

EfficientViM: Efficient Vision Mamba with Hidden State Mixer based State Space Duality

Sanghyeok Lee¹ Joonmyung Choi¹ Hyunwoo J. Kim^{2*}

¹Korea University ²KAIST

{cat0626, pizard}@korea.ac.kr hyunwoojkim@kaist.ac.kr

Abstract

For the deployment of neural networks in resource-constrained environments, prior works have built lightweight architectures with convolution and attention for capturing local and global dependencies, respectively. Recently, the state space model (SSM) has emerged as an effective operation for global interaction with its favorable linear computational cost in the number of tokens. To harness the efficacy of SSM, we introduce Efficient Vision Mamba (**EfficientViM**), a novel architecture built on hidden state mixer-based state space duality (**HSM-SSD**) that efficiently captures global dependencies with further reduced computational cost. With the observation that the runtime of the SSD layer is driven by the linear projections on the input sequences, we redesign the original SSD layer to perform the channel mixing operation within compressed hidden states in the HSM-SSD layer. Additionally, we propose multi-stage hidden state fusion to reinforce the representation power of hidden states and provide the design to alleviate the bottleneck caused by the memory-bound operations. As a result, the EfficientViM family achieves a new state-of-the-art speed-accuracy trade-off on ImageNet-1k, offering up to a 0.7% performance improvement over the second-best model SHViT with faster speed. Further, we observe significant improvements in throughput and accuracy compared to prior works, when scaling images or employing distillation training. Code is available at <https://github.com/mlvlab/EfficientViM>.

1. Introduction

Efficient vision architectures for resource-constrained environments have been consistently explored in computer vision tasks, including image classification [5, 23, 27, 48, 49, 64, 88], object detection [34, 55, 65, 73], segmenta-

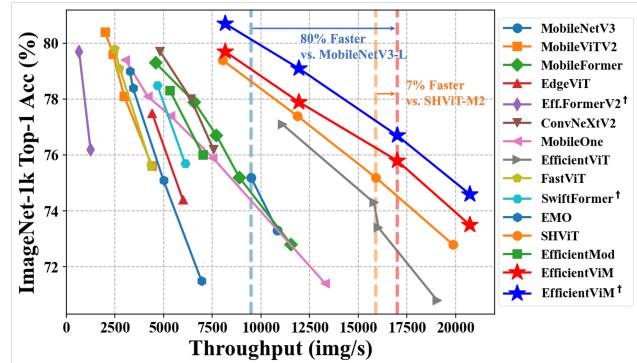


Figure 1. Comparison of efficient networks on ImageNet-1k [10] classification. The family of our EfficientViM, marked as red and blue stars, shows the best speed-accuracy trade-offs. † indicates the model trained with distillation following [67].

tion [24, 77, 79, 80, 89], etc. Earlier works [5, 18, 46] have explored the efficient convolutional neural networks (CNN). One such technique is depthwise separable convolution (DWConv), introduced in Xception [5], which has been widely adopted for modeling lightweight architectures, especially CNN-based networks including MobileNet [22, 23, 57] and following works [64, 71]. Meanwhile, with the advent of Vision Transformer (ViT) [12], attention mechanisms have become a key operation for capturing long-range dependencies within image patches. However, the quadratic cost of self-attention presents challenges for designing efficient architectures. To address this, prior works have attempted to approximate self-attention with reduced cost [6, 31, 74, 78], or restrict the number of tokens [1, 4, 7, 33, 42, 82]. For on-device deployment, several works [3, 40, 48, 49, 52, 70, 84, 87] devoted effort to developing hybrid ViTs combined with CNNs (DWConv). While prior works have demonstrated superior performance over traditional CNNs, the inherent quadratic complexity of self-attention in the number of tokens remains a major bottleneck limiting their efficiency and scalability. Recently, state space models (SSMs) [9, 14–16, 62] have emerged as a promising alternative to self-attention, of-

*Corresponding author.

fering a favorable linear computational complexity while maintaining a global receptive field. Mamba [9, 15] has introduced selective scanning mechanisms on SSM (S6) to process sequences with the hardware-aware parallel algorithm. Following Mamba, vision Mambas [26, 41, 81, 93] have extended the concept of SSM to vision tasks. These works have studied multi-path scanning mechanisms to address the causal constraints of SSM that are not desirable for image processing. More recent works like VSSD [60] and Linfusion [39] further eliminated the causal mask in the state space duality (SSD) of Mamba2 [9], introducing non-causal state space duality (NC-SSD). While vision Mambas demonstrate improved performance over previous SOTA methods, they still have relatively slower speeds than lightweight vision models. Further, we have observed that the major bottleneck of the previous SSD is the linear projection in gating operation and output projection.

In this paper, we present an **Efficient Vision Mamba (EfficientViM)**, a new family of mamba-based lightweight vision backbone built with a fast and effective SSD layer called **Hidden State Mixer-based SSD (HSM-SSD)**. In the HSM-SSD layer, we transfer the channel mixing operations of the standard SSD layer, including linear projection and gating function, from the image feature space to the hidden state space. These hidden states serve as compressed latent representations of the input. We observe that this design mitigates the major bottleneck of the SSD layer while maintaining the generalization ability of the model.

Also, we introduce a multi-stage hidden state fusion approach that generates the predictions by combining the original logits with those derived from the hidden states at each stage, leading to an enhanced representation power of hidden states. After breaking down the runtime of the HSM-SSD layer, we present the macro design that minimizes memory-bound operations, prioritizing practical performance in real-world applications over theoretical metrics like FLOPs. Through extensive experiments, we demonstrate that our EfficientViM achieves the new speed-accuracy state-of-the-art trade-off as shown in Figure 1. In particular, EfficientViM-M2 outperforms the previous SOTA model SHViT [84], and pioneering work ModelNetV3 [22] with a 0.6% performance improvement even bringing about 7% and 80% speed-ups, respectively. In summary, the contributions of EfficientViM are threefold:

- We propose a novel mamba-based lightweight architecture called EfficientViM, leveraging the linear computational cost of the global token mixer.
- We introduce HSM-SSD, which makes the major overhead of the SSD layer controllable by adjusting the number of hidden states.
- With a design minimizing memory-bound operations and incorporating multi-stage hidden state fusion, EfficientViM achieves the best speed-accuracy tradeoffs.

2. Preliminaries

State space models (SSM). Inspired by the linear time-invariant (LTI) continuous system, an SSM maps an input sequence $x(t) \in \mathbb{R}$ to an output sequence $y(t) \in \mathbb{R}$ as

$$h'(t) = \hat{\mathbf{A}}h(t) + \hat{\mathbf{B}}x(t), \quad y(t) = \hat{\mathbf{C}}h(t), \quad (1)$$

where $h(t) \in \mathbb{R}^{N \times 1}$ is a hidden state, $\hat{\mathbf{A}} \in \mathbb{R}^{N \times N}$, $\hat{\mathbf{B}} \in \mathbb{R}^{N \times 1}$, $\hat{\mathbf{C}} \in \mathbb{R}^{1 \times N}$ are the projection matrix, and N is the number of states. To adapt this continuous-time system for discrete data in deep learning, given the multivariate input sequences $\mathbf{x} = [\mathbf{x}_1^\top, \dots, \mathbf{x}_L^\top]^\top \in \mathbb{R}^{L \times D}$ with $\forall_t \mathbf{x}_t \in \mathbb{R}^{1 \times D}$, Mamba [15] first generates parameters as $\hat{\mathbf{B}}, \mathbf{C} = \mathbf{x}\mathbf{W}_\mathbf{B}, \mathbf{x}\mathbf{W}_\mathbf{C} \in \mathbb{R}^{L \times N}$, $\Delta = \mathbf{x}\mathbf{W}_\Delta \in \mathbb{R}^L$, where $\mathbf{W}_\mathbf{B}, \mathbf{W}_\mathbf{C} \in \mathbb{R}^{D \times N}$, $\mathbf{W}_\Delta \in \mathbb{R}^{D \times 1}$ are learnable matrices. Then, the discretized form of SSM with zero-order hold discretization is defined as

$$\mathbf{h}_t = \mathbf{A}_t \mathbf{h}_{t-1} + \mathbf{B}_t^\top \mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}_t \mathbf{h}_t, \quad (2)$$

where $\mathbf{y} \in \mathbb{R}^{L \times D}$, $\mathbf{h}_t \in \mathbb{R}^{N \times D}$, $\mathbf{A}_t = e^{\Delta_t \hat{\mathbf{A}}} \in \mathbb{R}^{N \times N}$, $\mathbf{B}_t = (\Delta_t \hat{\mathbf{A}})^{-1}(e^{\Delta_t \hat{\mathbf{A}}} - \mathbf{I}) \cdot \Delta_t \hat{\mathbf{B}}_t \approx \Delta_t \hat{\mathbf{B}}_t \in \mathbb{R}^{1 \times N}$. In this formulation, $\hat{\mathbf{A}} \in \mathbb{R}^{N \times N}$ is a learnable diagonal matrix, and all projection matrices $\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t$ enable the linear time-variant discrete system that selectively attends to inputs \mathbf{x} and hidden state \mathbf{h} of each timestamp t .

State space duality (SSD). Mamba2 [9] further simplifies the diagonal form of the evolution matrix $\hat{\mathbf{A}}$ into the scalar form as $\hat{a} \in \mathbb{R}$ resulting in $\mathbf{a} \in \mathbb{R}^L$ via the same discretization step. Then, the state space duality (SSD) reformulates Equation (2) as matrix transformation:

$$\mathbf{y} = \text{SSD}(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C}) = (\mathbf{M} \odot (\mathbf{C}\mathbf{B}^\top)) \mathbf{x},$$

$$\mathbf{M}_{ij} = \begin{cases} \prod_{k=j+1}^i \mathbf{a}_k & \text{if } i > j \\ 1 & \text{if } i = j \\ 0 & \text{if } i < j \end{cases}, \quad (3)$$

where $\mathbf{M} \in \mathbb{R}^{L \times L}$, and \odot indicates Hadamard product. Note that the lower triangular matrix \mathbf{M} acts as a causal mask, which is suboptimal for image processing. To address this, non-causal SSD (NC-SSD) [39, 93] has been studied as an alternative to SSD by defining the mask as $\mathbf{M}_{ij} = \prod_{k=j+1}^L \mathbf{a}_k$, resulting in $\mathbf{M}_{1j} = \mathbf{M}_{2j} = \dots = \mathbf{M}_{Lj}$. Further, in VSSD [93], it is simplified as $\mathbf{M}_{ij} = \mathbf{a}_j$ resulting in

$$\mathbf{y} = \text{NC-SSD}(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C}) = \mathbf{Ch},$$

$$\mathbf{h} = (\mathbf{a} \mathbb{1}_N^\top \odot \mathbf{B})^\top \mathbf{x} = \sum_{i=1}^L \mathbf{a}_i \mathbf{B}_i^\top \mathbf{x}_i, \quad (4)$$

where $\mathbb{1}_N \in \mathbb{R}^N$ is one vector for broadcasting \mathbf{a} . Since the cumulative multiplication of \mathbf{a} restricts the receptive fields as discussed in [60, 61], we adopt this version of NC-SSD as our starting point for an efficient token mixer.

Complexity	Method
$\mathcal{O}(LD^2 + L^2D)$	Attention [72]
$\mathcal{O}(LD^2)$	Linear Attention [29]
$\mathcal{O}(LD^2 + LND)$	SSD [9], NC-SSD [60]
$\mathcal{O}(ND^2 + LND)$	HSM-SSD

Table 1. Complexity comparison of global tokens mixers. L : # tokens, N : # states, D : # channels.

3. Method

We present a hidden state mixer-based SSD (**HSM-SSD**) for capturing global context with reduced costs, detailed in Section 3.1. Then, we discuss the additional techniques to improve both speed and performance with HSM-SSD in Section 3.2. After that, we outline the overall architecture and block designs to construct **EfficientViM** in Section 3.3.

3.1. Hidden State Mixer-based SSD

We start with a brief discussion on the computational cost of the NC-SSD layer depicted in Figure 2a. The entire process of the NC-SSD layer can be summarized as

$$\hat{\mathbf{B}}, \mathbf{C}, \Delta, \mathbf{x}, \mathbf{z} = \text{Linear}(\mathbf{x}_{\text{in}}) \quad (5)$$

$$\mathbf{a}, \mathbf{B} = \text{Discretization}(\hat{\mathbf{a}}, \hat{\mathbf{B}}, \Delta), \quad (6)$$

$$\mathbf{B}, \mathbf{C}, \mathbf{x} := \text{DWConv}(\mathbf{B}, \mathbf{C}, \mathbf{x}), \quad (7)$$

$$\mathbf{y} = \text{NC-SSD}(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C}), \quad (8)$$

$$\mathbf{x}_{\text{out}} = \text{Linear}(\mathbf{y} \odot \sigma(\mathbf{z})), \quad (9)$$

where $\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{out}}, \mathbf{x}, \mathbf{z} \in \mathbb{R}^{L \times D}$, and σ is the activation function. The computational costs of Equations (5) to (7) with a constant kernel size is $\mathcal{O}(LD^2 + LND)$. Subsequently, executing NC-SSD and the output projection requires $\mathcal{O}(LND)$ and $\mathcal{O}(LD^2)$, respectively. Given that the number of states N is typically much smaller than the number of channels D (i.e., $N \ll D$), the overall complexity is mainly driven by the linear projections involved in generating \mathbf{x} , \mathbf{z} , and \mathbf{x}_{out} , leading to $\mathcal{O}(LD^2)$. Therefore, optimizing the linear projection in SSD blocks is crucial for scalability.

We delve into optimizing these computations for efficient layer designs. NC-SSD (Equation (4)) can be factorized into two steps. First, it obtains shared global hidden state $\mathbf{h} \in \mathbb{R}^{N \times D}$ through a weighted linear combination of input states $\mathbf{B}_i^\top \mathbf{x}_i \in \mathbb{R}^{N \times D}$ using importance weights $\mathbf{a} \in \mathbb{R}^L$. Second, the outputs for each input are generated by projecting a hidden state with their corresponding $\mathbf{C} \in \mathbb{R}^{L \times N}$. Here, if we denote the projected input \mathbf{x} as $\mathbf{x}_{\text{in}} \mathbf{W}_{\text{in}}$ removing DWConv, the following holds:

$$\begin{aligned} \mathbf{h} &= (\mathbf{a} \mathbf{1}_N^\top \odot \mathbf{B})^\top (\mathbf{x}_{\text{in}} \mathbf{W}_{\text{in}}) \\ &= ((\mathbf{a} \mathbf{1}_N^\top \odot \mathbf{B})^\top \mathbf{x}_{\text{in}}) \mathbf{W}_{\text{in}} = \mathbf{h}_{\text{in}} \mathbf{W}_{\text{in}}, \end{aligned} \quad (10)$$

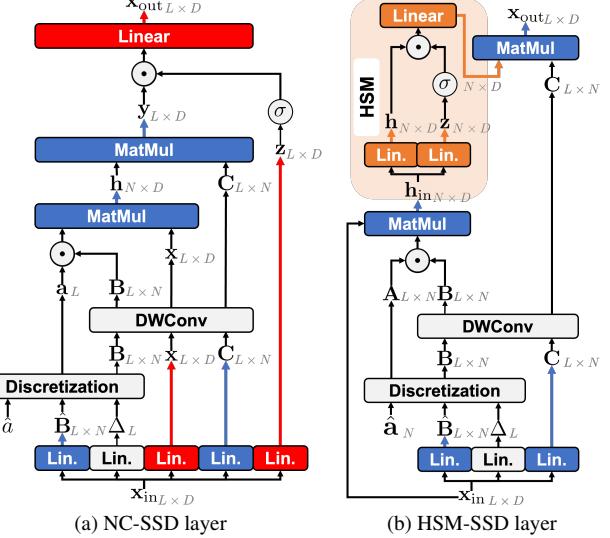


Figure 2. Illustration of (left) NC-SSD and (right) HSM-SSD layer. In the HSM-SSD layer, the computationally heavy projections are handled with the reduced hidden state in HSM as highlighted. Red, blue, and orange colors indicate the operation requiring the complexities of $\mathcal{O}(LD^2)$, $\mathcal{O}(LND)$, and $\mathcal{O}(ND^2)$.

where $\mathbf{W}_{\text{in}} \in \mathbb{R}^{D \times D}$, and $\mathbf{h}_{\text{in}} = (\mathbf{a} \mathbf{1}_N^\top \odot \mathbf{B})^\top \mathbf{x}_{\text{in}} \in \mathbb{R}^{N \times D}$. By computing \mathbf{h}_{in} first, we perform a linear projection onto the hidden state space. This approach reduces the cost from $\mathcal{O}(LD^2)$ to $\mathcal{O}(ND^2)$ which relies on the number of states. In other words, we could alleviate the major costs of the layer by adjusting the number of states such that $N \ll L$.

Hidden State Mixer. The next step is to alleviate the cost of gating and output projection in Equation (9), which still remains $\mathcal{O}(LD^2)$. To address this, we focus on a shared global hidden state \mathbf{h} . Note that the hidden state \mathbf{h} itself is the reduced latent array that compresses the input data with a significantly smaller sequence length N . Based on this observation, we propose a *Hidden State Mixer (HSM)* that performs channel mixing, including the gating and output projection, directly on the reduced latent array \mathbf{h} as highlighted in Figure 2b. To this end, we approximate the output of the NC-SSD layer as follows:

$$\begin{aligned} \mathbf{x}_{\text{out}} &= f(\mathbf{y}) \\ &= \text{Linear}(\mathbf{y} \odot \sigma(\mathbf{z})) \\ &= (\mathbf{C}\mathbf{h} \odot \sigma(\mathbf{x}_{\text{in}} \mathbf{W}_{\text{z}})) \mathbf{W}_{\text{out}} \\ &\approx \mathbf{C}((\mathbf{h} \odot \sigma(\mathbf{h}_{\text{in}} \mathbf{W}_{\text{z}})) \mathbf{W}_{\text{out}}) = \mathbf{C}f(\mathbf{h}), \end{aligned} \quad (11)$$

where $\mathbf{y} = \mathbf{C}\mathbf{h}$ from Equation (4), and f indicates channel mixing of gating function followed by linear projection with the learnable matrix $\mathbf{W}_{\text{z}}, \mathbf{W}_{\text{out}} \in \mathbb{R}^{D \times D}$. Contrary to the original NC-SSD layer where $\mathbf{C}\mathbf{h}$ is computed first and then fed to f , we apply the gating and projection directly to the hidden states with HSM. Then, the final output \mathbf{x}_{out} is generated by projecting updated hidden states with \mathbf{C} . Consequently, the total complexity of capturing global

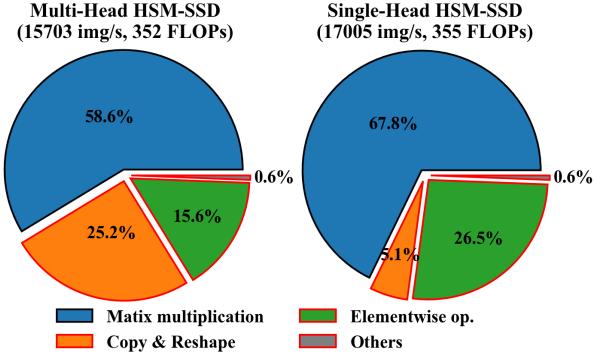


Figure 3. **Runtime breakdown of HSM-SSD with EfficientViM-M2.** The operations highlighted in red are memory-bound.

context in the HSM-SSD layer becomes $\mathcal{O}(ND^2 + LND)$, which is negligible as N gets smaller. Refer to Table 1 for a comparison of the big- \mathcal{O} complexities with previous global token mixers.

Proposition 1. Let $N = L$, $\mathbf{a} \mathbb{1}_L^\top \odot \mathbf{B} = \mathbb{I}_L$, and $\mathbf{C} \in \mathbb{R}^{L \times L}$ be diagonal. Then, $HSM-SSD(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C})$ is equivalent to $NC-SSD(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C})$ including gating and output projection, as $\mathbf{x}_{out} = f(\mathbf{y}) = \mathbf{C}f(\mathbf{h})$. See supplement for proof.

Remark 1. Although we utilize gating and linear projection for the HSM to emulate the operation in the original SSD layer, other methods are also available. For example, inspired by the Perceiver architecture [28], we could implement a global token mixer for the function f in Eq. (11) to efficiently handle high-dimensional inputs using a reduced set of latent variables. Moreover, HSM-SSD can recursively apply HSM-SSD itself as the hidden state mixer, further reducing computational complexity and forming a recurrent architecture unrolled across depth.

3.2. HSM-SSD layer

Multi-stage hidden state fusion. To further improve the performance of EfficientViM, we introduce a *multi-stage hidden-state fusion (MSF)* mechanism that fuses the prediction logits leveraging hidden states from multiple stages of the network. Let $\{\mathbf{h}^{(s)}\}_{s=1}^S$ denote the hidden states at the last block of each stage s , where S is the total number of stages. For each $\mathbf{h}^{(s)}$, we compute a global representation $\hat{\mathbf{h}}^{(s)}$ through a simple average over the hidden states:

$$\hat{\mathbf{h}}^{(s)} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_i^{(s)}. \quad (12)$$

Then, each global representation $\hat{\mathbf{h}}^{(s)} \in \mathbb{R}^D$ is normalized and projected to generate its corresponding logits $\mathbf{z}^{(s)} \in \mathbb{R}^c$, where c indicates the number of classes. We set the final logit \mathbf{z} of EfficientViM as a weighted sum of the logits from all stages, including the original logit $\mathbf{z}^{(0)}$ obtained

Algorithm 1 HSM-SSD Layer

Input: $\mathbf{x}_{in} \in \mathbb{R}^{L \times D}$
Output: $\mathbf{x}_{out} \in \mathbb{R}^{L \times D}$

1: $\hat{\mathbf{B}}, \mathbf{C}, \Delta \leftarrow \text{Linear}(\mathbf{x}_{in})$	$\triangleright \mathcal{O}(LND)$
2: $\hat{\mathbf{B}}, \mathbf{C} \leftarrow \text{DWConv}(\hat{\mathbf{B}}, \mathbf{C})$	$\triangleright \mathcal{O}(LNK^2D)$
3: $\mathbf{A}, \mathbf{B} \leftarrow \text{Discretization}(\hat{\mathbf{a}}, \hat{\mathbf{B}}, \Delta)$	$\triangleright \mathcal{O}(LD)$
4: $\mathbf{h}_{in} \leftarrow (\mathbf{A} \odot \mathbf{B})^\top \mathbf{x}_{in}$	$\triangleright \mathcal{O}(LND)$
5: $\mathbf{h}, \mathbf{z} \leftarrow \text{Linear}(\mathbf{h}_{in})$	$\triangleright \mathcal{O}(ND^2)$
6: $\mathbf{h} \leftarrow \text{Linear}(\mathbf{h} \odot \sigma(\mathbf{z}))$	$\triangleright \mathcal{O}(ND^2)$
7: $\mathbf{x}_{out} \leftarrow \mathbf{Ch}$	$\triangleright \mathcal{O}(LND)$
8: Return \mathbf{x}_{out}	

from the output of the last stage, which is defined as

$$\begin{aligned} \mathbf{z} &= \sum_{s=0}^S \hat{\beta}^{(s)} \mathbf{z}^{(s)}, \\ \hat{\beta}^{(s)} &= \frac{\exp(\beta^{(s)})}{\sum_{i=0}^S \exp(\beta^{(i)})}, \end{aligned} \quad (13)$$

where $\beta^{(s)}$ is learnable scalar. By training with this combined logit, we explicitly reinforce the representational power of the hidden states, as they contribute to the final predictions. It also enriches the information by integrating both low-level and high-level features, thereby enhancing the generalization ability of the model during inference.

Single-head HSM-SSD. The multi-head design in the attention mechanism [72] allows it to selectively attend to the features from independent representation subspaces within each head. SSD-based models [9, 60] generally adopt a multi-head variant called multi-input SSD, where the input \mathbf{x} and \mathbf{a} are defined for each head while \mathbf{B} , and \mathbf{C} are shared across the head. However, recent work [84] has pointed out that a significant portion of real runtime in multi-head self-attention is driven by memory-bound operations.

In our preliminary experiments, we also found that multi-head configuration has become a bottleneck for HSM-SSD as summarized in Figure 3. As shown in the figure, the real runtime of multi-head HSM-SSD is largely bounded by memory access, requiring almost a quarter of the total runtime. Hence, we eliminate all tensor manipulation caused by multi-head (e.g., reshape, copy operation). Concurrently, to mimic the capability of multi-head in capturing diverse relationships, we set $\Delta \in \mathbb{R}^{L \times N}$, $\hat{\mathbf{a}} \in \mathbb{R}^N$, enabling the importance weights $\mathbf{A} \in \mathbb{R}^{L \times N}$ to estimate the importance of the tokens per state. Then, the input for the hidden state mixer becomes

$$\mathbf{h}_{in} = (\mathbf{A} \odot \mathbf{B})^\top \mathbf{x}_{in}. \quad (14)$$

As a result, our single-head HSM-SSD with state-wise importance weights achieved higher throughput (17,005 img/s) with single-head compared to multi-head (15,703

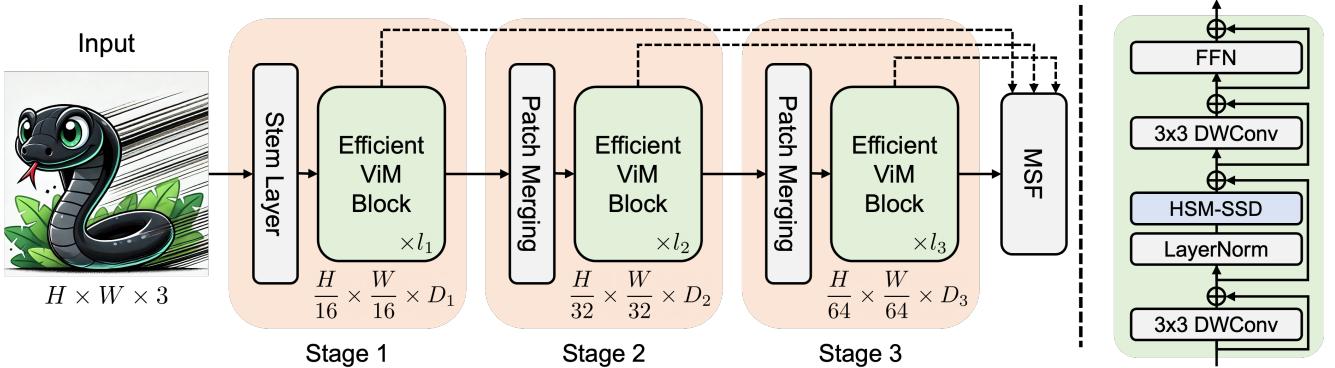


Figure 4. **(left)** Overall architecture and **(right)** block design of EfficientViM. The dotted line indicates a skip connection for multi-stage hidden state fusion (MSF). Illustration of the HSM-SSD layer in the EfficientViM block is presented in Figure 2.

img/s) along with competitive performance (see Table 9.b) under similar FLOPs. The pseudocode of the single-head HSM-SSD layer is provided in Algorithm 1.

3.3. EfficientViM

In this subsection, we present EfficientViM, an efficient vision mambas built upon the HSM-SSD layer. The overall architecture is illustrated in Figure 4.

Block design. In each block of EfficientViM, we sequentially stack the HSM-SSD layer and a feed-forward network (FFN) to facilitate global information aggregation and channel interaction, respectively. FFN layer consists of two consecutive 1×1 convolution layers, known as pointwise convolution, with an expansion ratio of 4. To capture the local context with minimal computational costs, we incorporate a 3×3 depthwise convolution (DWConv) layer before both the NC-SSD and FFN layers. Each layer is combined with a residual connection using a layer scale following [68, 69]. For normalization, we apply layer normalization (LN) only before the HSM-SSD layer for numerical stability, while batch normalization (BN) is used for DWConv and FFN considering its faster speed over LN.

Overall architecture. The stem layer first maps the $H \times W \times 3$ input image to the down-sized feature map $\frac{H}{16} \times \frac{W}{16} \times D_1$ through the four consecutive 3×3 convolutional layers with the stride of 2. Then, the resulting feature map is fed into the three stages built with EfficientViM blocks. To achieve hierarchical architecture and improve efficiency, we downscale the feature map while increasing the number of channels at the end of each stage via the down-sampling layer adopted from [40, 57, 93]. Regarding activation functions, we use SiLU only in the HSM-SSD layer, and others use ReLU [51] since the latencies of the complex activation functions (*e.g.*, Gelu [21], DynamicReLU [2], etc.) are largely dependent on devices, as discussed in previous works [40, 71, 84]. Detailed architecture specifications of EfficientViM variants are available in Table 2.

Model	# Blocks	# Channels	# States
EfficientViM-M1	[2, 2, 2]	[128, 192, 320]	[49, 25, 9]
EfficientViM-M2	[2, 2, 2]	[128, 256, 512]	[49, 25, 9]
EfficientViM-M3	[2, 2, 2]	[224, 320, 512]	[49, 25, 9]
EfficientViM-M4	[3, 4, 2]	[224, 320, 512]	[64, 32, 16]

Table 2. Specification of EfficientViM variants.

4. Experiments

In this section, we first demonstrate the effectiveness of EfficientViM on image classification (Section 4.1). Then, we conduct experiments to analyze the extensibility of EfficientViM on dense predictions (Section 4.2). We also provide ablation studies and analysis of EfficientViM (Section 4.3). See the supplement for implementation details and more experiments.

4.1. Image Classification.

Comparison with efficient vision backbones. For comparison of EfficientViM with prior works, we conduct ImageNet-1K [10] classification. To validate the effectiveness of EfficientViM in speed-accuracy trade-offs, we measure the throughput (im/s), and latency (ms) with the batch size of 256 on an NVIDIA RTX 3090 along with the accuracy, and present the results in Table 3. EfficientViM outperforms all previous efficient networks in both speed and accuracy. After training 450 epochs, EfficientViM-M1 shows a competitive performance with MobileNetV3-L 0.75 [22] and EfficientViT-M3 [40] while achieving about 90% and 30% speedup. Further, EfficientViM-M2 achieves about 4× faster speed, with a 0.2% performance improvement compared to MobileViTV2 0.75 [49] and FastViT-T8 [70]. EfficientViM-M3 and M4 achieve 77.9% and 79.7% accuracy, respectively, surpassing all previous works in throughput and accuracy within each section. Also, EfficientViM consistently outperforms the previous SOTA network, SHViT [84], across model sizes while reducing latency, which demonstrates the superiority of EfficientViM.

Method	Venue	Input Size	Epochs	Token Mixer	Throughput (im/s) \uparrow	Thr _{rel} \uparrow	Latency (ms) \downarrow	Top-1 (%) \uparrow	Params (M)	FLOPs (M)
MobileViTV2 0.5 [49]	Arxiv 2022	256 ²	300	Att.	6,702	$\times 0.32$	0.149	70.2	1.4	466
MobileOne-S0 [71]	CVPR 2023	224 ²	300	Conv	13,313	$\times 0.64$	0.075	71.4	2.1	275
EMO-1M [87]	ICCV 2023	224 ²	300	Att.	6,945	$\times 0.34$	0.144	71.5	1.3	261
MobileFormer-96M [3]	CVPR 2022	224 ²	450	Att.	11,554	$\times 0.56$	0.087	72.8	4.6	96
SHViT-S1 [84]	CVPR 2024	224 ²	300	Att.	19,868	$\times 0.96$	0.050	72.8	6.3	241
EfficientViM-M1	-	224 ²	300	SSD	20,731	$\times 1.00$	0.048	72.9	6.7	239
MobileNetV3-L 0.75 [22]	ICCV 2019	224 ²	600	Conv	10,846	$\times 0.52$	0.092	73.3	4.0	155
EfficientViT-M3 [40]	CVPR 2023	224 ²	300	Att.	16,045	$\times 0.77$	0.062	73.4	6.9	263
EfficientViM-M1	-	224 ²	450	SSD	20,731	$\times 1.00$	0.048	73.5	6.7	239
EfficientFormerV2-S0 [36]	NeurIPS 2022	224 ²	300	Att.	1,350	$\times 0.08$	0.741	73.7	3.5	407
EfficientViT-M4 [40]	CVPR 2023	224 ²	300	Att.	15,807	$\times 0.93$	0.063	74.3	8.8	299
EdgeViT-XXS [52]	ECCV 2022	224 ²	300	Att.	5,990	$\times 0.35$	0.167	74.4	4.1	556
EMO-2M [87]	ICCV 2023	224 ²	300	Att.	4,990	$\times 0.29$	0.200	75.1	2.3	439
MobileNetV3-L 1.0 [22]	ICCV 2019	224 ²	600	Conv	9,493	$\times 0.56$	0.105	75.2	5.4	217
MobileFormer-151M [3]	CVPR 2022	224 ²	450	Att.	8,890	$\times 0.52$	0.112	75.2	7.6	151
SHViT-S2 [84]	CVPR 2024	224 ²	300	Att.	15,899	$\times 0.93$	0.063	75.2	11.4	366
EfficientViM-M2	-	224 ²	300	SSD	17,005	$\times 1.00$	0.059	75.4	13.9	355
MobileViTV2 0.75 [49]	Arxiv 2022	256 ²	300	Att.	4,409	$\times 0.26$	0.227	75.6	2.9	1030
FastViT-T8 [70]	ICCV 2023	256 ²	300	Att.	4,365	$\times 0.26$	0.229	75.6	3.6	705
EfficientViM-M2	-	224 ²	450	SSD	17,005	$\times 1.00$	0.059	75.8	13.9	355
EfficientMod-XXS [47]	ICLR 2024	224 ²	300	Att.	7022	$\times 0.59$	0.142	76.0	4.7	583
ConvNeXtV2-A [76]	CVPR 2023	224 ²	300	Conv	7563	$\times 0.63$	0.132	76.2	3.7	552
EfficientViT-M5 [40]	CVPR 2023	224 ²	300	Att.	11,105	$\times 0.93$	0.090	77.1	12.4	522
MobileOne-S2 [71]	CVPR 2023	224 ²	300	Conv	5,360	$\times 0.45$	0.187	77.4	7.8	1299
SHViT-S3 [84]	CVPR 2024	224 ²	300	Att.	11,873	$\times 0.99$	0.084	77.4	14.2	601
EdgeViT-XS [52]	ECCV 2022	224 ²	300	Att.	4,405	$\times 0.37$	0.227	77.5	6.7	1136
EfficientViM-M3	-	224 ²	300	SSD	11,952	$\times 1.00$	0.084	77.6	16.6	656
MobileFormer-294M [3]	CVPR 2022	224 ²	450	Att.	6,576	$\times 0.55$	0.152	77.9	11.4	294
EfficientFormerV2-S1 [36]	NeurIPS 2022	224 ²	300	Att.	1,248	$\times 0.10$	0.801	77.9	6.1	668
EfficientViM-M3	-	224 ²	450	SSD	11,952	$\times 1.00$	0.084	77.9	16.6	656
ConvNeXtV2-F [76]	CVPR 2023	224 ²	300	Conv	6,405	$\times 0.78$	0.156	78.0	5.2	785
MobileViTV2 1.0 [49]	Arxiv 2022	256 ²	300	Att.	2,977	$\times 0.36$	0.336	78.1	4.9	1844
MobileOne-S3 [71]	CVPR 2023	224 ²	300	Conv	4,181	$\times 0.51$	0.239	78.1	10.1	1896
EfficientMod-XS [47]	ICLR 2024	224 ²	300	Att.	5,321	$\times 0.65$	0.188	78.3	6.6	778
EMO-6M [87]	ICCV 2023	224 ²	300	Att.	3,266	$\times 0.40$	0.306	79.0	6.1	961
FastViT-T12 [70]	ICCV 2023	256 ²	300	Att.	2,741	$\times 0.34$	0.365	79.1	6.8	1419
MobileFormer-508M [3]	CVPR 2022	224 ²	450	Att.	4,586	$\times 0.56$	0.218	79.3	14.0	508
MobileOne-S4 [71]	CVPR 2023	224 ²	300	Conv	3,041	$\times 0.37$	0.329	79.4	14.8	2978
SHViT-S4 [84]	CVPR 2024	256 ²	300	Att.	8,024	$\times 0.98$	0.124	79.4	16.5	986
EfficientViM-M4	-	256 ²	300	SSD	8,170	$\times 1.00$	0.122	79.4	19.6	1111
MobileViTV2 1.25 [49]	Arxiv 2022	256 ²	300	Att.	2,409	$\times 0.24$	0.415	79.6	7.5	2857
EfficientViM-M4	-	256 ²	450	SSD	8,170	$\times 1.00$	0.122	79.6	19.6	1111

Table 3. **Comparison of efficient networks on ImageNet-1K [10] classification.** Results are sorted by accuracy. We also denote the relative throughput Thr_{rel} of each method compared to EfficientViM in each split.

Method	Thr.	Thr _{rel}	Top-1	Params	FLOPs
EfficientViM-M2	17,005	$\times 2.08$	75.8	13.9M	355M
ViM-T [93]	1,612	$\times 0.20$	76.1	7.1M	1500M
LocalViM-T [93]	593	$\times 0.07$	76.2	8.0M	1500M
EfficientVMamba-T [47]	2,763	$\times 0.34$	76.5	6.0M	800M
MSVMamba-N [61]	2,060	$\times 0.25$	77.3	6.9M	864M
EfficientVMamba-S [47]	1,350	$\times 0.17$	78.7	11.0M	1300M
EfficientViM-M4	8,170	$\times 1.00$	79.6	19.6M	1111M
MSVMamba-M [61]	1,527	$\times 0.19$	79.8	11.9M	1507M

Table 4. **Comparison of EfficientViM with vision Mambas.** Thr_{rel} is relative throughput compared to EfficientViM-M4.

Comparison with vision Mambas. In Table 4, we compare our EfficientViM with the recent vision Mambas including ViM [93], LocalViM [26], EfficientVMamba [53], and MSVMamba [61]. EfficientViM brings promising speed improvements over the prior works. EfficientViM-M2 is almost 10 \times and 29 \times faster than ViM-T and LocalViM-T with comparable accuracy. EfficientViM-M4 shows about 3~14 \times higher throughput than other methods even with competitive performances. Notably, it outperforms MSVMamba-N and EfficientVMamba-S by 3.1% and 2.3%, achieving about 4 \times and 6 \times higher throughput, respectively. This reveals that EfficientViM is a highly efficient architecture among mamba-based vision models.

Method	Size	Thr.	Thr _{rel}	Top-1	Params	FLOPs
SHViT-S1 [84]	224 ²	19,868	$\times 0.96$	74.0	6.3M	241M
EfficientViM-M1	224 ²	20,731	$\times 1.00$	74.6	6.7M	239M
EfficientFormerV2-S0 [36]	224 ²	1,350	$\times 0.08$	75.7	3.5M	407M
SwiftFormer-XS [58]	224 ²	6,102	$\times 0.36$	75.7	3.5M	605M
SHViT-S2 [84]	224 ²	15,672	$\times 0.94$	76.2	11.4M	366M
FastViT-T8 [70]	256 ²	4,365	$\times 0.26$	76.7	3.6M	705M
EfficientViM-M2	224 ²	17,005	$\times 1.00$	76.7	13.9M	355M
EfficientMod-XXS [47]	224 ²	7,022	$\times 0.59$	77.1	4.7M	583M
SHViT-S3 [84]	224 ²	11,873	$\times 0.99$	78.3	14.2M	601M
SwiftFormer-S [58]	224 ²	4,675	$\times 0.39$	78.5	6.1M	988M
EfficientFormerV2-S1 [36]	224 ²	1,248	$\times 0.10$	79.0	6.1M	668M
EfficientViM-M3	224 ²	11,952	$\times 1.00$	79.1	16.6M	656M
EfficientMod-XS [47]	224 ²	5,321	$\times 0.65$	79.4	6.6M	778M
SHViT-S4 [84]	256 ²	8,024	$\times 0.98$	80.2	16.5M	986M
FastViT-T12 [70]	256 ²	2,741	$\times 0.34$	80.3	6.8M	1419M
EfficientViM-M4	256 ²	8,170	$\times 1.00$	80.7	19.6M	1111M

Table 5. Comparison of efficient networks after training with distillation objective in [67]. Thr_{rel} is relative throughput compared to EfficientViM in each split.

Head: Mask R-CNN [20]							
Method	Lat. (ms)	AP ^b	AP ₅₀ ^b	AP ₇₅ ^b	AP ^m	AP ₅₀ ^m	AP ₇₅ ^m
EfficientNet-B0 [64]	0.95	31.9	51.0	34.5	29.4	47.9	31.2
PoolFormer-S1 [83]	1.49	37.3	59.0	40.1	34.6	55.8	36.9
FastViT-SA12 [70]	1.63	38.9	60.5	42.2	35.9	57.6	38.1
SHViT-S4 [84]	0.52	39.0	61.2	41.9	35.9	57.9	37.9
EfficientViM-M4	0.45	39.3	60.2	42.5	35.8	57.1	37.4
Head: RetinaNet [56]							
Method	Lat. (ms)	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _i
PVTv2-B0 [75]	0.87	37.2	57.2	39.5	23.1	40.4	49.7
MobileFormer-508M [3]	1.09	38.0	58.3	40.3	22.9	41.2	49.7
EdgeViT-XXS [52]	0.94	38.7	59.0	41.0	22.4	42.0	51.6
SHViT-S4 [84]	0.52	38.8	59.8	41.1	22.0	42.4	52.7
EfficientViM-M4	0.45	38.8	59.6	41.1	22.1	42.4	52.8

Table 6. Instance segmentation and object detection and results on COCO-2017 [37].

Training with distillation. We compare EfficientViM with the prior works [36, 58, 70, 84] trained with distillation objectives in DeiT [67]. We train the model for 300 epochs, using RegNetY-160 [54] as the teacher model. Table 5 shows that distillation is effective for EfficientViM. Compared to FastViT-T8&T12 [70], EfficientViM-M2&M4 delivers more than 3× higher throughput along with comparable or even better performance. Further, EfficientViM outperforms SHViT [84] up to 0.8% running at a higher speed. With distillation, EfficientViM still outperforms all other models in speed-accuracy trade-offs and further establishes a promising Pareto front as shown in Figure 1.

4.2. Dense Predictions

Object detection and instance segmentation. We validate the effectiveness of EfficientViM on object detection and instance segmentation using the COCO-2017 [37] dataset. For training the models, we follow the settings of previous

Metric	PVTv2-B0	FastViT-SA12	EdgeViT-XXS	EfficientViM-M4
Latency (ms)	0.87	1.63	0.94	0.45
mIoU(%)	37.2	38.0	39.7	41.3

Table 7. Semantic segmentation with SemanticFPN [30] on ADE20K [91]

Method	Memory	Thr.	Thr _{rel}	Top-1	Params
EfficientViT-M4 [40]	870M	15,807	$\times 0.93$	74.3	8.8M
EMO-2M [87]	2656M	4,990	$\times 0.29$	75.1	2.3M
MobileNetV3-L.1.0 [22]	2643M	9,493	$\times 0.56$	75.2	5.4M
Mobile-Former-151M [3]	1800M	8,890	$\times 0.52$	75.2	7.6M
SHViT-S2 [84]	879M	15,899	$\times 0.94$	75.2	11.4M
FastViT-T8 [84]	2811M	4,365	$\times 0.26$	75.6	3.6M
EfficientViM-M2	969M	17,005	$\times 1.00$	75.8	13.9M

Table 8. Comparsion on peak memory usage during inference. Thr_{rel} is relative throughput compared to EfficientViM-M2.

works [40, 52, 70, 84], where Mask R-CNN [20] and RetinaNet [56] are used for instance segmentation and object detection. After training the model for 12 epochs (1x schedule) with a batch size of 16, we report the performances and backbone latencies with the resolution of 512², following [84]. Table 6 demonstrates that EfficientViM achieves competitive performances while maintaining a faster speed in dense prediction tasks. Specifically, in instance segmentation with Mask R-CNN, EfficientViM-M4 surpasses SHViT-S4 with the 0.3% improvements in AP^b while reducing latency by 0.7ms. Similarly, in object detection with RetinaNet, EfficientViM-M4 achieves the best average precision of 38.8% with the lowest latency.

Semantic segmentation. We also demonstrate the extensibility of EfficientViM on semantic segmentation using ADE20K [91] benchmark. Following previous works [52, 70], we replace the backbone of SemanticFPN [30] with EfficientViM-M4. The model is finetuned for 40K iterations with the batch size of 32 and the initial learning rate of 2×10^{-4} decayed using a polynomial scheduler with a power of 0.9. We report the mIoU and latency of the backbone with 512² resolutions in Table 7. We observe that EfficientViM-M4 largely surpasses all previous efficient backbones in semantic segmentation, which demonstrates the efficacy of EfficientViM in dense predictions along with the results of object detection and instance segmentation. It is worth noting that EfficientViM-M4 brings both 1.6% mIoU gains and about 2× speedup compared to the second-best model EdgeViT-XXS.

4.3. Analysis and Ablation studies

Memory Efficiency. Our models have a relatively large number of parameters compared to prior works. Yet, the memory usage of the model in the device is determined by the memory I/O during inference rather than just the number of parameters alone. We here analyze the peak memory usage of the networks and provide the results in Table 8.

Method	Thr.	Top-1	Params	FLOPs
(A) EfficientViM-M2 (Base)	17,005	75.4	13.9M	355M
(B) EfficientViM-M3	11,952	77.5	16.6M	656M
<i>a. Token Mixers (Base: HSM-SSD)</i>				
(C) (\rightarrow) NC-SSD [60]	9,786	76.2	13.0M	382M
(D) (\rightarrow) Self-Attention [72]	13,038	76.1	13.6M	416M
<i>b. Head designs (Base: Single-head with $\mathbf{A} \in \mathbb{R}^{L \times N}$)</i>				
(E) (\rightarrow) Multi-head	15,703	75.4	13.9M	352M
<i>c. Multi-stage fusion</i>				
(F) (\rightarrow) None	17,317	75.1	13.0M	354M

Table 9. **Ablation studies on EfficientViM.** All ablation studies were conducted with EfficientViM-M2 denoted as (Base).

Despite the highest parameter counts, EfficientViM shows a competitive memory efficiency while achieving the best throughput. Notably, we observe that EfficientViM requires only about 1/3 of the peak memory usage of the models having lower parameters, *e.g.*, EMO [87], MobileNetV3 [22], and FastViT [70]. Furthermore, the models with a relatively large number of parameters (*e.g.*, EfficientViT, SHViT, and EfficientViM) demonstrate high throughput and low memory consumption, highlighting that the number of parameters is not a critical factor for memory and time efficiency. Overall, our EfficientViM achieves the best speed and performance maintaining memory efficiency.

Ablation studies. We conduct ablation studies with EfficientViM-M2 after training 300 epochs and provide the results in Table 9. First, we compare HSM-SSD with other global token mixers, by replacing them with other methods including NC-SSD [60] and self-attention (SA) [72]. Considering that EfficientViM-M3 with HSM-SSD shows 77.5% with a throughput of 11,952 (im/s), NC-SSD and SA show a poorer speed-accuracy trade-off than HSM-SSD. Regarding head choices, we observe that our single-head design brings significant speed-up (15703 \rightarrow 17005 (im/s)) without performance degradation. Lastly, multi-stage fusion with hidden states (75.4%) surpasses the accuracy of EfficientViM without fusion (75.1%) under similar throughput. Refer to the supplement for more ablation studies.

5. Related Works

Efficient vision backbones. Earlier works [5, 22, 23, 27, 46, 57, 64, 71, 88] have studied improving the trade-off between accuracy and computational efficiency in CNN architectures. To reduce the complexity of convolution, Xception [5] introduced depthwise convolutions (DWConv), which have become major techniques used across modern efficient models. For instance, the pioneering work MobileNet [23] constructed lightweight architectures based on DWConv. ShuffleNet [88] and GhostNet [18] also have explored additional techniques to enhance CNN by shuffling

the channel and generating more feature maps.

Following ViTs [12], several works [3, 40, 48, 49, 52, 70, 84, 87] built efficient vision backbones with attention [72]. Some works sparsified the attention [7, 11, 42] to reduce query and key, while another line of works [6, 31, 74, 78] have approximated attention itself with reduced cost. More recently, EfficientFormer [35], EfficientViT [40], and SHViT [84] have proposed hardware-frienly ViT architectures for real-world applications, focusing on actual speed in practical use, rather than FLOPs.

Vision Mambas. State space model (SSM) [9, 14–16, 62] has become a popular global token mixer with its favorable linear cost. Especially, Mamba [15] has introduced a selective scan mechanism on SSM (S6) to enable time-variant selection. Following Mambas [9, 15], several works [26, 41, 47, 59, 66, 81, 93] have proposed SSM-based Vision backbones. ViM [93] applied bi-directional SSM to flattened patches considering the causal nature in Mamba. Similarly, VMamba [41], PlainMamba [81], and LocalMamba [26] have been proposed with the multi-path scanning mechanism. GroupMamba [59], and EfficientVMamba [53] further enhance the efficiency by splitting the channels and sharing the learnable parameters for each path. Recent works, VSSD [60] and Linfusion [39] introduced non-causal SSD to overcome causal properties in Mamba.

6. Conclusion

We propose a novel mamba-based architecture, Efficient Vision Mamba (EfficientViM), built with a hidden state mixer-based SSD (HSM-SSD). With the observation that the primary bottleneck of SSD stems from the linear projections with input tokens, we rearrange the channel mixing operation within the hidden states which serve as reduced latent representations. We also introduce multi-stage hidden state fusion that integrates both low-level and high-level features. To enhance practical runtime efficiency, we adopt a single-head design for HSM-SSD to minimize the memory-bound operations, which are typically overlooked when only considering FLOPs. Our comprehensive experiments demonstrate that EfficientViM significantly improves efficiency and performance over prior works across various tasks.

Acknowledgments. This work was supported by Korea University - KT (Korea Telecom) R&D Center, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2023R1A2C2005373), the Virtual Engineering Platform Project (Grant No. P0022336), funded by the Ministry of Trade, Industry & Energy (MoTIE, South Korea), and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00457882, AI Research Hub Project).

EfficientViM: Efficient Vision Mamba with Hidden State Mixer based State Space Duality

Supplementary Material

Configuration	Base	Distillation	FT
Epochs	300/450	300	30 + 30
Batch size	2048		1024
Weight decay	0.05		1e-8
Warmup Epochs	20		0
Cooldown Epochs	10		0
Learning rate	2e-3		1e-3
Min Learning rate	2e-5		1e-5
Optimizer (Momentum)	Adamw (0.9, 0.999)		
Gradient Clipping	0.02		
Learning rate scheduler	Cosine		
Rand Augment	rand-m9-mstd0.5-inc1		
Mixup	0.8		
Cutmix	1.0		
Mixup switch prob	0.5		
Random erasing prob	0.25		
Label smoothing	0.1		
EMA decay rate	0.9995		
Teacher model	None	RegNetY-160	None

Table A. **Settings for training EfficientViM.** FT: finetuning with higher resolution images (Section B).

A. Implementation Details

We use the ImageNet-1K [10] to validate the effectiveness of EfficientViM on the image classification task. For training EfficientViM, we follow the training recipes of previous works [60, 69, 84]. Specifically, all models are trained from scratch with a batch size of 2,048 for 300 epochs using AdamW optimizer [43] with a warmup of 20 epochs and a cooldown of 10 epochs. Following [3, 22, 36], we also report the results after training 450 epochs. During training, we adopt a cosine annealing [44] scheme with the initial learning rate of 2×10^{-3} decreasing to 2×10^{-5} . The weight decay of 0.05 and gradient clipping with a threshold of 0.02 are used. Also, MESA [13] and EMA with the decay rate of 0.9995 is adopted following [17, 60]. For data augmentation, we follow DeiT [67] using Mixup [86] & CutMix [85] with a Label smoothing [63], RandAugment [8], and Random Erasing [90]. We report the throughput and latency with the batch size of 256 on Nvidia RTX 3090 GPU.

Additionally, we finetune the model with the batch size of 1024, using cosine annealing with the initial learning rate of 1×10^{-3} , for 30 epochs at a resolution of 384^2 , followed by an additional 30 epochs at 512^2 . For a fair comparison, we employ pre-trained models trained for 300 epochs. Also, to report the throughput of the models with extremely

Method	Size	Thr.	Thr _{rel}	Top-1	Params	FLOPs
SHViT-S4 [84]	384^2	3,685	$\times 0.99$	81.0	16.5M	2225M
EfficientViM-M4	384^2	3,724	$\times 1.00$	80.9	21.3M	2379M
SHViT-S4 [84]	512^2	2,122	$\times 0.86$	82.0	16.5M	3973M
EfficientViM-M4	512^2	2,452	$\times 1.00$	81.9	21.3M	4154M

Table B. **Classification results on ImageNet-1K [10] after fine-tuning with higher resolutions.** Thr_{rel} is relative throughput compared to EfficientViM-M4.

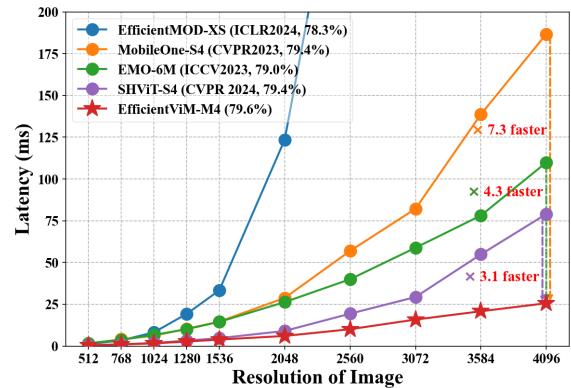


Figure A. **Latency comparison of recent efficient networks for an extremely high-resolution image.**

high-resolution images in Figure A, we start with a batch size of 256 and halve it once the memory exceeds the GPU limit as the resolution increases. Regarding training with distillation, all settings are the same as in Table 3. of the main paper except for the guidance from the teacher model of RegNetY-160 [54] following DeiT [67].

B. EfficientViM with high-resolution images.

Following [84], we also explore the applicability of EfficientViM on high-resolution images after fine-tuning 30 epochs at a resolution of 384^2 , followed by an additional 30 epochs at 512^2 . For a fair comparison, we use the EfficientViM-M4 pre-trained for 300 epochs. In 384^2 size, EfficientViM demonstrates competitive performance as presented in Table B. Interestingly, when the resolution increases, the throughput gap between EfficientViM and SHViT [84] gets larger, resulting in more than 15% speedup compared to SHViT in 512^2 while achieving comparable accuracy. We further investigate the scalability of our method in extremely high-resolution images beyond 512^2 , by comparing the latency of the models while scaling the resolution from 512^2 to 4096^2 . As depicted in Figure A, we observe an advantage of EfficientViM over the

Method	Thr.	Top-1	Params	FLOPs
(A) EfficientViM-M2 (Base)	17,005	75.4	13.9M	355M
(B) EfficientViM-M3	11,952	77.5	16.6M	656M
<i>a. Token Mixers (Base: HSM-SSD)</i>				
(C) (\rightarrow) NC-SSD [60]	9,786	76.2	13.0M	382M
(D) (\rightarrow) Self-Attention [72]	13,038	76.1	13.6M	416M
<i>b. Head designs (Base: Single-head (SH) with $\mathbf{A} \in \mathbb{R}^{L \times N}$)</i>				
(E) (\rightarrow) Multi-head	15,703	75.5	13.9M	352M
(F) (\rightarrow) SH w. $\mathbf{a} \in \mathbb{R}^L$	17,081	75.2	13.9M	352M
<i>c. Multi-stage fusion (Base: $\mathbf{h}^{(s)}$)</i>				
(G) (\rightarrow) Fusion with $\mathbf{x}^{(s)}$	17,041	75.3	13.4M	355M
(H) (\rightarrow) None	17,317	75.1	13.0M	354M
<i>d. # states (N) of each stage (Base: [49, 25, 9])</i>				
(I) (\rightarrow)[9, 25, 49]	16,476	75.4	14.0M	407M
(J) (\rightarrow)[25, 25, 25]	16,991	75.2	13.9M	373M
<i>e. Normalization (Base: Partial LN)</i>				
(L) (\rightarrow) Full BN	17,432	NaN	13.9M	355M

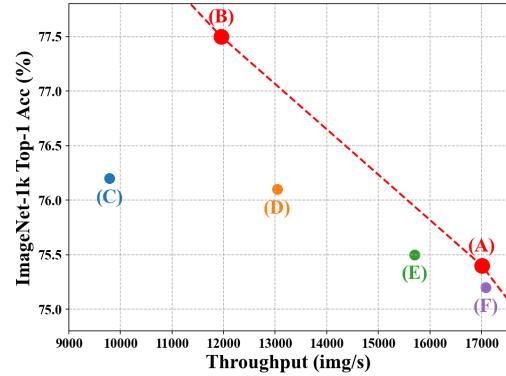
Table C. **Ablation studies on EfficientViM.** All ablations are conducted with EfficientViM-M2. See Figure B for a visualization comparing the ablated models with the Pareto front of EfficientViM.

recent state-of-the-art method on extremely high-resolution images. EfficientViM shows about $3\times$, $4\times$, and $7\times$ faster speed compared to SHViT, EMO [87], and MobileOne [71], respectively. This notable result highlights the scalability of EfficientViM for high-resolution images based on linear cost of HSM-SSD.

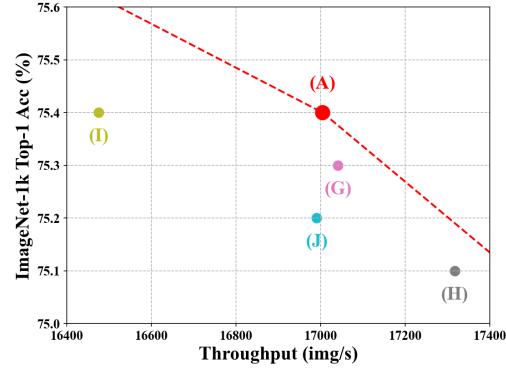
C. Ablation Studies

Here, we show the effectiveness of HSM-SSD by ablating the proposed components. The results are summarized in Table C, and Figure B. First, we compare HSM-SSD with other global token mixers, by replacing them with other methods including (C) NC-SSD [60] and (D) self-attention (SA) [72]. Considering that EfficientViM-M3 with HSM-SSD shows 77.5% with a throughput of 11,952 (im/s), NC-SSD and SA show a poorer speed-accuracy trade-off than HSM-SSD. Regarding head choices, we observe that our single-head design brings significant speed-up (17005 (im/s)) over (E) multi-head (15,703 (im/s)) without performance degradation. Additionally, defining the importance score per state as $\mathbf{A} \in \mathbb{R}^{L \times N}$ to mimic multi-head leads to the +0.2% gain with a minor increase in latency, compared to using the original score (F) $\mathbf{a} \in \mathbb{R}^L$. Note that all ablates models (C-F) are placed under the Pareto front of EfficientViM (Figure Ba), which proves the efficacy of HSM-SSD and our single-head design.

Also, multi-stage fusion with hidden states (75.4%) surpasses the accuracy of (G) the fusion with the output feature maps $\mathbf{x}^{(s)}$ (75.3%) and (H) the EfficientViM without fusion



(a) **Ablations on the token mixer and head design (C-F).**



(b) **Ablations on Multi-stage fusion and # states (G-J).**

Figure B. **Ablation studies on EfficientViM.** Refer to Table C for the corresponding models. Red line indicates the Pareto frontier of EfficientViM.

Method	Token Mixer	Thr.	Top-1	Params	FLOPs
VSSD-M [60]	NC-SSD	1459	82.5	14M	2.3G
VSSD-T [60]	NC-SSD	947	84.1	24M	4.5G
VSSD-T	\rightarrow HSM-SSD	1660	82.7	24M	3.7G

Table D. **Comparison of HSM-SSD with NC-SSD.**

(75.1%) under similar throughput. For the number of states N , we observe that an increasing schedule with respect to the stages is more effective than (I) a decreasing or (J) constant schedule. See (G)-(J) in Table C and Figure Bb for the ablation studies on multi-stage fusion and the number of states. Additionally, for normalization, using (L) batch normalization (BN) across all operations is simple and fast, but, this approach leads to numerical instability. Therefore, we apply layer normalization (LN) only before HSM-SSD, and BN for the rest.

D. Comparison of HSM-SSD with NC-SSD

To show the advantage of the proposed HSM-SSD over NC-SSD, we replace NC-SSD of VSSD-T [60] with HSM-SSD

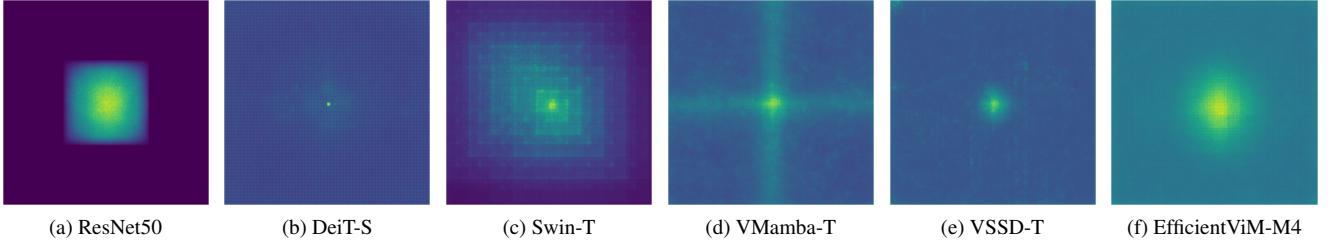


Figure C. Comparison of Effective Receptive Fields (ERF) [45]

and train the model following the original training configuration provided in the official repository. In Table D, after replacing the token mixer with HSM-SSD, VSSD-T with HSM-SSD demonstrates a significant increase in the throughput ($1.8\times$). Notably, compared to scaling down the models to smaller sizes (*e.g.*, VSSD-M), replacing the token mixer with HSM-SSD provides better speed-accuracy trade-offs (14% faster, +0.2% accuracy), highlighting the advantage of HSM-SSD over NC-SSD. We also provide the qualitative comparison of HSM-SSD with NC-SSD in the following section.

E. Effective Receptive Field of HSM-SSD

In this section, we qualitatively compare the HSM-SSD with previous token mixers including convolutions in ResNet50 [19], attentions in vision Transformers such as DeiT-S [67] and Swin-T [42], and SSM and SSD variants in vision Mambas like VMamba-T [41] and VSSD-T [60]. We here analyze the Effective Receptive Fields (ERF) [45] of each model, which quantifies the region of the input that contributes to the output. In Figure C, we visualize the ERF with respect to the central pixel of the output feature maps. Among various models, EfficientViM-M4 with HSM-SSD shows a global receptive field rather than focusing on a specific region. For instance, ResNet50 shows a relatively small ERF due to its intrinsic locality of convolution. The attention mechanism in DeiT-S predominantly focuses on the central pixel itself, and the shifted window attention in Swin-T limits the global receptive field. Further, since SSM is conducted after flattening the image patch both vertically and horizontally in VMamba-T, it generates an unnatural cross pattern in ERF. VSSD-T shows a relatively better global receptive field, yet it still largely depends on the close region. On the other hand, EfficientViM-M4 generates a global effective receptive field (ERF) similar to that of VSSD-T but extracts more information from all regions, enabling it to capture the global dependencies better.

F. CPU & Mobile Latency

To understand the applicability of EfficientViM in a resource-constrained environment, we here provide the la-

Method	Latency (ms)			Top-1
	GPU	CPU	Mobile	
MobileViTV2 1.0 [49]	0.34	138.8	1.1	78.1
EfficientMod-XS [47]	0.19	33.1	0.9	78.3 (79.4)
MobileFormer-508M [3]	0.22	29.7	2.1	79.3
FastViT-T12 [70]	0.37	81.3	1.8	79.1 (80.3)
MobileOne-S4 [71]	0.33	79.9	1.0	79.4
SHViT-S4 [84]	0.12	27.4	0.9	79.4 (80.2)
EfficientViM-M4	0.12	32.1	1.0	79.6 (80.7)

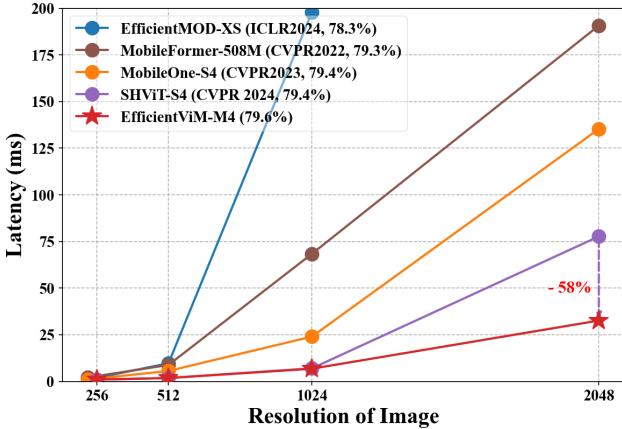
Table E. Latency comparison of EfficientViM-M4 with prior works. The number in parentheses indicates the performance with distillation.

tencies of vision backbones in GPU, CPU, and mobile devices (Table E). The latencies are measured with a batch size of 256 on an NVIDIA RTX 3090 GPU, 16 on an AMD EPYC 7742 CPU, and 1 on an iPhone 16 (iOS 18.1). For mobile latencies, we use CoreML [50] library. EfficientViM-M4 achieves the highest accuracy of 79.6% with the lowest latency on GPUs and competitive latency on CPU and iPhone 16. Although EfficientViM-M4 shows slightly higher latency than a few of the prior works in CPU and mobile, EfficientViM-M4 shows a significantly lower latency as the resolution increases (Figure D). In the resolution of 2048^2 , EfficientViM-M4 achieves at least 58% and 20% reductions in latency compared to previous works in iPhone 16 and CPU, respectively. To summarize, EfficientViM serves as a general solution suitable for both GPU and edge devices. Furthermore, EfficientViM is an effective backbone for real-world applications where high-resolution images are given, such as in image generation, object detection, and instance segmentation.

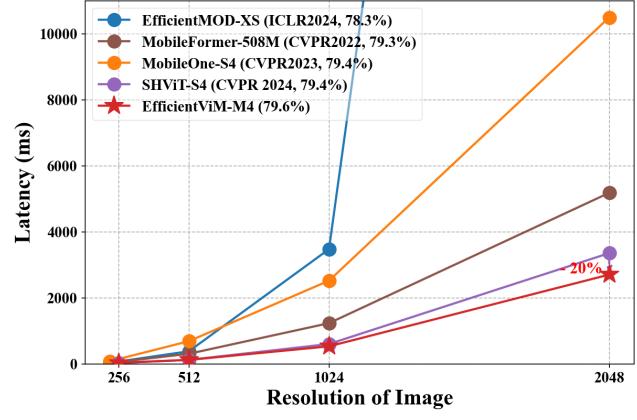
G. Proof for Proposition 1

Proposition 1. Let $N = L$, $\mathbf{a} \mathbb{1}_L^\top \odot \mathbf{B} = \mathbb{I}_L$, and $\mathbf{C} \in \mathbb{R}^{L \times L}$ be diagonal. Then, $HSM\text{-}SSD(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C})$ is equivalent to $NC\text{-}SSD(\mathbf{x}, \mathbf{a}, \mathbf{B}, \mathbf{C})$ including gating and output projection, as $\mathbf{x}_{out} = f(\mathbf{y}) = \mathbf{C}f(\mathbf{h})$.

Proof. It is sufficient to show that $\mathbf{C}f(\mathbf{h})$ of HSM-SSD is equivalent to $f(\mathbf{y}) = (\mathbf{Ch} \odot \sigma(\mathbf{x}_{in}\mathbf{W}_z))\mathbf{W}_{out}$ in order to prove the proposition. Here, based on the assumption, the



(a) Mobile latency comparison.



(b) CPU latency comparison.

Figure D. Mobile and CPU latency comparison.

following holds:

$$\begin{aligned}
 & \mathbf{C}f(\mathbf{h}) \\
 &= \mathbf{C}((\mathbf{h} \odot \sigma(\mathbf{h}_{in}\mathbf{W}_z))\mathbf{W}_{out}) \\
 &= \mathbf{C}((\mathbf{h} \odot \sigma(((\mathbf{a}\mathbf{1}_L^\top \odot \mathbf{B})^\top \mathbf{x}_{in})\mathbf{W}_z))\mathbf{W}_{out}) \\
 &= \mathbf{C}(\mathbf{h} \odot \sigma(\mathbf{x}_{in}\mathbf{W}_z))\mathbf{W}_{out} \\
 &= (\mathbf{Ch} \odot \sigma(\mathbf{x}_{in}\mathbf{W}_z))\mathbf{W}_{out} = f(\mathbf{y})
 \end{aligned}$$

□

Remarks. We assume that \mathbf{C} is a diagonal matrix but \mathbf{C} is dependent on \mathbf{x}_{in} since $\mathbf{C} = \mathbf{x}_{in}\mathbf{W}_C$. Unfortunately, there does not exist a weight matrix \mathbf{W}_C that makes \mathbf{C} diagonal for arbitrary inputs \mathbf{x}_{in} . We provide this proposition to understand the relationship between the HSM-SSD and NC-SSD operations. This implies that in specific conditions with particular data, the two methods yield the same result. However, one approach does not generalize the other.

H. Discussion of multi-stage hidden stage fusion

In this section, we briefly discuss how multi-stage hidden state fusion (MSF) provides a performance boost, although the earlier layers generally provide less accurate logits. Note that our MSF leverages the logits across the layer as *deep supervision* and *multi-scale representation* to improve the performance. MSF aligns with the concept of ‘deep supervision’ in pioneering works, such as DSN [32], and U-Net++ [92]. Specifically, during training, MSF can be interpreted as auxiliary classification tasks, encouraging even the earlier layers to learn more discriminative features. Further, the earlier layers generally capture fine-grained patterns, while later layers extract high-level semantics, i.e., DenseNet [25] and FPN [38]. By combining these complementary representations with the learnable coefficients,

we take advantage of the ensemble effect from ‘multi-scale representations’. In fact, it is well-known that an ensemble of the models often outperforms each model, highlighting the benefits of HSM. As a results, MSF brings substantial improvements on EfficientViM.

References

- [1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2022. 1
- [2] Yingpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic relu. In *ECCV*, 2020. 5
- [3] Yingpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobileformer: Bridging mobilenet and transformer. In *CVPR*, 2022. 1, 6, 7, 8, 9, 11
- [4] Joonmyung Choi, Sanghyeok Lee, Jaewon Chu, Minhyuk Choi, and Hyunwoo J Kim. vid-tldr: Training free token merging for light-weight video transformer. In *CVPR*, 2024. 1
- [5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 1, 8
- [6] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *ICLR*, 2021. 1, 8
- [7] 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. *NeurIPS*, 2021. 1, 8
- [8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 9
- [9] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured

- state space duality. In *International Conference on Machine Learning (ICML)*, 2024. 1, 2, 3, 4, 8
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 5, 6, 9
- [11] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 8
- [12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 8
- [13] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *NeurIPS*, 2022. 9
- [14] 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. *ICLR*, 2023. 1, 8
- [15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023. 2, 8
- [16] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *ICLR*, 2022. 1, 8
- [17] Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *arXiv:2405.16605*, 2024. 9
- [18] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *CVPR*, 2020. 1, 8
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 11
- [20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 7
- [21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415*, 2016. 5
- [22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019. 1, 2, 5, 6, 7, 8, 9
- [23] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017. 1, 8
- [24] Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, and Liujuan Cao. You only segment once: Towards real-time panoptic segmentation. In *CVPR*, 2023. 1
- [25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 12
- [26] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv:2403.09338*, 2024. 2, 6, 8
- [27] Forrest N Iandola. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv:1602.07360*, 2016. 1, 8
- [28] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, 2021. 4
- [29] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pages 5156–5165. PMLR, 2020. 3
- [30] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 7
- [31] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *ICLR*, 2020. 1, 8
- [32] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015. 12
- [33] Sanghyeok Lee, Joonmyung Choi, and Hyunwoo J Kim. Multi-criteria token fusion with one-step-ahead attention for efficient vision transformers. In *CVPR*, 2024. 1
- [34] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *CVPR*, 2017. 1
- [35] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *NeurIPS*, 2022. 8
- [36] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *ICCV*, 2023. 6, 7, 9
- [37] 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. 7
- [38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 12
- [39] Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. Linfusion: 1 gpu, 1 minute, 16k image. *arXiv:2409.02097*, 2024. 2, 8
- [40] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *CVPR*, 2023. 1, 5, 6, 7, 8
- [41] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *NeurIPS*, 2024. 2, 8, 11
- [42] 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 *ICCV*, 2021. 1, 8, 11
- [43] I Loshchilov. Decoupled weight decay regularization. *ICLR*, 2019. 9
- [44] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv:1608.03983*, 2016. 9
- [45] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *NeurIPS*, 2016. 11

- [46] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018. 1, 8
- [47] Xu Ma, Xiyang Dai, Jianwei Yang, Bin Xiao, Yinpeng Chen, Yun Fu, and Lu Yuan. Efficient modulation for vision networks. *ICLR*, 2024. 6, 7, 8, 11
- [48] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *ICLR*, 2022. 1, 8
- [49] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *arXiv:2206.02680*, 2022. 1, 5, 6, 8, 11
- [50] Core ML. <https://developer.apple.com/documentation/coreml>, 2017. 11
- [51] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5
- [52] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *ECCV*, 2022. 1, 6, 7, 8
- [53] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientmamba: Atrous selective scan for light weight visual mamba. *arXiv:2403.09977*, 2024. 6, 8
- [54] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *CVPR*, 2020. 7, 9
- [55] J Redmon. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1
- [56] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *CVPR*, 2017. 7
- [57] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1, 5, 8
- [58] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *ICCV*, 2023. 7
- [59] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. Groupmamba: Parameter-efficient and accurate group visual state space model. *arXiv:2407.13772*, 2024. 8
- [60] Yuheng Shi, Minjing Dong, Mingjia Li, and Chang Xu. Vssd: Vision mamba with non-casual state space duality. *arXiv:2407.18559*, 2024. 2, 3, 4, 8, 9, 10, 11
- [61] Yuheng Shi, Minjing Dong, and Chang Xu. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. *NeurIPS*, 2024. 2, 6
- [62] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *ICLR*, 2023. 1, 8
- [63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 9
- [64] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 1, 7, 8
- [65] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, 2020. 1
- [66] Lv Tang, HaoKe Xiao, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Scalable visual state space model with fractal scanning. *arXiv:2405.14480*, 2024. 8
- [67] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 7, 9, 11
- [68] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 5
- [69] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 5, 9
- [70] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *ICCV*, 2023. 1, 5, 6, 7, 8, 11
- [71] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Mobileone: An improved one millisecond mobile backbone. In *CVPR*, 2023. 1, 5, 6, 8, 10, 11
- [72] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 3, 4, 8, 10
- [73] Chengcheng Wang, Wei He, Ying Nie, Jianyuan Guo, Chuanjian Liu, Yunhe Wang, and Kai Han. Gold-yolo: Efficient object detector via gather-and-distribute mechanism. *NeurIPS*, 2024. 1
- [74] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv:2006.04768*, 2020. 1, 8
- [75] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvttv2: Improved baselines with pyramid vision transformer, 2022. 7
- [76] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023. 6
- [77] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 1
- [78] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *AAAI*, 2021. 1, 8
- [79] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pre-training for efficient segment anything. In *CVPR*, 2024. 1
- [80] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *CVPR*, 2023. 1
- [81] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley.

- Plainmamba: Improving non-hierarchical mamba in visual recognition. *BMVC*, 2024. 2, 8
- [82] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, 2022. 1
- [83] 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 *CVPR*, 2022. 7
- [84] Seokju Yun and Youngmin Ro. Shvit: Single-head vision transformer with memory efficient macro design. In *CVPR*, 2024. 1, 2, 4, 5, 6, 7, 8, 9, 11
- [85] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 9
- [86] Hongyi Zhang. mixup: Beyond empirical risk minimization. *ICLR*, 2018. 9
- [87] Jiangning Zhang, Xiangtai Li, Jian Li, Liang Liu, Zhucun Xue, Boshen Zhang, Zhengkai Jiang, Tianxin Huang, Yabiao Wang, and Chengjie Wang. Rethinking mobile block for efficient attention-based models. In *ICCV*, 2023. 1, 6, 7, 8, 10
- [88] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 1, 8
- [89] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018. 1
- [90] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 9
- [91] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 7
- [92] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI Workshop*, 2018. 12
- [93] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *ICML*, 2024. 2, 5, 6, 8