Efficient High-Resolution Visual Representation Learning with State Space Model for Human Pose Estimation

Hao Zhang, Yongqiang Ma, Wenqi Shao, Ping Luo, Nanning Zheng, *Fellow, IEEE,* Kaipeng Zhang

arXiv:2410.03174v2 [cs.CV] 15 Aug 2025

***Abstract*—Capturing long-range dependencies while preserving high-resolution visual representations is crucial for dense prediction tasks such as human pose estimation. Vision Transformers (ViTs) have advanced global modeling through self-attention but suffer from quadratic computational complexity with respect to token count, limiting their efficiency and scalability to high-resolution inputs, especially on mobile and resource-constrained devices. State Space Models (SSMs), exemplified by Mamba, offer an efficient alternative by combining global receptive fields with linear computational complexity, enabling scalable and resource-friendly sequence modeling. However, when applied to dense prediction tasks, existing visual SSMs face key limitations: weak spatial inductive bias, long-range forgetting from hidden state decay, and low-resolution outputs that hinder fine-grained localization. To address these issues, we propose the *Dynamic Visual State Space* (DVSS) block, which augments visual state space models with multi-scale convolutional operations to enhance local spatial representations and strengthen spatial inductive biases. Through architectural exploration and theoretical analysis, we incorporate deformable operation into the DVSS block, identifying it as an efficient and effective mechanism to enhance semantic aggregation and mitigate long-range forgetting via input-dependent, adaptive spatial sampling. We embed DVSS into a multi-branch high-resolution architecture to build HRVMamba, a novel model for efficient high-resolution representation learning. Extensive experiments on human pose estimation, image classification, and semantic segmentation show that HRVMamba performs competitively against leading CNN-, ViT-, and SSM-based baselines. The code is available at https://github.com/zhanghao5201/PoseVMamba.**

***Index Terms*—High-Resolution Representation Learning, Mamba, Human Pose Estimation, Image Classification, Semantic Segmentation**

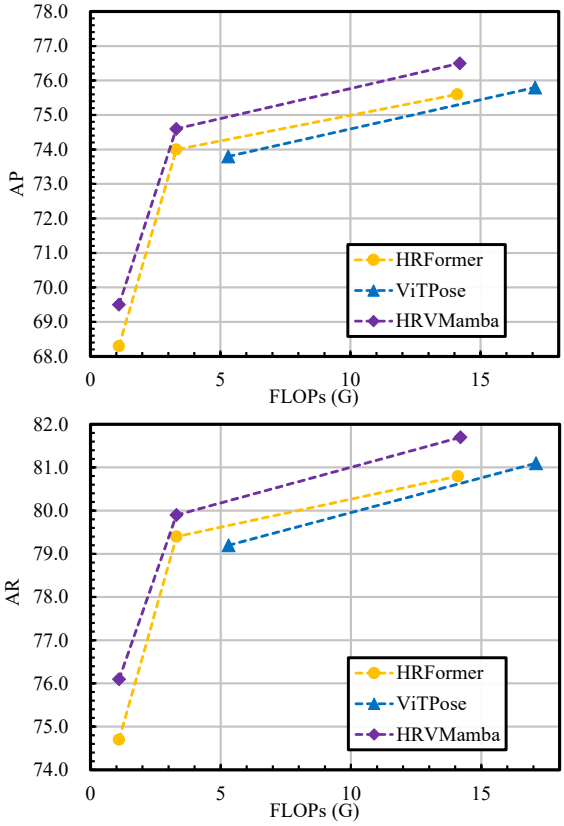


Fig. 1: **The trade-off between AP, AR and FLOPs on COCO pose estimation `val` set for human pose estimation.**

I. Introduction

Learning robust high-resolution representations is a fundamental yet challenging requirement for dense prediction tasks such as human pose estimation [1]–[3]. Conventional visual backbones like Convolutional Neural Networks (CNNs) [4]–[8] are effective at capturing local patterns, thanks to their strong spatial inductive biases and linear computational complexity. However, their limited receptive fields restrict their ability to model long-range dependencies, which are critical for complex visual understanding in dense prediction tasks. On the other hand, as illustrated in Fig. 1, Vision Transformers (ViTs) [9]–[12] exploit global self-attention to model long-range context. However, their quadratic complexity with respect to token count makes them computationally expensive, which in turn limits their performance, especially in high-resolution scenarios and when deployed on resource-constrained platforms such as mobile devices.

State Space Models (SSMs) have recently emerged as efficient alternatives to ViTs, offering linear computational complexity with respect to token length while modeling long-range global dependencies. By combining linear computational complexity with global receptive capabilities, SSMs enable scalable and computation-friendly modeling of long

Corresponding authors: Kaipeng Zhang; Nanning Zheng.

Hao Zhang, Yongqiang Ma, and Nanning Zheng are with National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: zhanghao520@stu.xjtu.edu.cn, musayq@xjtu.edu.cn, nnzheng@mail.xjtu.edu.cn).

Wenqi Shao, Ping Luo, Kaipeng Zhang are with Shanghai Artificial Intelligence Laboratory, Shanghai, 200000, China (e-mail: shaowenqi@pjlab.orn.cn, pluo@cs.hku.edu, zhangkaipeng@pjlab.org.cn).

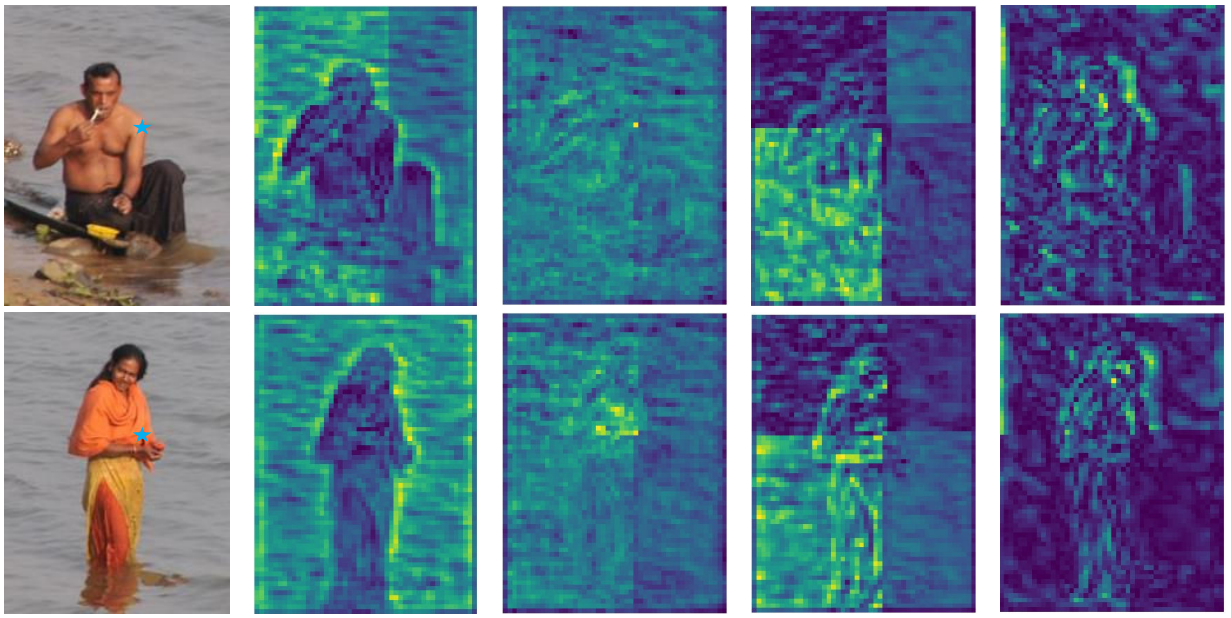


Fig. 2: **Activation maps of SSM for the query location** (marked by a blue pentagram). We embed the VSS block [14] and VSS+DSS2D (where the SS2D block is replaced by our proposed Deformable 2D-Selective-Scan (DSS2D) block) into HRVMamba-Small. Here, S_i denotes the i-th stage of HRVMamba. We visualize the SSM activation maps from the second block of the first block in the first branch of each S_i. In the early stage (S2), the DSS2D block attends to semantically meaningful regions that are relevant to the query patch, while the SS2D block primarily focuses on low-level edge features. In the later stage (S3), the DSS2D block effectively emphasizes human-related regions, whereas the SS2D block still responds to less informative background areas.

sequences. This advancement has led to a surge of visual SSM architectures, such as ViM [13], VMamba [14], LocalVMamba [15], VideoMamba [16], GroupMamba [17], and MambaVision [18].

Nevertheless, despite their improved efficiency, existing SSM-based visual models face several key limitations in dense prediction tasks such as human pose estimation. First, their tokenized sequence representation and bi-/multi-directional scanning mechanisms [13], [15] disrupt the spatial continuity of images, resulting in insufficient spatial inductive bias required for capturing local image details. Secondly, the recurrent nature of token processing causes gradual information decay in hidden states, leading to the phenomenon of long-range forgetting. Consequently, these models may lose crucial high-level semantic information relevant to distant tokens, defaulting instead to low-level features such as edges. This effect is visually demonstrated in Fig. 2, column 2[1]. Lastly, existing visual SSMs often produce low-resolution, single-scale feature maps, which inadequately represent fine-grained spatial details and multi-scale variability necessary for precise pose estimation.

To address these issues, we propose the *Dynamic Visual State Space* (DVSS) block, an improved variant of the Visual State Space (VSS) block in VMamba. DVSS incorporates convolutional kernels at multiple scales to robustly capture local spatial features and strengthen the model's inductive bias across different resolutions. To further alleviate long-range forgetting and enhance spatial dependency modeling, we systematically explore a range of architectural design alternatives. Through empirical comparisons and theoretical insights, we identify deformable operation as a particularly effective mechanism. It dynamically adjusts spatial aggregation based on the input and task-specific context, enabling adaptive sampling and enhancement of semantically relevant yet spatially distant regions. Our analysis demonstrates how this mechanism mitigates the decay of contextual information over distance and allows the model to focus selectively on meaningful spatial cues, thereby enhancing both representation robustness and performance. For example, as shown in Fig. 2, the left shoulder and chest features near the right shoulder are highlighted (row 1, column 3), the head features connected to the right shoulder are emphasized (row 1, column 5), along with the highlighted features of both hands and the chest (row 2, column 3).

Building on the multi-resolution parallel framework, we embed DVSS blocks into multi-branch architectures to form the High-Resolution Visual State Space Model named **HRVMamba**. This model effectively preserves high-resolution information and models multi-scale feature variations, thus excelling in dense visual tasks. HRVMamba achieves a better trade-off between computational complexity and accuracy, making it suitable for scenarios with limited computational budgets (e.g., under similar FLOP constraints), while still maintaining detailed spatial modeling.

[1]We visualize attention maps using VMamba's approach [14].

The contributions of this study are as follows:

- We propose **HRVMamba**, a high-resolution visual state space model that efficiently supports high-resolution representation learning. It adopts a multi-resolution branch architecture to preserve fine-grained details and capture multi-scale variations for human pose estimation.
- We introduce the DVSS block, which combines multi-scale convolutional kernels and deformable operations to enhance inductive bias and mitigate the long-range forgetting problem.
- HRVMamba demonstrates promising performance in human pose estimation, image classification, and semantic segmentation tasks. Experimental results show that HRVMamba achieves competitive results against existing CNN, ViT, and SSM benchmark models.

The remainder of this paper is organized as follows: Section II presents a concise overview of related work in CNNs, ViTs, SSMs, and high-resolution representation learning. Section III introduces the theoretical foundations of SSMs and the Selective State Space Model. Section IV details the proposed HRVMamba architecture, including its core component, i.e., the DVSS Block. Section V reports extensive experimental results and ablation studies across multiple benchmarks. Finally, Section VI concludes the paper.

II. Related work

In this section, we first review related work on Convolutional Neural Networks and Vision Transformers. Next, we discuss recent advances in State Space Models. Finally, we introduce works specifically focused on high-resolution representation learning.

A. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)

CNNs have long been the cornerstone of computer vision, evolving from early models like AlexNet [19] and ResNet [4] to more advanced architectures such as ConvNeXt [20], SCGNet [5], FlashInternImage [19], and FMGNet [7]. These models excel at local feature extraction, leveraging strong inductive biases and efficient computation, and have achieved remarkable performance across a wide range of tasks including image classification, semantic segmentation, and human pose estimation. ViTs introduce self-attention mechanisms from natural language processing, segmenting images into non-overlapping patches to effectively model long-range global dependencies. This paradigm shift forms the backbone of modern Large Vision-Language Models [21]–[23]. Numerous enhancements, such as the distillation strategies in DeiT [24], the hierarchical design in Swin Transformer [10], and the lightweight attention mechanisms in SwiftFormer [25], have been proposed to improve both performance and efficiency, promoting broader adoption of ViTs in various vision tasks. Meanwhile, hybrid architectures [26], [27] that combine the complementary strengths of CNNs and ViTs have gained increasing attention. These models leverage the local feature extraction efficiency and inductive bias of CNNs while incorporating the global context modeling capabilities of ViTs, marking a promising and significant direction in the development of modern backbone networks.

B. State Space Models (SSMs)

SSMs provide a mathematical framework for modeling dynamic systems and exhibit linear computational complexity with respect to sequence length, making them highly efficient for processing long sequential data. Advances in models such as S4 [28], S5 [29], and H3 [30] have significantly improved SSMs by incorporating structure-aware optimizations, parallel scan mechanisms, and hardware-friendly designs that enhance both accuracy and speed. Mamba [31] further advances the field by introducing input-dependent parameterization and a hardware-efficient parallel scanning mechanism known as S6, firmly establishing SSMs as a compelling and scalable alternative to Transformer-based architectures. Since Mamba's introduction, SSMs have been increasingly adopted in visual domains, with early efforts such as S4ND [32] treating images as continuous 2D signals and demonstrating the viability of SSMs for vision tasks. Building upon Mamba's success, several vision-oriented variants have been proposed. ViM [13] and VMamba [14] mitigate the directional bias of unidirectional scanning by incorporating bidirectional or four-way scan strategies. LocalVMamba [15] introduces localized window-based scanning to better preserve fine-grained spatial details, while PlainMamba [33] reworks 2D scanning for sequential data processing in a simpler form. GroupMamba [17] enhances training stability and convergence through a distillation-based optimization framework, and MambaVision [18] combines the strengths of SSMs and Transformers for hybrid feature modeling. Despite these innovations, most existing visual Mamba-based models still suffer from long-range forgetting and rely primarily on single-scale, low-resolution feature representations. This significantly limits their ability to preserve fine-grained visual cues and to capture the multi-scale variations that are crucial for dense prediction tasks such as human pose estimation.

C. High-Resolution Representation Learning

Learning effective high-resolution representations is essential for dense prediction tasks such as human pose estimation [1]–[3]. The High-Resolution Network (HRNet) [34] is the first to propose maintaining high-resolution representations throughout the entire network, demonstrating impressive performance in dense prediction tasks such as human pose estimation and semantic segmentation. HRNet employs a multi-resolution parallel architecture, where information from different scales is continuously fused through repeated multi-scale feature exchange, allowing it to effectively capture both fine-grained details and global context. Building on this foundation, HRFormer [9] incorporates the self-attention mechanism [10] into the high-resolution architecture, combining the benefits of Transformers with HRNet's strong structural inductive biases. This design further enhances performance across various dense prediction tasks by enabling better long-range context modeling. Subsequent efforts have focused on improving the efficiency and deployability of high-resolution

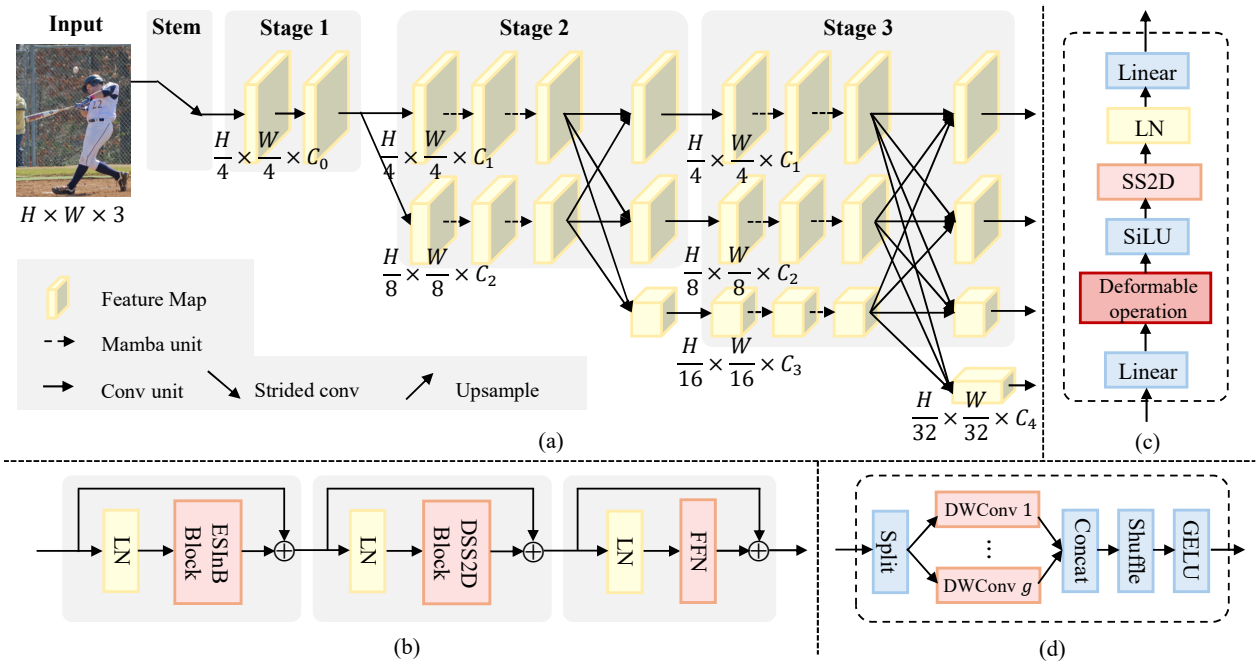


Fig. 3: **(a) Overall architecture of HRVMamba. (b) Dynamic Visual State Space block. (c) Deformable 2D-Selective-Scan (DSS2D) Block. (d) Enhanced Spatial Inductive Bias Block (ESInB) Block.** HRVMamba has four stages, but for demonstration purposes, we only show three. H and W represent the height and width of the image, while C_i denotes the number of channels in the i-th branch or position. LN, Linear, DWConv and SS2D represent LayerNorm, Linear Layer, depthwise convolution, and 2D-Selective-Scan SSM.

architectures. Lite-HRNet [35], Dite-HRNet [36], and HF-HRNet [6] introduce lightweight convolutional backbones using techniques such as depthwise convolution and dynamic convolution. Despite these advances, it remains unclear whether Mamba can fully exploit high-resolution structures for efficient high-resolution representation learning. Key challenges include addressing Mamba's inherent lack of spatial inductive bias and its tendency toward long-range forgetting, particularly in complex visual scenes. Effectively integrating Mamba into multi-resolution frameworks while preserving both fine-grained detail and high-level semantics remains an open and promising research direction.

III. Preliminaries

State Space Models (SSMs) map input stimulation $x \in \mathbb{R}^1$ to output response $y \in \mathbb{R}^1$ through a hidden state $\boldsymbol{h} \in \mathbb{R}^{N\times 1}$ based on continuous linear time-invariant (LTI) systems, where N represents the number of states. To integrate deep models and adapt to real-world data, discretization must be applied to convert the continuous differential equations of SSMs into discrete functions using the zero-order hold method. Specifically, with a discrete-time step $\Delta \in \mathbb{R}^1$, SSMs are discretized as follows:

$$\boldsymbol{h}(t) = \tilde{\mathbf{A}}\boldsymbol{h}(t-1) + \tilde{\mathbf{B}}\mathbf{X}(t), \tag{1}$$
$$\mathbf{Y}(t) = \mathbf{C}^\top \boldsymbol{h}(t), \tag{2}$$

where $\mathbf{X}(t) = x(\Delta t)$, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the system's evolution matrix, and $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{N\times 1}$ are the projection matrices. $\tilde{\mathbf{A}} = \exp(\Delta\mathbf{A})$, $\tilde{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \approx \Delta\mathbf{B}$, where $\mathbf{I}$ denotes the identity matrix.

Selective State Space Models (S6) are introduced in Mamba [31] to improve the extraction of strong contextual information. S6 allows $\mathbf{B}$, $\mathbf{C}$, and Δ to vary as functions of the input $\mathbf{X}(t)$, whereas in S4 [28], $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and Δ are input-independent, which limits the model's ability to extract crucial information from the input sequence. Formally, given an input sequence $\mathbf{X} \in \mathbb{R}^{B\times L\times C}$, where B, L, and C represent the batch size, sequence length, and feature dimension, respectively, the input-dependent parameters $\mathbf{B}$, $\mathbf{C}$, and Δ are computed as follows:

$$\mathbf{B} = \texttt{Linear}(\mathbf{X}) \in \mathbb{R}^{B\times L\times N}, \tag{3}$$
$$\mathbf{C} = \texttt{Linear}(\mathbf{X}) \in \mathbb{R}^{B\times L\times N}, \tag{4}$$
$$\Delta = \texttt{SoftPlus}(\tilde{\Delta} + \texttt{Linear}(\mathbf{X})) \in \mathbb{R}^{B\times L\times C}, \tag{5}$$

where $\tilde{\Delta} \in \mathbb{R}^{B\times L\times C}$ is a learnable parameter, and $\mathbf{A} \in \mathbb{R}^{C\times N}$ is the the system's evolution matrix.

IV. High-Resolution Visual State Space Model

A. Multi-resolution Parallel VMamba

Existing visual Mamba models typically produce single-scale, low-resolution feature maps, which leads to significant information loss and hampers their ability to capture the fine-grained details and multi-scale variations required for dense prediction tasks. To overcome this limitation, we incorporate the multi-resolution parallel design from HRNet [34] and propose High-Resolution Visual State Space

TABLE I: **The architecture configuration of HRVMamba.** ESInB, and DSS2D represent the ESInB block and DSS2D block respectively. (M_1, M_2, M_3, M_4): the number of blocks, (B_1, B_2, B_3, B_4): the number of blocks, (S_1, S_2, S_3, S_4): the SSM expansion ratios, (R_1, R_2, R_3, R_4): the MLP expansion ratios.

Res.	Stage 1	Stage 2	Stage 3	Stage 4
4×	$\begin{bmatrix} 1\times1, (C_{out}, C_{in}) = C_0, 64 \\ 3\times3, (C_{out}, C_{in}) = C_0, C_0 \\ 1\times1, (C_{out}, C_{in}) = 4\times C_0, C_0 \end{bmatrix} \times B_1 \times M_1$	$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_1 \\ \text{FFN}, R_1 \end{bmatrix} \times B_2 \times M_2$	$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_1 \\ \text{FFN}, R_1 \end{bmatrix} \times B_3 \times M_3$	$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_1 \\ \text{FFN}, R_1 \end{bmatrix} \times B_4 \times M_4$
8×		$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_2 \\ \text{FFN}, R_2 \end{bmatrix} \times B_2 \times M_2$	$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_2 \\ \text{FFN}, R_2 \end{bmatrix} \times B_3 \times M_3$	$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_2 \\ \text{FFN}, R_2 \end{bmatrix} \times B_4 \times M_4$
16×			$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_3 \\ \text{FFN}, R_3 \end{bmatrix} \times B_3 \times M_3$	$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_3 \\ \text{FFN}, R_3 \end{bmatrix} \times B_4 \times M_4$
32×				$\begin{bmatrix} \text{ESInB} \\ \text{DSS2D}, S_4 \\ \text{FFN}, R_4 \end{bmatrix} \times B_4 \times M_4$

TABLE II: **Architecture details of HRVMamba variants.** $(C_0, C_1, C_2, C_3, C_4)$ is defined in Section IV-A.

Model	#channels $(C_0, C_1, C_2, C_3, C_4)$	#blocks (B_1, B_2, B_3, B_4)	#blocks (M_1, M_2, M_3, M_4)	#SSM ratio (S_1, S_2, S_3, S_4)	#MLP ratio (R_1, R_2, R_3, R_4)
HRVMamba-Nano	$(16, 8, 16, 32, 64)$	$(2, 2, 2, 2)$	$(1, 1, 4, 2)$	$(2, 2, 2, 2)$	$(2, 2, 2, 2)$
HRVMamba-Tiny	$(32, 16, 32, 64, 128)$	$(2, 2, 2, 2)$	$(1, 1, 4, 2)$	$(2, 2, 2, 2)$	$(2, 2, 2, 2)$
HRVMamba-Small	$(64, 32, 64, 128, 256)$	$(2, 2, 2, 2)$	$(1, 1, 4, 2)$	$(2, 2, 2, 2)$	$(2, 2, 2, 2)$
HRVMamba-Base	$(64, 80, 160, 320, 640)$	$(2, 2, 2, 2)$	$(1, 1, 4, 2)$	$(2, 2, 2, 2)$	$(2, 2, 2, 2)$

Model (HRVMamba), which maintains high-resolution representations throughout the network.

The overall architecture of HRVMamba is illustrated in Fig. 3(a). Given an input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$, the network begins with a downsampling stem consisting of two 3×3 convolutional layers with stride 2, reducing the spatial resolution to $\frac{H}{4} \times \frac{W}{4}$. The backbone is composed of four stages, where each stage progressively introduces a lower-resolution branch while retaining all higher-resolution streams from the previous stage. By the final stage, the network maintains four parallel branches with spatial dimensions of $\frac{H}{4} \times \frac{W}{4} \times C_1$, $\frac{H}{8} \times \frac{W}{8} \times C_2$, $\frac{H}{16} \times \frac{W}{16} \times C_3$, and $\frac{H}{32} \times \frac{W}{32} \times C_4$, respectively. Motivated by prior work [26], [27], which highlights the effectiveness of convolutional operations on higher-resolution feature maps in early stages, we employ a standard Bottleneck block (as in HRNet) in the first stage. In the remaining stages, we use our proposed Dynamic Visual State Space (DVSS) block (Fig. 3(c)) as the primary building unit. To enable information exchange across resolutions, we follow HRNet's multi-scale fusion strategy, which leverages a series of upsampling and downsampling blocks to aggregate features from different branches.

B. Dynamic Visual State Space (DVSS) Block

We introduce the Dynamic Visual State Space (DVSS) Block. As illustrated in Fig. 3 (b), the DVSS Block incorporates the Enhanced Spatial Inductive Bias Block (ESInB), the Deformable 2D-Selective-Scan (DSS2D) block, and a Feed-Forward Network (FFN) as its primary feature extraction blocks.

Enhanced Spatial Inductive Bias Block (ESInB) Block. While Visual Mamba establishes a global receptive field by treating images as token sequences and scanning them bidirectionally or along four directions, this sequential processing inherently disrupts the natural 2D spatial layout of visual data. As a result, the model lacks spatial inductive bias, which is crucial for capturing local textures, edges, and fine-grained details, particularly important in dense prediction tasks such as human pose estimation.

To compensate for this deficiency, we propose the Enhanced Spatial Inductive Bias Block, which explicitly incorporates local context through multi-scale convolutional operations. By leveraging multiple convolutional kernels of varying sizes, ESInB is designed to effectively model visual patterns at different spatial granularities, thereby restoring spatial awareness and enhancing the model's capacity to generalize to various scales and poses. Concretely, given an input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we first partition it along the channel dimension into G groups, where $G = 4$ by default. Each group $\mathbf{X}_g$ is processed independently by a depthwise convolution with a distinct kernel size $K_g = 2g+1$, allowing each group to focus on a specific spatial scale. This design introduces scale-specific locality into the model without significantly increasing computation. The outputs of these convolutions, denoted as $\mathbf{Y}_g$, are then concatenated along the channel dimension and passed through a channel shuffle operation, which facilitates information exchange across different groups and prevents isolated feature learning. Finally, a GELU nonlinearity is applied to improve expressiveness:

$$[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_G] = \text{Split}(\mathbf{X}, \text{axis=-1}), \tag{6}$$

$$\mathbf{Y}_g = \text{DWConv}_g(K_g \times K_g)(\mathbf{X}_g), \tag{7}$$

$$\mathbf{Y} = \text{GELU}(\text{Shuffle}(\text{Concat}([\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_G], \text{axis=-1})), \tag{8}$$

where Shuffle denotes the channel shuffle operation. The ESInB block provides a lightweight yet effective enhancement

TABLE III: **Comparison on the COCO pose estimation `val` set.** "*Trans.*" means transformer architecture. – means the numbers are not provided in the original paper. † marks a model that is not pretrained on ImageNet [37], while ‡ signifies that the backbone uses the classic decoder from ViTPose [2]. The #param. and FLOPs of HRFormer [9] are based on the implementation from MMPOSE [38].

Arch.	Method	Input Size	#param.	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
CNN	HRNet-W48 [34]	256 × 192	63.6M	14.6G	75.1	90.6	82.2	71.5	81.8	80.4
	FlashInternImage-B‡ [19]	256 × 192	100.7M	17.0G	74.1	90.6	82.0	70.3	80.4	79.3
Trans.	PRTR [39]	512 × 384	57.2M	37.8G	73.3	89.2	79.9	69.0	80.9	80.2
	TransPose-H-A6 [40]	256 × 192	17.5M	21.8G	75.8	–	–	–	–	80.8
	TokenPose-L/D24 [41]	256 × 192	27.5M	11.0G	75.8	90.3	82.5	72.3	82.7	80.9
	HRFormer-Tiny [9]	256 × 192	2.1M	1.1G	68.3	87.9	76.0	65.0	74.7	74.7
	HRFormer-Small [9]	256 × 192	7.7M	3.3G	74.0	90.2	81.2	70.4	80.7	79.4
	HRFormer-Base [9]	256 × 192	43.2M	14.1G	75.6	90.8	82.8	71.7	82.6	80.8
	Swin-B‡ [10]	256 × 192	94.0M	19.0G	73.7	90.5	82.0	70.2	80.4	79.3
	PVTv2-B2‡ [42]	256 × 192	29.1M	5.1G	73.7	90.5	81.2	70.0	80.6	79.1
	ViTPose-B [2]	256 × 192	90.0M	17.9G	75.8	90.7	83.2	68.7	78.4	81.1
	HRFormer-Small [9]	384 × 288	7.7M	7.3G	75.6	90.3	82.2	71.6	82.5	80.7
	HRFormer-Base [9]	384 × 288	43.2M	30.9G	77.2	91.0	83.6	73.2	84.2	82.0
SSM	Vim-S‡ [13]	256 × 192	28.0M	6.1G	69.8	89.2	78.2	67.2	75.5	76.0
	VMamba-T‡ [14]	256 × 192	34.7M	6.0G	74.4	90.4	82.3	70.8	81.0	79.6
	VMamba-B‡ [14]	256 × 192	93.8M	16.3G	74.8	90.7	82.1	71.2	81.5	80.1
	LocalVMamba-S‡ [15]	256 × 192	54.2M	14.1G	74.1	90.4	81.8	70.9	80.4	79.9
	MambaVision-B‡ [18]	256 × 192	102.9M	24.6G	73.4	90.1	80.9	69.7	80.2	78.9
	GroupMamba-B‡ [17]	256 × 192	57.7M	15.0G	73.2	90.3	81.1	69.8	79.8	78.7
SSM (Ours)	HRVMamba-Tiny	256 × 192	2.3M	1.1G	69.5	88.3	77.0	66.2	75.8	76.1
	HRVMamba-Small	256 × 192	8.0M	3.3G	74.6	90.5	81.7	71.1	81.0	79.9
	HRVMamba-Base	256 × 192	47.1M	14.2G	76.5	90.9	83.6	73.0	82.8	81.7
	HRVMamba-Small†	384 × 288	8.0M	7.4G	75.2	90.3	82.1	71.7	81.6	80.3
	HRVMamba-Small	384 × 288	8.0M	7.4G	76.4	90.9	83.3	72.7	83.0	81.3
	HRVMamba-Base†	384 × 288	47.1M	32.0G	77.6	91.1	**84.2**	**74.0**	84.2	82.4
	HRVMamba-Base	384 × 288	47.1M	32.0G	**77.7**	**91.2**	**84.2**	**74.0**	**84.3**	**82.5**

mechanism for capturing local spatial dependencies, which are often diminished in pure state space architectures.

Deformable 2D-Selective-Scan (DSS2D) Block. To further alleviate long-range forgetting and enhance spatial dependency modeling, we systematically explored a range of architectural design alternatives in Section V-D2. Specifically, we investigated the use of dilated convolutions and deformable operations, both capable of enhancing long-range spatial interactions. The experimental results in Section V-D2, along with our subsequent theoretical analysis, identify the deformable operation as a particularly effective mechanism.

This choice is theoretically grounded in the characteristics of SSMs. As described in prior work [43], an SSM applied to a sequence $\boldsymbol{x} \in \mathbb{R}^{1\times L\times C}$ defines the contribution of the m-th token to the n-th token ($m < n$) as:

$$\mathbf{C}_n^\top \prod_{i=m}^{n} \tilde{\mathbf{A}}_i \tilde{\mathbf{B}}_m = \mathbf{C}_n^\top \tilde{\mathbf{A}}_{(m\to n)} \tilde{\mathbf{B}}_m, \tag{9}$$

where $\tilde{\mathbf{A}}_{(m\to n)} = \exp\left(\sum_{i=m}^{n} \Delta_i \mathbf{A}\right)$ acts as an exponential decay factor along the sequence. Due to most $\Delta_i \mathbf{A}$ being negative, $\tilde{\mathbf{A}}_{(m\to n)}$ diminishes with larger $|n-m|$, leading to the long-range forgetting issue. This decay causes the model to underutilize distant contextual information and overemphasize low-level signals (see Fig. 2, column 2).

To overcome this degradation, we propose the **DSS2D block**, which injects a deformable sampling operation [19] into the State Space 2D-Selective-Scan (SS2D) structure. By allowing the model to dynamically sample from semantically important but spatially distant regions, DSS2D enhances the spatial expressiveness of SSMs and supports adaptive, context-aware feature integration.

Concretely, given a spatial input $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, the deformable operation with $K = 9$ sampling points per group is defined as:

$$\mathbf{Y}_g = \sum_{k=1}^{K} \mathbf{m}_{gk} \mathbf{X}_g(p_o + p_k + \Delta p_{gk}), \tag{10}$$

$$\mathbf{Y} = \mathrm{Concat}([\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_G], \mathrm{axis} = -1), \tag{11}$$

where G is the number of groups, p_k are regular grid offsets, Δp_{gk} are learnable input-dependent offsets, and $\mathbf{m}_{gk}$ are modulation scalars. These allow the model to sample informative features beyond fixed neighborhoods.

This operation is embedded into the state update rule of the SSM, replacing the fixed token input $x(p_o)$ with a spatially aggregated representation:

$$\boldsymbol{h}_g(p_o) = \tilde{\mathbf{A}}\boldsymbol{h}_g(p_o - 1) + \tilde{\mathbf{B}} \sum_{k=1}^{K} \mathbf{m}_{gk} \mathbf{X}_g(p_o + p_k + \Delta p_{gk}). \tag{12}$$

This integration mitigates the exponential decay of long-range dependencies by injecting and amplifying salient signals from relevant spatial regions into the SSM recurrence, enabling each hidden state update to incorporate flexible and task-adaptive spatial context, even from spatially distant regions.

C. HRVMamba Architecture Instantiation

We present the architectural configurations of the proposed HRVMamba model in Table I. The design follows a hierarchical multi-stage structure, where each stage progressively

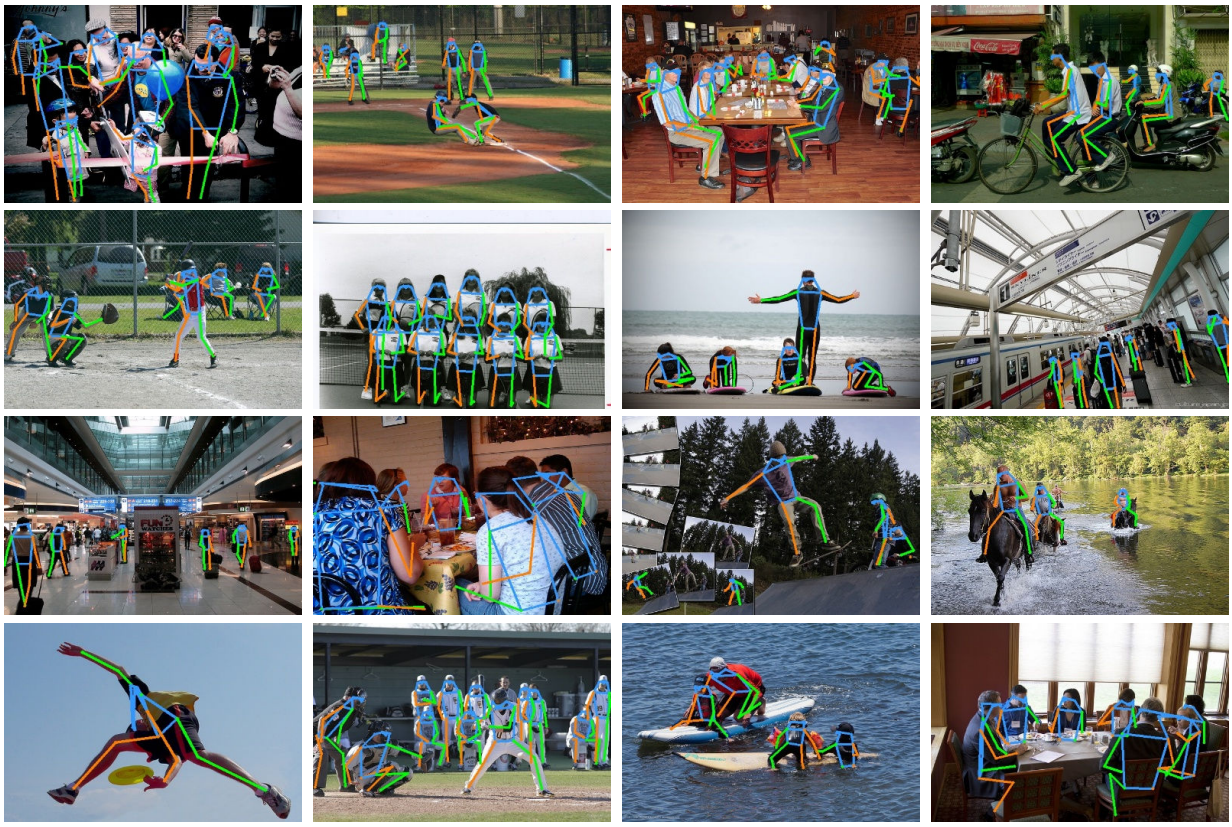


Fig. 4: **Example qualitative results on COCO pose estimation.** HRVMamba-Base is used as the backbone for pose estimation. The input size is 384×288.

refines the spatial resolution and semantic abstraction. Specifically, in the i-th stage, we denote B_i as the number of DSS2D blocks, S_i as the expansion ratio of the state space block (SSM), R_i as the MLP expansion ratio, and M_i as the number of stacked blocks, which jointly control the model's representation capacity and computational complexity.

To accommodate different deployment scenarios and computational budgets, we instantiate HRVMamba in four variants: HRVMamba-Nano, HRVMamba-Tiny, HRVMamba-Small, and HRVMamba-Base. These versions differ primarily in terms of width, depth, and expansion ratios used in the SSM and MLP, enabling a flexible trade-off between accuracy and efficiency. The detailed configuration for each variant is summarized in Table II. We also integrate LPU block [44] and SE block [45] for HRVMamba-Nano and HRVMamba-Tiny. Our architectural characteristics are particularly beneficial for dense prediction tasks such as human pose estimation, where multi-scale spatial precision and long-range feature propagation are both crucial.

V. Experiments

In this section, we compare the performance of HRVMamba with other state-of-the-art networks across several tasks: human pose estimation (COCO), image classification (ImageNet-1K), and semantic segmentation (ADE20K). We then analyze the ablation effects of the Multi-Resolution Parallel Architecture, the Deformable 2D-Selective-Scan block, and the Enhanced Spatial Inductive Bias Block.

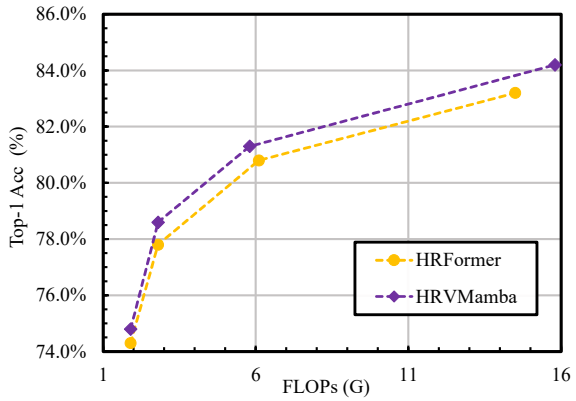


Fig. 5: **The trade-off between Top-1 accuracy and FLOPs on ImageNet val set for high-resolution models.**

A. Human Pose Estimation

Training setting. We evaluate HRVMamba on COCO dataset [47] for human pose estimation, which comprises over 200,000 images and 250,000 labeled person instances with 17 keypoints. Our experiments are trained on the COCO `train` 2017 dataset, which includes 57,000 images and 150,000 person instances. The performance of our model is assessed on the `val` 2017 and `test-dev` 2017 sets, comprising 5,000 and 20,000 images, respectively. For training and evaluation, we follow the implementation of MMPOSE [38]. The AdamW

TABLE IV: **Comparison on the COCO pose estimation `test-dev` set.** † marks a model that is not pretrained, while ‡ signifies that the backbone uses the classic decoder from ViTPose.

Method	Input Size	#param.	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HRNet-W48 [34]	384 × 288	63.6M	32.9G	75.5	92.5	83.3	71.9	81.5	80.5
PRTR [39]	512 × 384	57.2M	37.8G	72.1	90.4	79.6	68.1	79.0	79.4
TransPose-H-A6 [40]	256 × 192	17.5M	21.8G	75.0	92.2	82.3	71.3	81.1	–
TokenPose-L/D24 [41]	384 × 288	29.8M	22.1G	75.9	92.3	83.4	72.2	82.1	80.8
HRFormer-Small [9]	384 × 288	7.7M	7.3G	74.5	92.3	82.1	70.7	80.6	79.8
HRFormer-Base [9]	384 × 288	43.2M	30.9G	76.2	92.7	83.8	72.5	82.3	81.2
HRFormer-Base† [9]	384 × 288	43.2M	30.9G	76.0	92.6	83.6	72.9	81.5	81.0
Swin-L‡ [10]	384 × 288	207.9M	88.2G	75.4	92.6	83.3	72.0	80.9	80.5
ViTPose-B [2]	256 × 192	90.0M	17.9G	75.1	92.5	83.1	72.0	80.7	80.3
VMamba-B‡ [14]	384 × 288	93.8M	36.6G	75.3	92.7	83.3	72.0	80.9	80.3
HRVMamba-Small	384 × 288	8.0M	7.4G	75.3	92.5	83.1	72.1	80.9	80.3
HRVMamba-Base†	384 × 288	47.1M	32.0G	76.5	92.6	84.2	**73.5**	81.8	81.4
HRVMamba-Base	384 × 288	47.1M	32.0G	**76.7**	**92.8**	**84.4**	73.4	**82.2**	**81.5**

TABLE V: **Comparison with the state-of-the-art on ImageNet.** "iso.", "hie.", "hig." represent isotropic architecture without downsampling layers, hierarchical architecture, high-resolution architecture, respectively. † indicates that this implementation aligns the structure definition and classification header of our HRVMamba and adopts the basic block of HRFormer provided by MMPOSE [38].

Type	Arch.	Model	Ref.	Input Size	#Param (M)	FLOPs (G)	Top-1 Acc
iso.	*CNN*	ConvNeXt-S [20]	CVPR'2022	224^2	22	4.3	79.7
		ConvNeXt-B [20]	CVPR'2022	224^2	87	16.9	82.0
	Trans.	DeiT-S [24]	PMLR'2021	224^2	22	4.6	79.8
		DeiT-B [24]	PMLR'2021	224^2	87	17.6	81.8
	SSM	S4ND-ViT-B [32]	NeurIPS'2022	224^2	89	-	80.4
		Vim-Ti [13]	ICML'2024	224^2	7	1.1	76.9
		Vim-S [13]	ICML'2024	224^2	26	4.3	80.5
		VideoMamba-S [16]	ECCV'2024	448^2	26	16.9	83.3
		PlainMamba-L3 [33]	BMVC'2024	224^2	50	14.4	82.3
		VideoMamba-M [16]	ECCV'2024	576^2	75	83.1	**84.0**
hie.	*CNN*	ConvNeXt-T [20]	CVPR'2022	224^2	29	4.5	82.1
		ConvNeXt-B [20]	CVPR'2022	224^2	89	15.4	83.8
		MambaOut-B [46]	CVPR'2025	224^2	85	15.8	84.2
	Trans.	Swin-T [10]	CVPR'2021	224^2	28	4.5	81.3
		Swin-B [10]	CVPR'2021	224^2	88	15.4	83.5
	SSM	VMamba-B [14]	NeurIPS'2024	224^2	89	15.4	83.9
		LocalVMamba-S [15]	ECCV'2024	224^2	50	11.4	83.7
		MambaVision-B [18]	CVPR'2025	224^2	50	15.0	84.2
		GroupMamba-B [17]	CVPR'2025	224^2	57	14.0	**84.5**
hig.	*Trans.*	HRFormer-Nano† [9], [38]	NeurIPS'2021	256^2	12	1.9	74.3
		HRFormer-Tiny† [9], [38]	NeurIPS'2021	256^2	14	2.8	77.8
		HRFormer-Small† [9], [38]	NeurIPS'2021	256^2	20	6.1	80.8
		HRFormer-Base† [9], [38]	NeurIPS'2021	224^2	57	14.5	83.3
		HRFormer-Base [9]	NeurIPS'2021	224^2	50	13.7	82.8
	SSM	HRVMamba-Nano	our	256^2	12	1.9	74.8
		HRVMamba-Tiny	our	256^2	14	2.8	78.6
		HRVMamba-Small	our	256^2	20	5.8	81.3
		HRVMamba-Base	our	224^2	61	15.8	**84.2**

optimizer is used, configured with a learning rate of 5e-4, betas of (0.9, 0.999), and a weight decay of 0.01.

Results. Table III presents the results on the COCO `val` dataset. HRVMamba consistently outperforms other CNN models, ViT models, and recent state-of-the-art SSM methods. With an input size of 256 × 192, HRVMamba-Small achieves 74.6 AP, exceeding FlashInternImage-B (74.1 AP) while using only one-fifth of the FLOPs. HRVMamba-Base achieves 76.5 AP, surpassing state-of-the-art SSM methods like Vim-S, VMamba-B, MambaVision-B, and GroupMamba-B. At similar computational complexity, HRVMamba-Base improves by 3.3 AP and 3.0 AR over GroupMamba-B. Additionally, HRVMamba-Base outperforms ViTPose-B by 0.7 AP and 0.6 AR, with 50% fewer parameters and 20% fewer FLOPs. With an input size of 384 × 288, HRVMamba-Small achieves a 0.8 AP improvement over HRFormer-Small. HRVMamba-Base achieves a 0.5 AP gain and a 0.5 AR improvement over HRFormer-Base with pretraining on ImageNet, while it gains 0.6 AP and 0.6 AR over HRFormer-Base without pretraining on ImageNet. We present the pose estimation results of HRVMamba-Base pretrained on ImageNet in Fig. 4, which demonstrate that HRVMamba-Base effectively handles challenges such as viewpoint change, occlusion, and multiple persons.

We also provide comparisons on the COCO `test-dev` set in Table IV. Our HRVMamba-Small achieves an AP of 75.3, outperforming ViTPose-B by 0.2 while using only one-eleventh of its parameters. It matches the performance of

TABLE VI: **Performance comparison for semantic segmentation.** We report the mIoUs on ADE20K `val`. 'SS' and 'MS' denote evaluations performed at single-scale and multi-scale levels, respectively. All FLOPs are calculated with the input resolution fixed at 512×2048. Results of other methods are taken directly from their original papers. "-" indicates that the corresponding result was not provided in the original paper.

Method	#Param.	FLOPs	ADE20K	
			mIoU (SS)	mIoU (MS)
Swin-B [10]	121M	1188G	48.1	49.7
ConvNeXt-B [20]	122M	1170G	49.1	49.9
HRFormer-Base [9]	56M	1120G	48.7	50.0
VMamba-B [14]	122M	1170G	51.0	51.6
LocalVMamba-S [15]	81M	1095G	50.0	51.0
MambaOut-B [46]	112M	1178G	49.6	51.0
MambaVision-B [18]	126M	1342G	49.1	-
HRVMamba-Base	99M	1184G	**51.4**	**52.2**

VMamba-B, but with just one-fifth of the FLOPs. Furthermore, HRVMamba-Base surpasses HRFormer-Base by 0.5 in AP and 0.4 in AR without pretraining on ImageNet. After pretraining on ImageNet, HRVMamba-Base achieves 76.7 AP and 81.5 AR, setting a new state-of-the-art performance.

B. ImageNet Classification

Training setting. We conduct experiments on the ImageNet-1K dataset [37], which consists of 1.28M training images and 50K validation images across 1000 categories. HRVMamba is trained using the Swin Transformer [10] training framework on 80GB A100 GPUs. Specifically, we adopt the AdamW optimizer for 300 epochs with a cosine learning rate decay schedule and a 20-epoch linear warm-up phase. The initial learning rate is set to 0.002, and the weight decay is 0.05. We also apply exponential moving average (EMA), with MixUp and CutMix augmentation strategies set to 0.8 and 1.0, respectively. Due to inconsistencies between the official HRFormer implementation [9] and the HRFormer blocks employed in MMPOSE [38], we adopt the MMPOSE version, which is more commonly used in dense prediction tasks. As MMPOSE does not provide a classification head for HRFormer, we utilize a unified classification head based on HRVMamba for all comparisons. For HRFormer variants that are not defined in MMPOSE, such as HRFormer-Nano and HRFormer-Tiny, we follow the architectural specifications $(C_0, C_1, C_2, C_3, C_4)$ and (B_1, B_2, B_3, B_4) of HRVMamba as detailed in Table II. To ensure a fair comparison, all HRFormer results reported in this section are based on our reimplementation under identical training settings as HRVMamba. The pretrained weights and source code are publicly released via our GitHub repository.

Results. Table V compares HRVMamba with several representative CNN, ViT, and SSM methods. HRVMamba consistently demonstrates competitive performance across isotropic architectures [16], [24], hierarchical architectures [10], [14], and high-resolution architectures [9]. Specifically, HRVMamba-Base achieves a Top-1 accuracy of 84.2%, while requiring only 19% of the FLOPs consumed by VideoMamba-M, which reaches a comparable 84.0% Top-1 accuracy. Notably, HRVMamba-Base achieves this result without relying on advanced training strategies, such as the LAMB optimizer used in MambaVision-B [18] (84.2% Top-1 accuracy) or the distillation approach adopted in GroupMamba-B [17] (84.5% Top-1 accuracy). Although HRVMamba-Base slightly underperforms GroupMamba-B in image classification task under similar FLOPs, it substantially outperforms it in dense prediction tasks; as shown in Table III, HRVMamba-Base achieves 76.5 AP, compared to 73.2 AP by GroupMamba-B. Among high-resolution models, HRVMamba establishes a new state-of-the-art, with HRVMamba-Tiny, HRVMamba-Small, and HRVMamba-Base surpassing MMPOSE [38]-reimplemented HRFormer-Nano, HRFormer-Tiny, HRFormer-Small, and HRFormer-Base by 0.5, 0.8, 0.5, and 0.9 points, respectively, and exceeding the original HRFormer-Base [9] by 1.4 points, all under comparable FLOPs. Furthermore, as illustrated in Figure 5, HRVMamba consistently delivers higher accuracy than HRFormer under equivalent computational budgets, underscoring its superior efficiency and strong potential for deployment on resource-constrained platforms.

C. Semantic Segmentation

Training setting. We adopt UPerNet [48] as the decoder head, with backbone networks initialized from ImageNet-1K pretrained weights [37]. The ADE20K dataset serves as our primary benchmark, offering 150 fine-grained semantic classes across 20K training, 2K validation, and 3K test images. Model optimization is performed using the AdamW optimizer with a base learning rate of 6×10^{-5} and a weight decay of 5×10^{-4}. Training is conducted over 160k iterations with a batch size of 16. A linear learning rate warm-up is applied during the first 1,500 iterations. Input images are resized to a fixed resolution of 512×512. Data augmentation strategies are consistent with standard practices, including random horizontal flips, scale jittering within a [0.5, 2.0] range, and photometric distortions. We report both single-scale and multi-scale performance on the ADE20K validation set for comprehensive evaluation.

Results. Table VI presents the results on the ADE20K `val` set. Under single-scale testing, HRVMamba-Base achieves a mIoU of **51.4**, surpassing recent SSM-based backbones such as MambaVision-B [18] by 2.3 mIoU, VMamba-B [14] by 0.4 mIoU, and the CNN-based MambaOut-B [46] by 1.8 mIoU. With multi-scale testing, HRVMamba-Base further improves to **52.2** mIoU, outperforming HRFormer [9] by 2.2 mIoU and VMamba-B [14] by 0.6 mIoU. Overall, HRVMamba-Base consistently delivers higher segmentation accuracy under comparable computational budgets (FLOPs), demonstrating its strong capability for dense prediction tasks and highlighting its potential as an efficient and effective backbone for semantic segmentation.

D. Ablation Experiments

1) Multi-resolution Parallel architecture: The results in Table III demonstrate that HRVMamba, utilizing the Multi-resolution Parallel architecture, achieved state-of-the-art performance in pose estimation. In particular, comparisons with

TABLE VII: **Performance comparison on the COCO pose estimation `val` set.** All models are trained from scratch without ImageNet pretraining.

Model	Input Size	Operation Strategy	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HRVMamba-Small	256×192	3×3 dilation convolution; dilation factors:1	71.5	89.1	78.7	68.4	77.5	77.0
HRVMamba-Small	256×192	3×3 dilation convolution; dilation factors:3	70.5	89.0	78.1	67.3	76.6	76.0
HRVMamba-Small	256×192	3×3 dilation convolution; dilation factors:5	70.1	88.5	77.9	67.0	76.3	75.8
HRVMamba-Small	256×192	3×3 dilation convolution; dilation factors:1,3,5	71.2	88.9	79.1	68.1	77.4	76.7
HRVMamba-Small	256×192	deformable operation	**73.1**	**89.6**	**80.8**	**69.9**	**79.2**	**78.4**

TABLE VIII: **Ablation Experiments Results on COCO `val` set.** All models are not pretrained on the ImageNet. The HRVMamba-Small in Table II is the basic architecture setting. The input size is 256×192. † denotes their is only 3×3 depthwise convolution in ESInB Block.

SS2D Block	DSS2D Block	ESInB Block	ESInB in FFN	AP	AR
✔	✗	✗	✗	70.3	75.9
✗	✔	✗	✗	72.7	78.0
✗	✔	✗	✔	72.9	78.3
✗	✔	✔	✗	**73.1**	**78.4**
✗	✔	✔†	✗	72.8	78.3

other state-of-the-art SSM models such as VMamba [14], VMamba-B [14], MambaVision-B [18], and GroupMamba-B [17] highlight the advantages of the Multi-resolution Parallel architecture for dense prediction tasks.

2) Operation Strategies to Alleviate Long-Range Forgetting: To investigate the impact of spatial operator design on mitigating long-range forgetting, we evaluate several operation strategies within the HRVMamba-Small framework. As shown in Table VII, we explore dilated convolutions with different dilation factors and deformable operations, aiming to enhance long-range spatial interactions through adaptive spatial sampling.

We observe that using a standard dilation factor of 1 provides a solid baseline (AP: 71.5), while increasing the dilation factor to 3 or 5 leads to slight performance drops (AP: 70.5 and 70.1, respectively). This suggests that excessive dilation may disrupt local continuity and limit the model's ability to maintain consistent spatial relationships across joints. Employing multiple dilation factors (1, 3, 5) slightly recovers the performance (AP: 71.2). Remarkably, introducing deformable operations yields the best performance (AP: 73.1, AR: 78.4), highlighting their effectiveness in adaptively capturing spatially distributed dependencies. The learnable offsets in deformable convolutions enable the model to dynamically focus on semantically relevant positions, thereby strengthening the spatial coherence between distant keypoints and mitigating long-range forgetting.

These findings demonstrate that enhancing long-range spatial communication (deformable operation) rather than merely increasing receptive field size (dilated convolutions with different dilation factors) is crucial for robust pose estimation in visual SSMs. Deformable operations serve as a key mechanism to reinforce such dependencies in a data-driven, context-aware manner.

3) Deformable 2D-Selective-Scan Block: As shown in Table VIII, the DSS2D block improves AP by 2.4 points compared to the SS2D block (row 2 vs. row 1), demonstrating that incorporating deformable operations enhances Mamba's spatial feature extraction. Specifically, as shown in Fig. 2, DSS2D focuses on high-level features related to the query patch in the early stage (S2), while SS2D targets low-level edge features. In the later stage (S3), DSS2D highlights human-related details, whereas SS2D tends to capture irrelevant background information. We think deformable operations enhance the features of high-level semantic relations between patches, allowing them to influence each other despite long-range decay (long-range forgetting issue).

4) Enhanced Spatial Inductive Bias Block: HRFormer introduces depthwise convolution in the FFN to enhance the model's inductive bias. However, our experimental results in Table VIII demonstrate that directly embedding the ESInB Block into the FFN (row 3 vs. row 2) yields only marginal improvement, suggesting limited effectiveness in this setting. In contrast, using the ESInB Block as a standalone block (row 4) leads to a notable performance boost, achieving the best AP of 73.1 and AR of 78.4. Furthermore, replacing the multi-scale convolutional kernels in the ESInB Block with a simple 3×3 depthwise convolution (row 5) reduces performance (AP: 72.8), highlighting the importance of multi-scale convolutions in enhancing local spatial representations and inductive bias.

5) ImageNet Pretraining: Table III shows that pretraining on ImageNet notably improves the performance of the smaller model, HRVMamba-Small, in pose estimation. However, for the larger variant, HRVMamba-Base, the gain is only marginal. This observation suggests that the pretraining strategy suitable for SSM-based models may differ from those commonly employed for CNNs and Vision Transformers, likely due to the unique recurrence and information propagation mechanisms of SSMs. These findings highlight a promising direction for future research: developing tailored pretraining methods for HRVMamba to fully exploit its potential. Exploring alternative strategies may lead to further performance gains across various model scales.

VI. Conclusion

Visual Mamba faces notable challenges in dense prediction tasks, including weak spatial inductive bias, long-range forgetting due to hidden state decay, and degraded spatial precision from low-resolution outputs. To address these limitations, we propose the Dynamic Visual State Space (DVSS) block, which

enhances spatial inductive bias via multi-scale convolutions and alleviates long-range forgetting through input-adaptive deformable operations. Building upon the multi-resolution parallel architecture, we develop HRVMamba, a high-resolution visual state space model that preserves fine-grained spatial representations and supports efficient multi-scale feature learning. Importantly, HRVMamba achieves a superior trade-off between computational complexity and accuracy, making it well-suited for deployment in resource-constrained scenarios, such as mobile devices, where computational budgets are limited. Extensive experiments across human pose estimation, image classification, and semantic segmentation demonstrate that HRVMamba delivers competitive performance compared to state-of-the-art CNN-, ViT-, and SSM-based models.

REFERENCES

[1] J. Zhang, D. Zhang, H. Yang, Y. Liu, J. Ren, X. Xu, F. Jia, and Y. Zhang, "Mvpose: Realtime multi-person pose estimation using motion vector on mobile devices," *IEEE Transactions on Mobile Computing*, vol. 22, no. 6, pp. 3508–3524, 2023. 1, 3

[2] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose++: Vision transformer for generic body pose estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1212–1230, 2024. 1, 3, 6, 8

[3] H. Zhang, L. Xu, S. Lai, W. Shao, N. Zheng, P. Luo, Y. Qiao, and K. Zhang, "Open-vocabulary animal keypoint detection with semantic-feature matching," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5741–5758, 2024. 1, 3

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 1, 3

[5] H. Zhang, S. Lai, Y. Wang, Z. Da, Y. Dun, and X. Qian, "Scgnet: Shifting and cascaded group network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4997–5008, 2023. 1, 3

[6] H. Zhang, Y. Dun, Y. Pei, S. Lai, C. Liu, K. Zhang, and X. Qian, "Hf-hrnet: a simple hardware friendly high-resolution network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 4

[7] H. Zhang, Y. Ma, K. Zhang, N. Zheng, and S. Lai, "Fmgnet: An efficient feature-multiplex group network for real-time vision task," *Pattern Recognition*, p. 110698, 2024. 1, 3

[8] E. J. Roh, H. Baek, D. Kim, and J. Kim, "Fast quantum convolutional neural networks for low-complexity object detection in autonomous driving applications," *IEEE Transactions on Mobile Computing*, vol. 24, no. 2, pp. 1031–1042, 2025. 1

[9] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: high-resolution transformer for dense prediction," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021, pp. 7281–7293. 1, 3, 6, 8, 9

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 1, 3, 6, 8, 9

[11] Q. Wang, Z. Zheng, Q. Wang, D. Deng, and J. Zhang, "Generalizations of wearable device placements and sentences in sign language recognition with transformer-based model," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 10 046–10 059, 2024. 1

[12] J. Gong, Y. Liu, T. Li, J. Ding, Z. Wang, and D. Jin, "STTF: A spatiotemporal transformer framework for multi-task mobile network prediction," *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 4072–4085, 2025. 1

[13] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: efficient visual representation learning with bidirectional state space model," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 62 429–62 442. 2, 3, 6, 8

[14] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024. 2, 3, 6, 8, 9, 10

[15] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "Localmamba: Visual state space model with windowed selective scan," in *European Conference on Computer Vision*. Springer, 2024, pp. 12–22. 2, 3, 6, 8, 9

[16] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," in *European Conference on Computer Vision*, 2024, pp. 237–255. 2, 8, 9

[17] A. Shaker, S. T. Wasim, S. Khan, J. Gall, and F. S. Khan, "Groupmamba: Efficient group-based visual state space model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 912–14 922. 2, 3, 6, 8, 9, 10

[18] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 261–25 270. 2, 3, 6, 8, 9, 10

[19] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao *et al.*, "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5652–5661. 3, 6

[20] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986. 3, 8, 9

[21] H. Zhang, W. Shao, H. Liu, Y. Ma, P. Luo, Y. Qiao, N. Zheng, and K. Zhang, "B-avibench: Towards evaluating the robustness of large vision-language model on black-box adversarial visual-instructions," *IEEE Transactions on Information Forensics and Security*, 2024. 3

[22] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu *et al.*, "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," *arXiv preprint arXiv:2404.16006*, 2024. 3

[23] S. Liu, K. Ying, H. Zhang, Y. Yang, Y. Lin, T. Zhang, C. Li, Y. Qiao, P. Luo, W. Shao *et al.*, "Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models," *arXiv preprint arXiv:2403.20194*, 2024. 3

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357. 3, 8, 9

[25] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 425–17 436. 3

[26] S. Yun and Y. Ro, "Shvit: Single-head vision transformer with memory efficient macro design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5756–5767. 3, 5

[27] X. Ma, X. Dai, J. Yang, B. Xiao, Y. Chen, Y. Fu, and L. Yuan, "Efficient modulation for vision networks," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3, 5

[28] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021. 3, 4

[29] J. T. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022. 3

[30] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *arXiv preprint arXiv:2212.14052*, 2022. 3

[31] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023. 3, 4

[32] E. Nguyen, K. Goel, A. Gu, G. Downs, P. Shah, T. Dao, S. Baccus, and C. Ré, "S4nd: Modeling images and videos as multidimensional signals with state spaces," *Advances in neural information processing systems*, vol. 35, pp. 2846–2861, 2022. 3, 8

[33] C. Yang, Z. Chen, M. Espinosa, L. Ericsson, Z. Wang, J. Liu, and E. J. Crowley, "Plainmamba: Improving non-hierarchical mamba in visual recognition," *arXiv preprint arXiv:2403.17695*, 2024. 3, 8

[34] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020. 3, 4, 6, 8

[35] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 440–10 450. 4

[36] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation," in *Proceedings of the Thirty-First International Joint Conference on Artificial*

Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, L. D. Raedt, Ed. ijcai.org, 2022, pp. 1095–1101. 4

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 6, 9

[38] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020. 6, 7, 8, 9

[39] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1944–1953. 6, 8

[40] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 802–11 812. 6, 8

[41] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 11 313–11 322. 6, 8

[42] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022. 6

[43] Y. Shi, M. Dong, and C. Xu, "Multi-scale vmamba: Hierarchy in hierarchy visual state space model," *arXiv preprint arXiv:2405.14174*, 2024. 6

[44] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 175–12 185. 7

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *CVPR*, 2018, pp. 7132–7141. 7

[46] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?" in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4484–4496. 8, 9

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014, pp. 740–755. 7

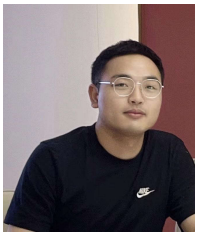
[48] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434. 9



Hao Zhang received a B.S. degree in information engineering from Xi'an Jiaotong University in 2021. He is currently pursuing a Ph.D. degree in artificial intelligence at Xi'an Jiaotong University. His research interests include neural network architecture design and Large Vision-Language Models.



Yongqiang Ma received an M.S. degree in software engineering from Xi'an Jiaotong University in 2015, and a Ph.D. degree in control science and engineering with Xi'an Jiaotong University in 2021. He is currently an assistant professor at Xi'an Jiaotong University. His research focuses on neuromorphic computing, spiking neural network, and cognitive Computing Model.



Wenqi Shao received the Ph.D. degree from Multimedia Lab, the Chinese University of Hong Kong (CUHK) in 2022. Now he is a researcher at Shanghai Artificial Intelligence Lab, Shanghai, China. His research interests lie in the pre-training, evaluation, applications of multimodal foundation models, as well as compression techniques and hardware codesign for large models.



Ping Luo received the Ph.D. degree in information engineering from the Chinese University of Hong Kong (CUHK). He is currently an associate professor with the Department of Computer Science, University of Hong Kong (HKU). He was a postdoctoral fellow in CUHK from 2014 to 2016. His research interests include machine learning and computer vision. He has published more than 100 peer-reviewed articles in top-tier conferences and journals.



Nanning Zheng graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975, and received the M.S. degree in information and control engineering from Xi'an Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. He joined Xi'an Jiaotong University in 1975, where he is currently a professor and the director of the Institute of Artificial Intelligence and Robotics. His research interests include computer vision, pattern recognition, and machine learning. Dr. Zheng became a member of the Chinese Academy of Engineering in 1999. He is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition.



Kaipeng Zhang received an M.S. degree from National Taiwan University, Taipei, Taiwan in 2018, and a Ph.D. degree from the University of Tokyo, Tokyo, Japan in 2022. Now he is a researcher at Shanghai Artificial Intelligence Lab, Shanghai, China. His current research interests include face analysis, active learning, and foundation vision models.