

MambaVision: A Hybrid Mamba-Transformer Vision Backbone

Ali Hatamizadeh, Jan Kautz

NVIDIA

{ahatamizadeh, jkautz}@nvidia.com

Abstract

We propose a novel hybrid Mamba-Transformer backbone, MambaVision, specifically tailored for vision applications. Our core contribution includes redesigning the Mamba formulation to enhance its capability for efficient modeling of visual features. Through a comprehensive ablation study, we demonstrate the feasibility of integrating Vision Transformers (ViT) with Mamba. Our results show that equipping the Mamba architecture with self-attention blocks in the final layers greatly improves its capacity to capture long-range spatial dependencies. Based on these findings, we introduce a family of MambaVision models with a hierarchical architecture to meet various design criteria. For classification on the ImageNet-1K dataset, MambaVision variants achieve state-of-the-art (SOTA) performance in terms of both Top-1 accuracy and throughput. In downstream tasks such as object detection, instance segmentation, and semantic segmentation on MS COCO and ADE20K datasets, MambaVision outperforms comparably sized backbones while demonstrating favorable performance. Code: <https://github.com/NVlabs/MambaVision>

1. Introduction

In recent years, Transformers [31] have become the de facto architecture across different domains including computer vision, natural language processing, speech processing, and robotics. The versatility of the Transformer architecture, primarily due to its attention mechanism and flexibility, makes it highly suitable for multimodal learning tasks, where integrating and processing information from various modalities is essential. Despite these benefits, the quadratic complexity of the attention mechanism with respect to sequence length makes Transformers computationally expensive to train and deploy. Recently, Mamba [7] proposed a new State Space Model (SSM) that achieves linear time complexity and outperforms or matches Transformers in different language modeling tasks [7]. The core contribution of Mamba is a novel selection mechanism that enables efficient input-dependent processing of long sequences with hardware-aware con-

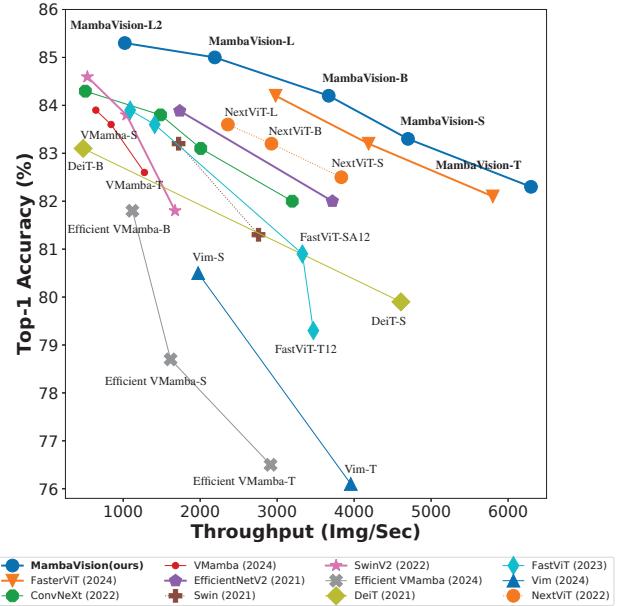


Figure 1 – Top-1 accuracy vs. image throughput comparisons on ImageNet-1K dataset. The MambaVision models achieve a new Pareto front for Top-1 accuracy and image throughput tradeoff. Specifically, MambaVision variants outperform Mamba-based models such as VMamba and Vim, sometimes by a significant margin. For all models, image throughput is measured on an A100 NVIDIA GPU with a batch size of 128.

siderations. Recently, a number of Mamba-based backbones [25, 40] have been proposed to leverage the strengths of its SSM formulation in vision tasks such as image classification and semantic segmentation. However, the Mamba’s autoregressive formulation, while effective for tasks requiring sequential data processing, faces limitations in computer vision tasks that benefit from a full receptive field: (1) Unlike sequences where order matters, image pixels do not have a sequential dependency in the same way. Instead, spatial relationships are often local and need to be considered in a more parallel and integrated manner. Hence, this results in inefficiency for processing spatial data (2) An autoregressive model like Mamba processes data step-by-step,

limiting its ability to capture and utilize global context in one forward pass. In contrast, vision tasks often require understanding the global context to make accurate predictions about local regions. Vision Mamba (Vim) [40] and others have proposed modifications such as bidirectional SSMs to address lack of global context and spatial understanding. While bidirectional SSMs have the potential to capture more comprehensive context, they introduce significant latency due to the need to process the entire sequence before making predictions. Additionally, the increased complexity can lead to challenges in training, risk of overfitting, and may not always result in better accuracy. Due to these pitfalls, backbones with Vision Transformer (ViT) and Convolutional Neural Network (CNN) architectures still outperform best Mamba-based vision models on different vision tasks.

In this work, we systematically re-design the Mamba block to make it more suitable for vision tasks. We propose a hybrid architecture that consists of our proposed formulation (*i.e.* MambaVision Mixer and MLP) as well as Transformer blocks. Specifically, we study different integration patterns such as adding the Transformer blocks in an iso-parameter manner to earlier, middle, and final layers as well as every l layers. Our analysis shows that leveraging several self-attention blocks at the final stages can significantly enhance the capability to capture global context and long-range spatial dependencies. As shown in Sec. 5, using a hybrid architecture also results in higher image throughput compared to both pure Mamba and ViT-based models. We introduce the MambaVision model which consists of a multi-resolution architecture and leverages CNN-based residual blocks for fast feature extraction of larger resolution features. As shown in Fig. 1, the MambaVision achieves a new SOTA Pareto front in terms of ImageNet-1K Top-1 accuracy and image throughput, outperforming Mamba, CNN, and ViT-based models, sometimes by a significant margin. In downstream tasks such as object detection, instance segmentation, and semantic segmentation, models with MambaVision backbones outperform comparably-sized counterparts on MS COCO and ADE20 datasets, respectively. Hence, it validates the effectiveness and versatility of MambaVision as an efficient backbone.

To the best of our knowledge, MambaVision is the *first* effort to study and develop a hybrid architecture comprising of both Mamba and Transformers for computer vision applications. Our main contributions in this work are summarized as follows:

- We introduce a redesigned vision-friendly Mamba block, improving accuracy and image throughput over the original Mamba architecture.
- We present a systematic investigation of integration patterns for Mamba and Transformer blocks, and demonstrate that incorporating self-attention blocks at the final stages significantly improves the model’s ability to capture global

context and long-range spatial dependencies.

- We introduce MambaVision, which is a novel hybrid Mamba-Transformer model. The hierarchical MambaVision achieves a new SOTA Pareto front on the ImageNet-1K dataset in terms of Top-1 and image throughput.

2. Related work

Conv-Based. CNNs have been the cornerstone of computer vision since the introduction of AlexNet [16]. Recent efforts have focused on modernizing CNN architectures with Transformer-inspired principles. ConvNeXt [23] demonstrated competitive performance with Transformers by redesigning ResNet [12] with increased width, larger kernels, and layer normalization. RegNetY [26] introduced systematic network design through design space analysis, while EfficientNetV2 [27] leveraged neural architecture search and progressive learning for better efficiency trade-offs. Despite strong performance, these CNN models inherently lack the global receptive field needed for capturing long-range dependencies.

Transformer-Based. ViTs [5] marked a significant shift in computer vision by introducing self-attention mechanisms for enlarged receptive fields. However, ViTs initially lacked the inherent advantages of CNNs and required extensive training data. To address these limitations, DeiT [28] introduced distillation-based training for improved accuracy on smaller datasets, while Swin Transformer [21] proposed a hierarchical architecture using shifted windows for self-attention, effectively balancing local and global context. Models such as Twins [3] and PVT [32] further enhanced efficiency through spatially separable self-attention and hierarchical structures with patch embedding. Despite these advances, the quadratic complexity of self-attention operations in these models continued to pose efficiency challenges.

Conv-Transformer. The complementary strengths of CNNs and ViTs inspired hybrid architectures. CoAT [36] and CrossViT [1] demonstrated enhanced feature learning by combining convolutions with self-attention, while NextViT [17] systematically incorporated CNN-like processing into Transformers. Recent efforts like EfficientFormer [18] and FasterViT [10] focused on optimizing efficiency-accuracy trade-offs, achieving competitive performance with high throughput through carefully designed hybrid architectures.

Mamba-Based. Since the introduction of Mamba, a number of efforts have been proposed to leverage its capability for vision applications. Vim [40] introduced a bidirectional SSM formulation that processes tokens in both forward and backward directions to capture global context and improve spatial understanding. However, this bidirectional approach faces significant limitations: it increases computational overhead, slows down training and inference times, and struggles to effectively combine information from

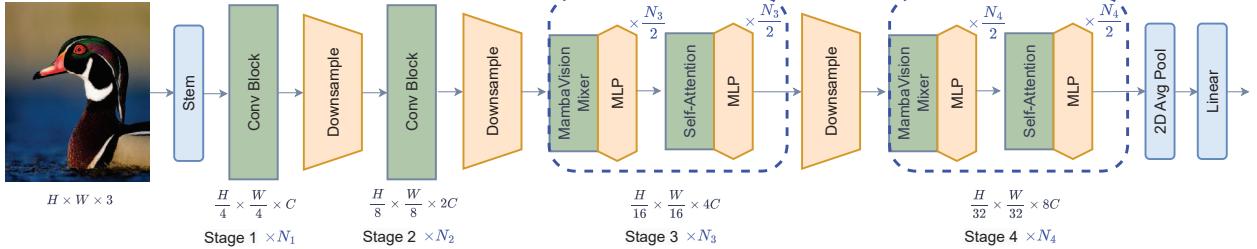


Figure 2 – The architecture of hierarchical MambaVision models. The first two stages use residual convolutional blocks for fast feature extraction. Stages 3 and 4 employ both MambaVision and Transformer blocks. Specifically, given N layers, we use $\frac{N}{2}$ MambaVision and MLP blocks, which are followed by additional $\frac{N}{2}$ Transformer and MLP blocks. The Transformer blocks in the final layers allow for recovering lost global context and capturing long-range spatial dependencies.

multiple directions without losing global context. In contrast, MambaVision achieves superior results using a single forward pass with our redesigned Mamba block, demonstrating significantly better ImageNet Top-1 accuracy and throughput. VMamba [20] proposed a generic Mamba-based vision backbone featuring a Cross-Scan Module (CSM). This module implements a four-way selective scan methodology (from upper-left and lower-right to opposite directions) to integrate information from surrounding tokens and capture global context. VMamba also incorporates architectural modifications like depth-wise convolutions and a hierarchical multi-resolution structure. While the CSM module is designed for vision tasks, its receptive field remains constrained by the cross-scan paths. MambaVision offers several advantages over VMamba: our mixer design is simpler yet captures both short and long-range dependencies more effectively, we employ CNN-based layers for fast feature extraction rather than using uniform block structures across all stages, and we achieve superior performance with significantly higher throughput. EfficientVMamba [25] uses SSMs for larger resolutions and CNNs for lower ones, while MambaVision takes the opposite approach with CNNs at higher resolutions and SSM/self-attention at lower ones, leading to significantly better accuracy and throughput. Similarly, while SiMBA [24] addresses Mamba’s stability through EinFFT channel modeling, it doesn’t fully address spatial understanding limitations, resulting in lower performance compared to MambaVision’s comprehensive design.

3. Methodology

3.1. Macro Architecture

In this section, we introduce MambaVision which is our proposed novel architecture with SOTA performance on ImageNet-1K dataset. As illustrated in Fig. 2, MambaVision has a hierarchical architecture consisting of 4 different stages. The first two stages consist of CNN-based layers for fast feature extraction at higher input resolutions, while stage 3 and 4 include the proposed MambaVision and Transformer

blocks. Specifically, given an image of size $H \times W \times 3$, the input is first converted into overlapping patches with size $\frac{H}{4} \times \frac{W}{4} \times C$ and projected into a C dimensional embedding space by the stem which consists of two consecutive 3×3 CNN layers with stride of 2. The downampler in between stages consists of a 3×3 CNN layer with stride 2 which reduces the image resolution by half. Furthermore, the CNN blocks in stages 1 and 2 follow a generic residual block formulation according to the following

$$\begin{aligned}\hat{\mathbf{z}} &= \text{GELU}(\text{BN}(\text{Conv}_{3 \times 3}(\mathbf{z}))), \\ \mathbf{z} &= \text{BN}(\text{Conv}_{3 \times 3}(\hat{\mathbf{z}})) + \mathbf{z},\end{aligned}\quad (1)$$

GELU and BN denote Gaussian Error Linear Unit activation function [14] and batch normalization [15], respectively. Please see the supplementary materials for further details regarding MambaVision macro architecture.

3.2. Micro Architecture

In this section, we first revisit the preliminaries of Mamba and SSMs. We then present the micro design of the architecture in stages 3 and 4 and discuss MambaVision formulation in more details.

3.2.1. Mamba Preliminaries

In Mamba, a 1D continuous input $x(t) \in \mathbb{R}$ is transformed into $y(t) \in \mathbb{R}$ via a learnable hidden state $h(t) \in \mathbb{R}^M$ with parameters $\mathbf{A} \in \mathbb{R}^{M \times M}$, $\mathbf{B} \in \mathbb{R}^{M \times 1}$ and $\mathbf{C} \in \mathbb{R}^{1 \times M}$ according to

$$\begin{aligned}h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t),\end{aligned}\quad (2)$$

Discretization The continuous parameters \mathbf{A} , \mathbf{B} and \mathbf{C} in the above formulation are further converted into discrete parameters for better computational efficiency [8]. Specifically, assuming a timescale Δ , a zero-order hold rule can be applied to obtain discrete parameters $\bar{\mathbf{A}} \in \mathbb{R}^{M \times M}$,

Algorithm 1 PyTorch-like pseudo-code for MambaVision mixer

```

import torch
import math
import torch.nn as nn
import torch.nn.functional as F
from einops import rearrange, repeat

class MambaVisionMixer(nn.Module):
    def __init__(self, dim, d_state=16, kernel_size=3):
        super().__init__()
        self.d_state = d_state
        self.dt_rank = math.ceil(dim / 16)
        self.in_proj = nn.Linear(self.d_model, self.d_inner)
        self.convld_x = nn.Conv2d(dim//2, dim//2, kernel_size=kernel_size, padding='same', groups=dim//2)
        self.convld_z = nn.Conv2d(dim//2, dim//2, kernel_size=kernel_size, padding='same', groups=dim//2)
        self.dt_proj = nn.Linear(self.dt_rank, dim//2)
        dt = torch.exp(torch.rand(dim//2) * (math.log(dt_max) - math.log(dt_min)) + math.log(dt_min))
        A_log = torch.log(repeat(torch.arange(1, self.d_state + 1), n -> d n, d=dim//2))
        self.A_log = nn.Parameter(A_log)
        self.D = nn.Parameter(torch.ones(dim//2))
        self.out_proj = nn.Linear(dim, dim)

    def forward(self, hidden_states):
        xz = rearrange(self.in_proj(hidden_states), b 1 d -> b d 1)
        x, z = xz.chunk(2, dim=1)
        A = -torch.exp(self.A_log)
        x = F.silu(self.convld_x(x))
        z = F.silu(self.convld_z(z))
        seqlen = hidden_states.shape[1]
        x_dbl = self.x_proj(rearrange(x, b d 1 -> (b 1) d))
        dt, B, C = torch.split(x_dbl, [self.dt_rank, self.d_state, self.d_state], dim=-1)
        dt = rearrange(self.dt_proj(dt), (b 1) d -> b d 1, l=seqlen)
        B = rearrange(B, (b 1) dstate -> b dstate 1, l=seqlen)
        C = rearrange(C, (b 1) dstate -> b dstate 1, l=seqlen)
        x_ssm = selective_scan_fn(x, dt, A, B, C, D)
        hidden_states = rearrange(torch.cat([x_ssm, z], dim=1), b d 1 -> b 1 d)
        return self.out_proj(hidden_states)

```

$\bar{\mathbf{B}} \in \mathbb{R}^{M \times 1}$ and $\bar{\mathbf{C}} \in \mathbb{R}^{1 \times M}$ according to

$$\begin{aligned}\bar{\mathbf{A}} &= \exp(\Delta \mathbf{A}), \\ \bar{\mathbf{B}} &= (\Delta \mathbf{A})^{-1} (\exp(\Delta \mathbf{A}) - \mathbf{I}) \cdot (\Delta \mathbf{B}), \\ \bar{\mathbf{C}} &= \mathbf{C},\end{aligned}\quad (3)$$

The Eq. 2 can then be expressed with discrete parameters as

$$\begin{aligned}h(t) &= \bar{\mathbf{A}} h(t-1) + \bar{\mathbf{B}} x(t), \\ y(t) &= \bar{\mathbf{C}} h(t),\end{aligned}\quad (4)$$

In addition, for an input sequence with size T , a global convolution with kernel $\bar{\mathbf{K}}$ can be applied for computing the output of Eq. 4 as in the following

$$\begin{aligned}\bar{\mathbf{K}} &= (\mathbf{C} \bar{\mathbf{B}}, \mathbf{C} \bar{\mathbf{A}} \bar{\mathbf{B}}, \dots, \mathbf{C} \bar{\mathbf{A}}^{T-1} \bar{\mathbf{B}}), \\ \mathbf{y} &= \mathbf{x} * \bar{\mathbf{K}},\end{aligned}\quad (5)$$

Selectivity Mamba further extends the S4 formulation by introducing a selection mechanism which allows for input-dependant sequence processing. This allows the model's parameters \mathbf{B} , \mathbf{C} and Δ to be adjusted dynamically according to the inputs and filter out irrelevant information. Further discretization details are provided in [7].

3.2.2. Layer Architecture

Assuming an input $X \in \mathbb{R}^{T \times C}$ with sequence length T with embedding dimension C , the output of layer n in stages 3 and 4 can be computed as in

$$\begin{aligned}\hat{X}^n &= \text{Mixer}(\text{Norm}(X^{n-1})) + X^{n-1}, \\ X^n &= \text{MLP}(\text{Norm}(\hat{X}^n)) + \hat{X}^n,\end{aligned}\quad (6)$$

Norm and Mixer denote the choices of layer normalization and token mixing blocks, respectively. Without loss of generality, Layer Normalization is used for Norm. Given N layers, the first $\frac{N}{2}$ layers employ MambaVision mixer blocks while the remaining $\frac{N}{2}$ layers employ self-attention. We describe the details of each mixer in the following.

MambaVision Mixer As shown in Fig. 3, we redesigned the original Mamba mixer to make it more suitable for vision tasks. First, we propose to replace the causal convolution with regular convolution, since it limits the influence to one direction, which is unnecessary and restrictive for vision tasks. In addition, we added a symmetric branch without SSM, consisting of an additional convolution and Sigmoid Linear Unit (SiLU) [6] activation, to compensate for any content lost due to the sequential constraints of SSMs. We then concatenate the output of both branches and project it via a final linear layer. This combination ensures that the

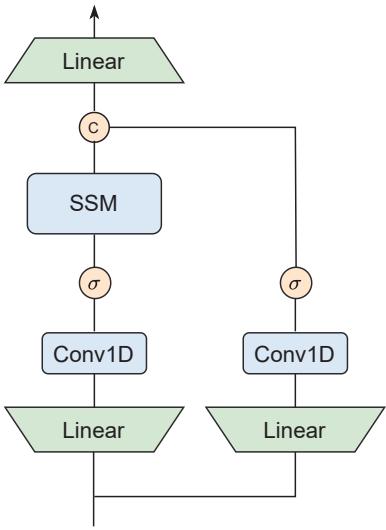


Figure 3 – Architecture of MambaVision block. In addition to replacing causal Conv layer with their regular counterparts, we create a symmetric path without SSM as a token mixer to enhance the modeling of global context.

final feature representation incorporates both the sequential and spatial information, leveraging the strengths of both branches. We note that the output of each branch is projected into an embedding space with size $\frac{C}{2}$ (*i.e.* half the size of original embedding dimension) to maintain similar number of parameters to the original block design. Given an input X_{in} , the output of MambaVision mixer X_{out} is computed according to

$$\begin{aligned} X_1 &= \text{Scan}(\sigma(\text{Conv}(\text{Linear}(C, \frac{C}{2})(X_{in})))), \\ X_2 &= \sigma(\text{Conv}(\text{Linear}(C, \frac{C}{2})(X_{in}))), \\ X_{out} &= \text{Linear}(\frac{C}{2}, C)(\text{Concat}(X_1, X_2)), \end{aligned} \quad (7)$$

$\text{Linear}(C_{in}, C_{out})(\cdot)$ denotes a linear layer with C_{in} and C_{out} as input and output embedding dimensions, Scan is the selective scan operation as in [7] and σ is the activation function for which SiLU is used. In addition, Conv and Concat represent 1D convolution and concatenation operations. In Algorithm 1, we present a PyTorch-like pseudo-code for MambaVision mixer. In general, our proposed modification leads to richer feature representations, better generalization, and improved performance on computer vision tasks. We have also experimentally validated the effectiveness of each of our design choices in Sec. 5.3.

Self-attention We use a generic multihead self-attention mechanism in accordance to

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V. \quad (8)$$

Q, K, V denote query, key and value respectively and d_h is the number of attention heads. In addition, our framework allows for computing the attention in a windowed manner similar to previous efforts [21, 22] (see the supplementary materials for window size ablation study).

4. Experiments

Image classification experiments are conducted on the ImageNet-1K dataset [4]. We followed the standard training recipe of previous efforts [11, 21, 37] to allow for a comparable analysis of performance across different models. Specifically, all models have been trained for 300 epochs using 32 A100 GPUs. The self-attention formulation in stages 3 and 4 of all MambaVision variants use a window size of 14 and 7, respectively. For detailed training configurations, please see the provided anonymous code repository. To evaluate the performance of downstream tasks, we used our pre-trained models as backbones for object detection, instance segmentation, and semantic segmentation tasks using the MS COCO dataset [19] and ADE20K dataset [39], respectively. Specifically, for object detection and instance segmentation, we used Cascade Mask-RCNN [13] head with hyperparameters such as $\times 3$ LR schedule. For semantic segmentation, we used a UperNet network [34] head and 8 A100 GPUs for all experiments.

5. Results

5.1. Image classification

In Table 1, we present the ImageNet-1K classification results. Specifically, we compare against different families of models such as Conv-based, Transformer-based, Conv-Transformer, and Mamba-based architectures and demonstrate that our model outperforms the previous efforts by a large margin, considering ImageNet Top-1 accuracy and image throughput. For example, MambaVision-B achieves higher accuracy (84.2%) compared to ConvNeXt-B (83.8%) and Swin-B (83.5%), while also having significantly better image throughput. We observe similar trends in comparison to Mamba-based models. Specifically, MambaVision-B (84.2%) outperforms VMamba-B (83.9%) despite having considerably higher image throughput. We observe similar trends in performance comparisons with respect to other Mamba-based models. In addition, we would also like to note that although our main design goal has been to optimize the accuracy and throughput tradeoff, the MambaVision model variants have much lower FLOPs when compared to similarly-sized counterparts. For instance, MambaVision-B has 56% less GFLOPs than MaxViT-B.

5.2. Object Detection and Segmentation

We evaluate our model’s object detection and instance segmentation performance on the MS COCO dataset [19], as

Table 1 – Comparison of classification benchmarks on **ImageNet-1K** dataset [4]. Image throughput is measured on A100 GPU with a batch size of 128.

Model	Image Size	#Params	FLOPs	Throughput	Top-1 (%)
	(Px)	(M)	(G)	(Img/Sec)	
Conv-Based					
ConvNeXt-T [23]	224	28.6	4.5	3196	82.0
ConvNeXt-S [23]	224	50.2	8.7	2008	83.1
ConvNeXt-B [23]	224	88.6	15.4	1485	83.8
RegNetY-040 [26]	288	20.6	6.6	3227	83.0
ResNetV2-101 [33]	224	44.5	7.8	4019	82.0
EfficientNetV2-S [27]	384	21.5	8.0	1735	83.9
Transformer-Based					
Swin-T [21]	224	28.3	4.4	2758	81.3
Swin-S [21]	224	49.6	8.5	1720	83.2
SwinV2-T [22]	256	28.3	4.4	1674	81.8
SwinV2-S [22]	256	49.7	8.5	1043	83.8
SwinV2-B [22]	256	87.9	15.1	535	84.6
TNT-S [9]	224	23.8	4.8	1478	81.5
Twins-S [3]	224	24.1	2.8	3596	81.7
Twins-B [3]	224	56.1	8.3	1926	83.1
Twins-L [3]	224	99.3	14.8	1439	83.7
DeiT-B [28]	224	86.6	16.9	2035	82.0
DeiT3-L [29]	224	304.4	59.7	535	84.8
PoolFormer-M58 [38]	224	73.5	11.6	884	82.4
Conv-Transformer					
CoaT-Lite-S [36]	224	19.8	4.1	2269	82.3
CrossViT-S [1]	240	26.9	5.1	2832	81.0
CrossViT-B [1]	240	105.0	20.1	1321	82.2
Visformer-S [2]	224	40.2	4.8	3676	82.1
NextViT-S [17]	224	31.7	5.8	3834	82.5
NextViT-B [17]	224	44.8	8.3	2926	83.2
NextViT-L [17]	224	57.8	10.8	2360	83.6
EfficientFormer-L1 [18]	224	12.3	1.31	6220	79.2
EfficientFormer-L3 [18]	224	31.4	3.9	2845	82.4
EfficientFormer-L7 [18]	224	82.2	10.2	1359	83.4
MaxViT-B [30]	224	120.0	23.4	507	84.9
MaxViT-L [30]	224	212.0	43.9	376	85.1
FasterViT-1 [10]	224	53.4	5.3	4188	83.2
FasterViT-2 [10]	224	75.9	8.7	3161	84.2
FasterViT-3 [10]	224	159.5	18.2	1780	84.9
Mamba-Based					
Vim-T [40]	224	7.0	-	3957	76.1
Vim-S [40]	224	26.0	-	1974	80.5
EfficientVMamba-T [25]	224	6.0	0.8	2904	76.5
EfficientVMamba-S [25]	224	11.0	1.3	1610	78.7
EfficientVMamba-B [25]	224	33.0	4.0	1482	81.8
SiMBA-S [24]	224	15.3	2.4	826	81.7
SiMBA-B [24]	224	22.8	4.2	624	83.5
VMamba-T [20]	224	30.0	4.9	1282	82.6
VMamba-S [20]	224	50.0	8.7	843	83.6
VMamba-B [20]	224	89.0	15.4	645	83.9
MambaVision					
MambaVision-T	224	31.8	4.4	6298	82.3
MambaVision-T2	224	35.1	5.1	5990	82.7
MambaVision-S	224	50.1	7.5	4700	83.3
MambaVision-B	224	97.7	15.0	3670	84.2
MambaVision-L	224	227.9	34.9	2190	85.0
MambaVision-L2	224	241.5	37.5	1021	85.3

shown in Table 2. To comprehensively validate MambaVision’s effectiveness, we trained models of varying sizes and compared them with popular vision backbones of comparable scale under identical conditions. Using a Cascade Mask

R-CNN [13] head, all variants of MambaVision demonstrated superior performance compared to their counterparts. Specifically, MambaVision models outperform ConvNeXt-T by +0.7 and +0.6, ConvNeXt-S by +0.4 and +0.2 and ConvNeXt-B by +0.1 and +0.1 in terms of box Average Precision (AP) and mask AP, respectively. Similarly, MambaVision outperforms Swin-T by +0.7 and +0.6, Swin-S by +0.4 and +0.2 and Swin-B by +0.9 and +0.7 in terms of box AP and mask AP, respectively. For semantic segmentation, we evaluated the performance on the ADE20K dataset [39] using UPerNet [34], as shown in Table 3. We observe that MambaVision models outperform similarly-sized competing models for different variants. For instance, MambaVision-T, MambaVision-S, and MambaVision-B outperform Swin-T, Swin-S, and Swin-B by +1.5, +0.6, and +1.0 in terms of mIoU, respectively. Notably, these improvements were achieved without extensive hyperparameter optimization for downstream tasks, highlighting MambaVision’s potential as a robust backbone for various vision tasks, particularly in high-resolution scenarios. Moreover, our approach consistently attains higher mIoU than Focal Transformers across all scales while having comparable model sizes.

5.3. Ablation

Large-scale Training on ImageNet-21K For the first time in any Mamba-based approach, our work (MambaVision) has scaled training to the large ImageNet-21K dataset with significantly bigger model sizes. As demonstrated in Fig. 4 the results are promising. Specifically, we observe meaningful improvements for the smaller MambaVision-B model (97.7M parameters), whose Top-1 accuracy increases from 84.2% to 84.9% at 224 resolution. In addition, pre-training and fine-tuning MambaVision-L raises its Top-1 accuracy from 85% to 86.1% at 224 resolution. We have also introduced a larger variant, MambaVision-L3 (739.6M parameters), which attains Top-1 accuracies of 87.3% and 88.1% at 256 and 512 resolutions, respectively. These results validate the scalability of our model across larger datasets, varying model sizes, and different image resolutions. To the best of our knowledge, MambaVision represents the first successful scaling of a Mamba-based vision architecture to ImageNet-21K with strong performance. This ability to scale is critical for real-world scenarios that rely on massive datasets, where bigger, more capable models are needed to achieve robust performance. We anticipate that MambaVision’s proven scalability will further encourage the adoption of Mamba-based models in industrial and large-scale research applications.

Design of Token Mixer We conducted a comprehensive ablation study to systematically design the MambaVision token mixer. Our investigation focused on adapting the Mamba block for computer vision tasks, evaluating performance across classification, object detection, instance segmentation,

Backbone	Params (M)	FLOPs (G)	AP ^{box}	AP ₅₀ ^{box}	AP ₇₅ ^{box}	AP ^{mask}	AP ₅₀ ^{mask}	AP ₇₅ ^{mask}
DeiT-Small/16 [28]	80	889	48.0	67.2	51.7	41.4	64.2	44.3
ResNet-50 [12]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T [21]	86	745	50.4	69.2	54.7	43.7	66.6	47.3
ConvNeXt-T [23]	86	741	50.4	69.1	54.8	43.7	66.5	47.3
MambaVision-T	86	740	51.1	70.0	55.6	44.3	67.3	47.9
X101-32 [35]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S [21]	107	838	51.9	70.7	56.3	45.0	68.2	48.8
ConvNeXt-S [23]	108	827	51.9	70.8	56.5	45.0	68.4	49.1
MambaVision-S	108	828	52.3	71.1	56.7	45.2	68.5	48.9
X101-64 [35]	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B [21]	145	982	51.9	70.5	56.4	45.0	68.1	48.9
ConvNeXt-B [23]	146	964	52.7	71.3	57.2	45.6	68.9	49.5
MambaVision-B	145	964	52.8	71.3	57.2	45.7	68.7	49.4

Table 2 – Object detection and instance segmentation benchmarks using Cascade Mask R-CNN [13] on MS COCO dataset [19]. All models are trained by using a $3 \times$ schedule and a crop resolution of 1280×800 .

Backbone	Param (M)	FLOPs (G)	mIoU
DeiT-Small/16 [28]	52	1099	44.0
Swin-T [21]	60	945	44.5
ResNet-101 [12]	86	1029	44.9
Focal-T [37]	62	998	45.8
MambaVision-T	55	945	46.0
Swin-S [21]	81	1038	47.6
Twins-SVT-B [3]	89	-	47.7
Focal-S [37]	85	1130	48.0
MambaVision-S	84	1135	48.2
Swin-B [21]	121	1188	48.1
Twins-SVT-L [3]	133	-	48.8
Focal-B [37]	126	1354	49.0
MambaVision-B	126	1342	49.1

Table 3 – Semantic segmentation results with UperNet [34] model using ADE20K dataset. All models are trained using a crop resolution of 512×512 .

and semantic segmentation. All experiments used a model architecture based on MambaVision-T configuration. As shown in Table 4, we began with the original Mamba formulation, which includes a causal convolution layer in the SSM branch (conv1) but lacks the additional convolution layer in our proposed symmetric branch (conv2). This baseline configuration achieved suboptimal performance across all metrics, with ImageNet Top-1 accuracy of 80.9% (-1.8%), MS COCO box AP of 44.8 (-1.6) and mask AP of 40.2 (-1.6), and ADE20K mIoU of 44.2% (-1.4). We then replaced the causal convolution in the SSM branch (conv1) with a regular convolution layer, which improved performance across all metrics. Subsequently, we added conv2 layer while maintaining Mamba’s original gating mechanism instead of concatenation, resulting in ImageNet Top-1 accuracy of 81.3%, MS COCO box AP of 45.3 and mask AP of 41.0, and ADE20K mIoU of 45.7%. Finally, implementing concatenation led to

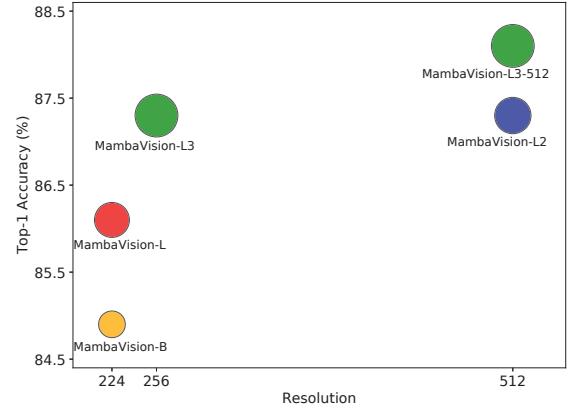


Figure 4 – Performance scalability of MambaVision ImageNet-21K pretrained models with varying model sizes and resolutions.

substantial improvements across all metrics, with gains of +1.0% in ImageNet Top-1, +1.1 in box AP and +0.8 in mask AP for MS COCO, and +0.9 in mIoU for ADE20K. These results validate our hypothesis that concatenating outputs from both branches (SSM and non-SSM) enables the model to learn richer feature representations and enhance global context understanding.

Hybrid Pattern We conducted a comprehensive study examining various hybrid integration patterns between self-attention and MambaVision token mixers. All experiments maintained the MambaVision-T architecture layout with iso-parameter models for fair comparison, implementing hybrid functionality in stages 3 and 4. Initial experiments with a random integration pattern yielded suboptimal results with a Top-1 accuracy of 81.3%, confirming our intuition that arbitrary self-attention placement may be ineffective. When we positioned self-attention blocks in the first N/2 layers of each

Table 4 – Systematic design of MambaVision token mixer. w/o and concat refer to "without" and concatenation. Conv1 and conv2 denote the conv operations in the SSM and additional symmetric branch as shown in Fig. 3. COCO experiments are performed using Mask-RCNN [13] head and $\times 1$ LR schedule.

	ImageNet top-1	COCO		ADE20k
		AP ^{box}	AP ^{mask}	mIoU
causal conv1 - w/o conv2	80.5	44.8	40.4	44.2
conv1 - w/o conv2	80.9	45.0	40.8	44.7
conv1 - conv2 - w/o concat	81.3	45.3	41.0	45.7
conv1 - conv2 - concat	82.3	46.4	41.8	46.0

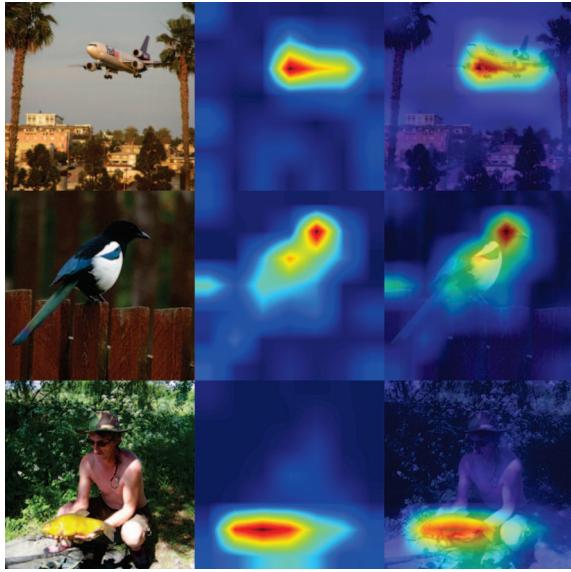


Figure 5 – Visualizations of MambaVision’s self-attention layers showing how the model learns to focus on semantically meaningful regions via attention maps (middle) and overlays (right).

stage (where N represents the total number of stage layers), performance improved by +0.2% (81.5%). A mixed layer pattern alternating between self-attention and MambaVision mixer blocks showed a slight performance decrease of -0.1% (81.4%), while reversing this order to MambaVision/self-attention improved accuracy to 81.6%. Placing self-attention blocks in only the last N/4 layers of each stage yielded a significant improvement of +0.3% (81.9%), supporting our hypothesis that self-attention is most effective in the final layers. Further optimization revealed that extending self-attention to the last N/2 layers of each stage achieved the best performance at 82.3%, indicating the importance of carefully balancing self-attention blocks with MambaVision layers for optimal representation learning.

Interpretability To better understand how MambaVision processes visual information, we visualize the attention maps from the self-attention layers in the final stages. As shown

Table 5 – Ablation study of on the effectiveness of different hybrid integration patterns. S and M denote self-attention and MambaVision token mixer blocks, respectively.

Model	Pattern	Params (M)	Top-1
Random	-	31.8	81.3
First N/2 layers	SSSSMMMM	31.8	81.5
Mixed layers-1	SMSMSMSM	31.8	81.4
Mixed layers-2	MSMSMSMS	31.8	81.6
Last N/4 layers	MMMMMMSS	31.8	81.9
Last N/2 layers	MMMMSSSS	31.8	82.3

in Fig. 5, these visualizations reveal that the model learns to focus on semantically meaningful regions without explicit supervision. In the aircraft example, the attention clearly highlights the entire plane body, suggesting effective capture of object boundaries. For the bird image, we observe concentrated attention on distinctive features like the head and tail regions, demonstrating the model’s ability to identify fine-grained details. In the case of object-human interaction (bottom row), the attention map shows strong activation on both the subject and the object being held, indicating that the self-attention layers successfully model relationships between different elements in the scene. These visualizations support our architectural design choice of using self-attention blocks in the final stages to capture global context and long-range dependencies.

6. Conclusion

In this work, we introduced MambaVision which is the first Mamba-Transformer hybrid backbone specifically tailored for vision applications. We proposed re-design of Mamba formulation to enhance global context representation learning capability. MambaVision achieves a new SOTA Pareto front in terms of Top-1 accuracy and image throughput, outperforming Transformer and Mamba-based models by a significant margin. Through extensive experimentation across multiple vision tasks, including classification, detection and segmentation, we demonstrated the versatility and effectiveness of our approach. Our systematic analysis of integration patterns revealed that positioning self-attention blocks in the final layers significantly improves the model’s ability to capture long-range dependencies while maintaining efficiency. Furthermore, we successfully scaled MambaVision to ImageNet-21K pretraining, achieving strong performance that matches SOTA models, demonstrating its potential for large-scale vision applications. The success of MambaVision in addressing the limitations of pure Mamba-based architectures while leveraging their strengths opens new possibilities for vision backbone design. We hope these findings could be the foundation for a new class of hybrid vision models.

References

- [1] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. [2](#), [6](#)
- [2] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021. [6](#)
- [3] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. [2](#), [6](#), [7](#)
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [5](#), [6](#)
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. [2](#)
- [6] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107: 3–11, 2018. [4](#)
- [7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. [1](#), [4](#), [5](#)
- [8] 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. *Advances in neural information processing systems*, 34:572–585, 2021. [3](#)
- [9] Kai Han, An Xiao, Enhua Wu, Jianyu Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. [6](#)
- [10] Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention. *arXiv preprint arXiv:2306.06189*, 2023. [2](#), [6](#)
- [11] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pages 12633–12646. PMLR, 2023. [5](#)
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [2](#), [7](#)
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. [5](#), [6](#), [7](#), [8](#)
- [14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. [3](#)
- [15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. [3](#)
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. [2](#)
- [17] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022. [2](#), [6](#)
- [18] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022. [2](#), [6](#)
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. [5](#), [7](#)
- [20] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. [3](#), [6](#)
- [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. [2](#), [5](#), [6](#), [7](#)
- [22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. [5](#), [6](#)
- [23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. [2](#), [6](#), [7](#)
- [24] Badri N Patro and Vijay S Agneeswaran. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024. [3](#), [6](#)
- [25] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. [1](#), [3](#), [6](#)
- [26] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. [2](#), [6](#)
- [27] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. [2](#), [6](#)

- [28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. [2](#), [6](#), [7](#)
- [29] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 516–533. Springer, 2022. [6](#)
- [30] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 459–479. Springer, 2022. [6](#)
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. [1](#)
- [32] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. [2](#)
- [33] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. [6](#)
- [34] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. [5](#), [6](#), [7](#)
- [35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. [7](#)
- [36] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990, 2021. [2](#), [6](#)
- [37] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. [5](#), [7](#)
- [38] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829, 2022. [6](#)
- [39] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. [5](#), [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. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62429–62442. PMLR, 2024. [1](#), [2](#), [6](#)