with other SSM based methods, the recurrent form must be parallelised. The parallel associative scan from the original Mamba formulation cannot be extended to 2D, due to the recurrent cumulative multiplication of $\overline{\mathbf{A}}_t$ and $\overline{\mathbf{A}}_z$. This re-

sults in a computationally intractable number of paths to traverse across a 2D plane. Therefore, we solve this issue by implementing wavefront parallelism (as depicted in Fig. 1) which processes each diagonal across the 2D plane in parallel through a custom CUDA kernel, for both the forward and backward passes.

Because of the improved pixel connectivity of Mamba2D, and the lack of data flattening, we are able to exploit the findings of [41]. In particular, we use a single causal 2D scan to aggregate all features across all spatial neighbourhoods. As a result, the relatively expensive computation of a "reverse" scan (traditionally used in 1D Mamba to remove causality) is avoided, leading to improved efficiency of Mamba2D.

3.3. Mamba2D model

Block Structure. In accordance with the current state of the art, we employ a MetaFormer-style block structure for our architectures: Given an input batch $x \in \mathbb{R}^{B \times H \times W \times C}$, this is transformed to a tensor of activations $y \in \mathbb{R}^{B \times H \times W \times C}$ as follows:

$$\hat{x} = \text{Mixer}(\text{Norm}(x)) + x$$
$$y = \text{MLP}(\text{Norm}(\hat{x})) + \hat{x},$$
(16)

where Norm and Mixer denote our choices of normalization layer (i.e. LayerNorm) and token-mixer (discussed below) respectively, while MLP denotes a 2-layer MLP with a GELU non-linearity.

Mamba2D Mixer. In order to efficiently apply our M2D-SSM to visual tasks we re-design the token-mixer construct proposed in [7]. First, we remove the symmetric branch included in the original Mamba architecture, which served as an additional gating mechanism. This adjustment was motivated by the observation that the selective capabilities of our M2D-SSM already effectively regulate information flow, rendering it redundant. Instead, we replace it with a local processing path comprised of a depthwise separable convolution for the efficient extraction of high-detail local features. This modification not only enhances the extraction of fine-grained spatial information but also introduces valuable local priors into the feature representation, which contributes to improved model convergence speed.

Overall Architecture. We implement the Mamba2D mixer described above as a choice of token mixer within a hierarchical transformer structure, as shown in Figure 3, inspired by the baselines laid out in MetaFormer. As such, our overall model consists of 4 stages comprising of a strided convolution-based stem/downsample followed by $N_{1...4}$ M2D blocks within each stage.
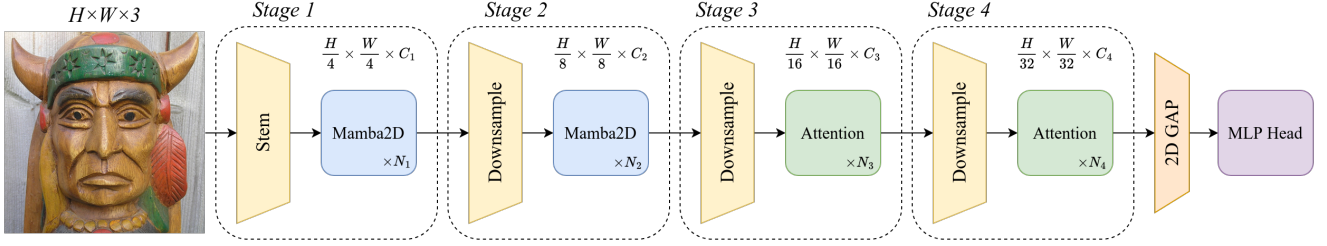
5

Figure 3. Architecture of our Mamba2D Model. A convolutional stem performs an initial patch embedding of the input image, followed by 4 stages of further feature extraction. Each stage consists of $N_{1...4}$ blocks containing a token mixer followed by an FFN. As shown, we opt to use Mamba2D as the token mixer for the first two stages where spatial relations are more impactful. The final two stages comprise of vanilla attention for a lossless encoding of the channel-wise relations as the spatial size of the features are diminished.

Table 1. Top-1 classification accuracy across numerous vision backbones on ImageNet-1K.

Model	Type	#Param.	Top-1 acc. (%)
RegNetY-4G [26]	Conv	21 M	80.0
ConvNeXt-T [22]	Conv	26 M	80.1
RegNetY-8G [26]	Conv	39 M	81.7
DeiT-S [30]	Attn	22 M	79.8
Swin-T [20]	Attn	29 M	81.3
SwinV2-T [21]	Attn	29 M	81.8
PVT-M [34]	Attn	44 M	81.2
ViM-S [40]	SSM	26 M	80.1
VMamba-T [18]	SSM	30 M	82.6
MambaVision-T [14]	SSM	32 M	82.3
EfficientVMamba-B [25]	SSM	33 M	81.8
VSSD-T [27]	2D-SSM	24 M	83.7
V2M-S [33]	2D-SSM	26 M	80.5
2DVMamba-T [38]	2D-SSM	31 M	82.8
S4ND-ViT-B [24]	2D-SSM	89 M	80.4
M2D-T	2D-SSM	27 M	82.4

For the task of classification, we employ Global Average Pooling (GAP) before feeding the features to an MLP classification head. Additionally, we propose a hybrid selection of token mixers at different points of the network, exploiting the strengths of each mechanism. Specifically, for the first two stages, our Mamba2D token mixer effectively captures long-range spatial dependencies whereas we opt for an attention-based token mixer in the latter two stages given the significantly reduced spatial dimensions of the feature maps.

4. Experiments

In this section, we present an experimental evaluation of the proposed framework. Firstly, we construct our M2D-T model with [3,3,9,3] blocks and [64,128,320,512] channels in each stage respectively, leading to a total model size of 26.7M parameters. Our model was trained on the ImageNet-1K dataset following the typical practices established in prior works [18, 19, 31]. Specifically, we employ a range of data augmentation schemes including: random cropping (to 224^2), random horizontal flipping, auto random augmentation, MixUp and random erasing [39]. Whilst we also train for 300 epochs, we adopt the schedule free AdamW optimiser [4] with an effective batch size of 4096, a learning rate of 0.004, weight decay set to 0.05 and a 5% warmup w.r.t optimisation steps. Additionally, in-line with contemporary works, we also make use of EMA [29] during training.

4.1. Image Classification on ImageNet-1K

Table 1 presents the results for Imagenet-1K classification. We compare against a range of model types, from Conv-based, Transformer, Mamba and hybrids of all three mechanisms. As can be seen, our model performance is well in line with the current state-of-the-art SSM techniques. M2D

generally outperforms all techniques with lower parameter counts, and remains more compact and efficient than models with higher accuracy.

5. Conclusion

This paper presents Mamba2D (M2D), a novel reformulation of the S6/Mamba State Space Model (SSM) methodology to natively process 2D visual data. M2D does not rely on simplifications or abstractions of the S6 scan and instead incorporates information from both image dimensions simultaneously while constructing a unified hidden state within the SSM model.

Critically, we retain the full expressibility of the original S6 scan along with the input-dependent selectivity mechanism that originally brought Mamba to the forefront of the SSM field. To maintain efficiency, we utilise a 2D wavefront scan computation paradigm, in place of the 1D parallel scan employed by traditional SSM models.

Evaluations of the model on the task of classification demonstrate the competitive performance of M2D, surpassing a range of well-established Convolution, Attention and SSM based models at similar parameter counts.

6. Future Work

Although this preliminary experimentation shows the promise of the proposed M2D model, we also aim to evaluate the application of M2D as a generic vision backbone for use in varied downstream tasks. Additionally, if larger model sizes are practical, it would be interesting to scale up the model sizes and perform a more detailed hyperparameter search to fully explore the upper-bound of model performance. Finally, implementing further hardware-aware optimisations for the custom M2D-SSM wavefront kernel could significantly enhance the efficiency of M2D, providing a more favourable balance between compute and performance.

References

[1] Ali Behrouz, Michele Santacatterina, and Ramin Zabih. Chimera: Effectively modeling multivariate time series with 2-dimensional state space models, 2024. 3

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second, 2024. 1

[3] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying Convolution and Attention for All Data Sizes, 2021. 1

[4] Aaron Defazio, Xingyu Alice Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The Road Less Scheduled, 2024. 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. 1

[6] Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry Hungry Hippos: Towards Language Modeling with State Space Models, 2023. 2

[7] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2023. 1, 2, 5

[8] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. HiPPO: Recurrent Memory with Optimal Polynomial Projections, 2020. 2, 4

[9] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining Recurrent, Convolutional, and Continuous-time Models with Linear State-Space Layers, 2021. 2

[10] Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces, 2022. 2, 3, 4

[11] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A Simple Baseline for Image Restoration with State-Space Model, 2024. 3

[12] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. CMT: Convolutional Neural Networks Meet Vision Transformers, 2022. 1

[13] Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal State Spaces are as Effective as Structured State Spaces, 2022. 2

[14] Ali Hatamizadeh and Jan Kautz. MambaVision: A Hybrid Mamba-Transformer Vision Backbone, 2024. 2, 3, 6

[15] Anil Kag, Huseyin Coskun, Jierun Chen, Junli Cao, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, and Jian Ren. AsCAN: Asymmetric Convolution-Attention Networks for Efficient Recognition and Generation, 2024. 1

[16] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. VideoMamba: State Space Model for Efficient Video Understanding, 2024. 3

[17] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Point-Mamba: A Simple State Space Model for Point Cloud Analysis, 2024. 3

[18] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. VMamba: Visual State Space Model, 2024. 3, 6

[19] 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, 2021. 6

[20] 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, 2021. 1, 6

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution, 2022. 1, 6

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s, 2022. 6

[23] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long Range Language Modeling via Gated State Spaces, 2022. 2

[24] Eric Nguyen, Karan Goel, Albert Gu, Gordon W. Downs, Preey Shah, Tri Dao, Stephen A. Baccus, and Christopher Ré. S4ND: Modeling Images and Videos as Multidimensional Signals Using State Spaces, 2022. 2, 6

[25] Xiaohuan Pei, Tao Huang, and Chang Xu. EfficientV-Mamba: Atrous Selective Scan for Light Weight Visual Mamba, 2024. 6

[26] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces, 2020. 6

[27] Yuheng Shi, Minjing Dong, Mingjia Li, and Chang Xu. VSSD: Vision Mamba with Non-Casual State Space Duality, 2024. 2, 3, 6

[28] Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified State Space Layers for Sequence Modeling, 2023. 2

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018. 6

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. 6

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. 6

[32] Chengkun Wang, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. V2m: Visual 2-dimensional mamba for image representation learning, 2024. 3

[33] Chengkun Wang, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. V2M: Visual 2-Dimensional Mamba for Image Representation Learning, 2024. 6

[34] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, 2021. 6

[35] Yijun Yang, Zhaohu Xing, Chunwang Huang, and Lei Zhu. Vivim: A Video Vision Mamba for Medical Video Object Segmentation, 2024. 3

[36] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer Is Actually What You Need for Vision, 2022. 1

[37] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S. Hosseini. 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification, 2024. 3

[38] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S. Hosseini. 2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification, 2024. 6

[39] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation, 2017. 6

[40] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model, 2024. 3, 4, 6

[41] Qinfeng Zhu, Yuan Fang, Yuanzhi Cai, Cheng Chen, and Lei Fan. Rethinking scanning strategies with vision mamba in semantic segmentation of remote sensing imagery: An experimental study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:18223–18234, 2024. 5