DTMPose: Depth Transform-enhanced Mamba Pose Estimation Framework for Efficient 2D Keypoint Detection

Guanting Dong[1 [0009−0000−2970−0180]] and Kei Kawamura[1]

Graduate School of Sciences and Technology for Innovation,
Yamaguchi University, 2-16-1 Tokiwadai, Yamaguchi 753-0841, Japan
redocting2022@outlook.com, kay@yamaguchi-u.ac.jp

Abstract. 2D keypoint detection plays an important role in the fields of group behavior analysis, motion capture, human-computer interaction, and security monitoring. However, in high-density crowd environments or edge devices with limited computational resources, it is still a major challenge to improve inference efficiency while ensuring detection accuracy. To this end, this paper proposes a keypoint detection framework called 'DTMPose', whose core innovation is to replace the computationally intensive attention module with a Mamba-based state-space model (SS2D mechanism) and to introduce a sense-field-enhanced convolution (e.g., 'DPConv') in the key parts, to improve the detection of local occlusion and edge details. Compared to models which only rely on the self-attention mechanism, DTMPose reduces the computational overheads, whilst still capturing global dependencies, and effectively mitigates local keypoint ambiguities through enhanced convolution. Experimental results on the COCO dataset show that DTMPose maintains a low parameter count with an accuracy of about 76% AP, demonstrating its deployment potential in high-density crowd scenarios and mobile edge devices, as well as providing a new feasible solution for applications such as people flow monitoring and group behavior analysis.

Keywords: 2D keypoint detection · State Space Model · Selective Scan (SS2D) · Pose Estimation · Lightweight Architecture.

1 Introduction

2D keypoint detection plays a crucial role in diverse application scenarios, including behavioral understanding, human-computer interaction, and motion capture. With the advancement of deep learning, two mainstream paradigms have emerged. Convolutional Neural Networks (CNNs), such as ResNet [1] and HRNet [2], offer efficient local perception and perform well at small to medium resolutions. In contrast, Transformer architectures [3], with global self-attention, demonstrate superior long-range dependency modeling and achieve high accuracy, as shown in ViTPose [4]. Nevertheless, practical limitations remain: CNNs

face high computational and memory costs in large-resolution or dense scenarios, while Transformers suffer from quadratic complexity with respect to feature map size, hindering deployment in resource-limited environments.

In recent years, state-space modeling (SSM)[5, 6] has introduced new perspectives to pose detection by enabling efficient modeling of long-range dependencies with linear or quasi-linear time complexity, especially in long sequences or high-resolution scenarios. Mamba[6], a refined SSM architecture, enhances this by integrating row and column features through a selective scan mechanism, thereby reducing parameters and computation while preserving modeling capacity. However, its reliance on global modeling may overlook local occlusions and fine-scale features, limiting performance in detail-sensitive tasks.

To better balance detection accuracy and computational efficiency, this paper proposes a novel keypoint detection framework, DTMPose. It integrates state-space modeling with locally enhanced convolution. Unlike conventional CNNs or Transformers, DTMPose employs a Mamba-based SS2D backbone to replace large-scale self-attention, reducing computational cost while preserving global dependency modeling. Additionally, depthwise partial convolution (e.g., DPConv) is inserted at key stages to improve representation of small-scale targets and locally occluded regions. A multi-scale feature fusion neck further facilitates the preservation of semantic and detailed features throughout tensor transformations during model training, thereby improving adaptability to diverse pose scenarios.

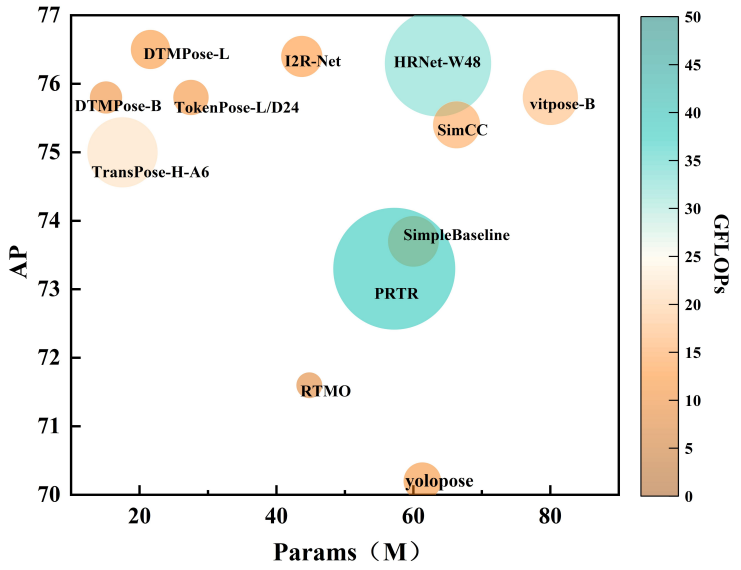

Fig. 1. Performance and parameter comparison of different attitude estimation models on the COCO val2017 dataset.

To visually highlight the distinctions between this study and existing methods, Fig.1 compares the overall performance of representative pose detection models[2, 4, 7–14]. The x-axis indicates model parameters (Params), the y-axis shows detection accuracy (AP: Average Precision), and circle area denotes model size, with a color gradient from orange (low) to cyan (high) reflecting GFLOPs, as detailed in the right-hand color bar. As illustrated, CNN-based models (e.g., ResNet, HRNet) are limited in accuracy, while Transformer-based models (e.g., ViTPose) offer higher accuracy but with greater computational and memory demands. In contrast, DTMPose achieves comparable or superior accuracy with significantly fewer parameters. This significantly reduces the overall number of parameters, providing a more scalable and practical solution for human posture recognition in resource-constrained environments.

The core contributions of this paper are as follows:

1. A novel backbone architecture that integrates state-space modeling with locally enhanced convolution is proposed, combining global dependency modeling and fine-grained feature extraction to improve detection accuracy.
2. Multi-scale feature fusion is employed to enhance local keypoint detection and improve robustness in occluded scenarios.
3. DTMPose maintains or improves detection accuracy while significantly reducing parameter count and computational overhead, offering a more lightweight and efficient solution for pose estimation tasks.

2 Related work

2D keypoint detection focuses on locating human joint positions in images or videos and plays a vital role in applications such as behavioral analysis, human-computer interaction, and security monitoring. With advances in deep learning, extensive research has aimed to improve inference efficiency while maintaining high accuracy. Existing methods mainly fall into two categories: convolutional neural network (CNN)-based and Transformer-based frameworks.

2.1 CNN-based Keypoint Detection Methods

CNN-based architectures have long dominated keypoint detection due to their strong local perception and computational efficiency. Early methods typically pair a backbone (e.g., ResNet) with upsampling or multi-resolution modules to enhance localization. For instance, SimpleBaseline [13] restores spatial resolution via deconvolution layers, while HRNet employs multi-resolution parallel branches to fuse features and maintain high-resolution representations, resulting in improved localization accuracy through stronger local feature representation.

2.2 Transformer-based Keypoint Detection Methods

The Transformer architecture, with its global self-attention mechanism, has demonstrated strong capability in modeling cross-region information and has

been increasingly adopted in keypoint detection tasks. For example, TransPose [14] applies Transformer layers to globally model CNN-extracted features, TokenPose encodes keypoint positions as tokens, and ViTPose leverages the Vision Transformer backbone to fully exploit global context modeling. While Transformer-based methods achieve competitive accuracy, particularly in high-density or high-resolution scenarios and when trained on large-scale datasets, their computational complexity scales quadratically with input size, often leading to redundant parameters and inefficiencies in model size and inference cost

2.3 State Space Modeling (SSM) Methods

Recently, State Space Models (SSMs) have shown favorable time complexity and generalization for sequence modeling and signal processing, offering new perspectives for dense prediction tasks like keypoint detection. These models capture long-range dependencies with lower computational cost and fewer parameters. For instance, Mamba uses a selective scan to integrate row and column features, avoiding the high overhead of self-attention in high-resolution images while maintaining global context modeling. However, due to limited local detail perception, complementary techniques—such as convolution, attention, or multiscale fusion—are often required to achieve balanced and accurate predictions.

3 Method

Starting from the idea of modules in the overall network structure, this chapter introduces the Mamba row-scanning mechanism (an alternative to self-attention for global dependency capture) and then describes the overall multilevel design of the DTMPose model, before detailing the introduction of feel-field-enhanced convolution (DPConv), as well as the Stem and Neck modules of the network, in order to balance the local details with the global modeling requirements.

3.1 Mamba: 2D-Selective Scan mechanism

The Transformer architecture models global dependencies via self-attention but suffers from $\mathcal{O}(n^2)$ complexity, limiting its scalability to high-resolution vision tasks. In contrast, the Mamba series introduces a row-column scanning mechanism based on State Space Models (SSMs), enabling global context modeling with approximately linear complexity and offering a more efficient infrastructure for visual representation learning.

Mamba-1: Selective State Space Model Mamba-1 [5] extends the classical linear time-invariant state-space model (LTI-SSM) by introducing input-dependent dynamic parameterization for enhanced adaptability. The continuous-time system is formulated as:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t), \tag{1}$$

which is discretized (via Zero-Order Hold) into:

$$h_t = \bar{A}h_{t-1} + \bar{B}_t x_t, \quad y_t = C_t h_t, \tag{2}$$

where $\bar{A}$ and $\bar{B}_t$ are learned or derived from A and B. Mamba introduces dynamic modulation:

$$\bar{B}_t = s_B(x_t), \quad C_t = s_C(x_t), \quad \Delta_t = \tau_A(\text{Param} + s_A(x_t)), \tag{3}$$

with $s_{\{\cdot\}}$ as learned functions and τ_A a softplus activation. This flexible design enables Mamba to handle diverse temporal dynamics. However, its sequential nature limits parallelism during training.

Mamba-2: Parallel Acceleration via State Space Duality To improve computational throughput, Mamba-2 [15] reformulates the state-space recurrence into a fully parallelizable matrix form using State Space Duality (SSD):

$$y = Mx, \quad M_{ji} = C_j^\top A_j \cdots A_{i+1} B_i, \tag{4}$$

and under tied transitions, the matrix M admits factorization:

$$M = L \circ (CB^\top), \tag{5}$$

where L is lower-triangular and $\circ$ denotes element-wise multiplication. This allows Mamba-2 to match attention-level expressiveness with linear time complexity, achieving 2–8× speedup over Mamba-1 in practice.

Two-dimensional selective scanning mechanism (2D-Selective Scan, SS2D) To better adapt the two-dimensional structure of images in image tasks, VMamba [16, 17] proposed a 2D-Selective Scan (SS2D) mechanism based on one-dimensional state-space modeling, to extend the state-space model to image space. As shown in Figure 2, SS2D achieves spatial sensory field enhancement by propagating the state along the row direction and column direction, respectively.

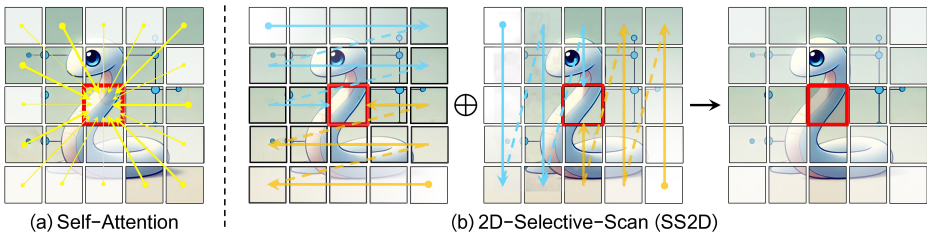

Fig. 2. Schematic diagram comparing the traditional self-attention mechanism with the 2D-Selective Scan (SS2D) mechanism.

The mechanisms are illustrated as follows:

(a) Self-Attention: Each pixel (in red) attends to all others via a global attention mechanism, enabling direct modeling of long-range dependencies but incurring high computational cost, particularly in high-resolution images.
(b) 2D-Selective Scan (SS2D): SS2D propagates state information horizontally and vertically (blue and orange arrows), enabling each pixel to capture long-range context along both axes. Unlike self-attention, SS2D adopts a state-space model with linear complexity.

In the final fusion (indicated by "⊕"), horizontal and vertical states are combined to approximate a global perceptual field with significantly lower computational cost. Compared to self-attention, SS2D achieves higher efficiency and stronger structural generalization, making it well-suited for 2D vision tasks such as recognition, segmentation, and pose estimation.

3.2 Overall model architecture: DTMPose

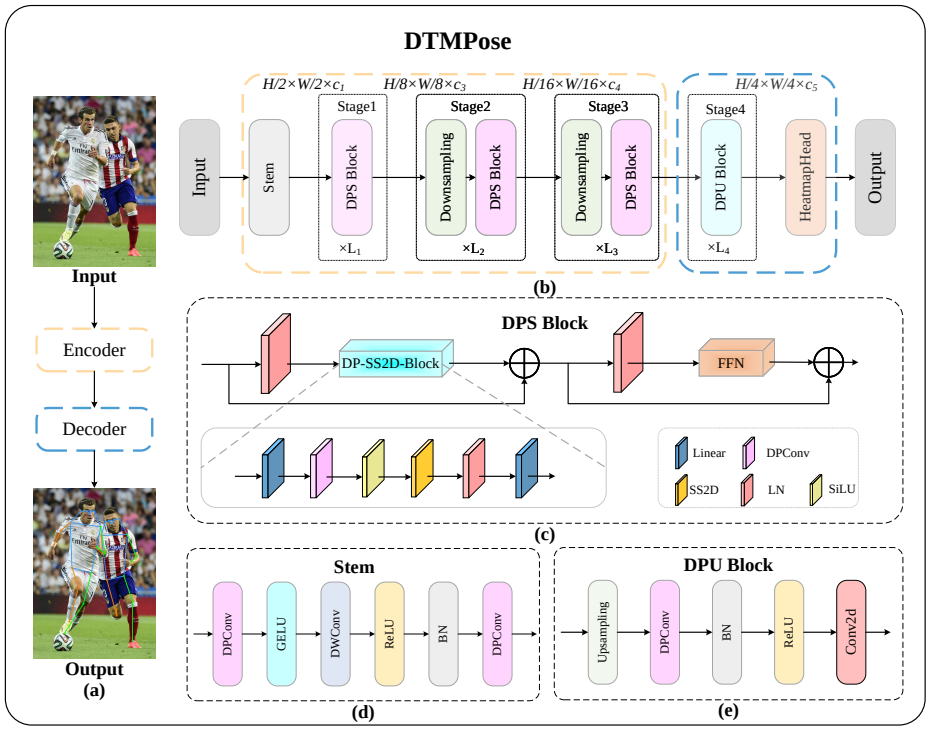

Fig. 3. (a) Overall processing flowchart of DTMPose; (b) Overall architecture of DTMPose; (c) Structure of DPS Block; (d) Structure of Stem Block; (e) Structure of DPU Block for downsampling and feature fusion.

To balance detection accuracy and computational efficiency in pose estimation tasks, we propose DTMPose, a dual-domain modeling framework that integrates state-space modeling with locally enhanced convolution. As shown in Figure 3, DTMPose consists of four main components:

- **Stem Module:** Performs initial downsampling and enhances low-level structural features.
- **Multi-stage Backbone:** Combines SS2D for global modeling and DPConv for local feature enhancement.
- **Decoder Path:** Gradually restores spatial resolution while refining pose-related representations.
- **Heatmap Head:** Generates 2D keypoint heatmaps for final pose estimation.

The DTMPose framework first applies the Stem module to extract low-dimensional features, followed by multi-stage encoding via DPS Blocks. Spatial resolution is then recovered using DPU Blocks, ultimately yielding the predicted human pose. This design effectively balances global modeling and local perception with high computational efficiency.

3.3 Backbone Network Design: DPS Block

The backbone of DTMPose adopts multi-stage stacked DPS Blocks, which integrate row-and-column state-space modeling (SS2D) with spatial-domain local enhancement (DPConv) to jointly capture global context and fine-grained structures. Each DPS Block comprises three parallel paths:

1. **SS2D Path (DPSD Block):** leverages Mamba's 2D Selective Scan to model long-range dependencies along rows and columns with linear complexity;
2. **Feedforward Path (FFN):** enhances channel-wise representation via nonlinear mapping;
3. **Locally Enhanced Path (DPConv):** employs receptive field enhanced convolution module to improve recognition in occluded, small-scale, and boundary regions.

A cross-path residual fusion mechanism integrates the outputs while preserving their respective advantages. The multi-stage structure enables hierarchical perception of semantic and spatial cues across scales, enhancing robustness in multi-gesture and complex-background scenarios.

3.4 Receptive Field Enhanced Convolution Module: DPConv

In human pose estimation, local regions such as limb overlaps and occlusions are common sources of error. To address this, DTMPose introduces DPConv—a lightweight receptive field enhancement module inspired by separable convolution [18]—to augment spatial perception and compensate for the limited spatial discrimination of state-space modeling.

DPConv (Figure 4) operates in three phases:

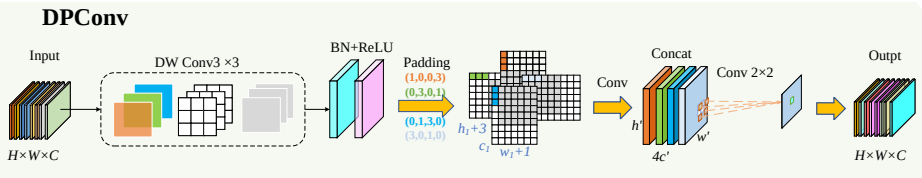

Fig. 4. DPConv Module Architecture.

1. **Multiscale modeling:** depthwise separable convolutions (3×3 DWConv) capture spatial patterns with reduced parameters;
2. **Directional enhancement:** asymmetric padding and direction convolutions extract boundary features to simulate bionic perception;
3. **Channel fusion:** directional features are concatenated and compressed to restore the original channel size while preserving spatial structure.

DPConv can be flexibly applied, particularly in the Stem and Decoder stages, to enhance recognition of occlusions and structural edges—critical for DTM-Pose's high accuracy and robustness.

3.5 Stem and DPU Block Implementation Details

Stem Module As shown in Figure 3(d), the Stem stage preprocesses the input and reduces spatial resolution. DPConv is introduced to enhance shallow feature responses to local structures. Combined with GELU activation, DWConv, and BN, it constructs a lightweight yet expressive low-dimensional feature space for the backbone.

DPU Block In the decoding stage, the DPU Block enhances spatial detail recovery by combining moderate upsampling with DPConv-based downsampling. This structure strengthens edge information perception while maintaining spatial consistency, improving localization accuracy in complex scenes.

Finally, the Heatmap Head projects the decoded features to keypoint heatmaps for high-precision 2D pose estimation.

4 Experimentation

4.1 Experimental setup

The purpose of this chapter is to provide a systematic evaluation of the performance of the DTMPose framework across multiple datasets and scenarios. We developed the empirical analysis around the following core questions:

1. Can DTMPose significantly reduce model complexity and inference cost while maintaining high accuracy under different architectures and module configurations?

2. Does DTMPose strike a better trade-off between cross-dataset performance, computational efficiency and accuracy than existing mainstream keypoint detection methods?
3. Can the integration of the DPConv local enhancement module within the Mamba state-space framework effectively expand the receptive field and improve the model's capacity for local feature modeling?

To comprehensively evaluate the performance of the model in human keypoint detection tasks, we conducted experiments on two widely used standard datasets: the COCO 2017 Keypoint Dataset [19] and the MPII Human Pose Dataset [20]. The experiments cover a variety of dimensions, such as modular ablation, method comparison, and heatmap visualization, to ensure the comprehensiveness and rigor of the evaluation.

The experiments were conducted using a single NVIDIA RTX 4090 GPU under the PyTorch framework, with the input resolution set to 256×192.The implementation was based on the MMPose [21] toolbox. The optimizer was AdamW, with an initial learning rate of 3×10^{-4} and a 500-step warm-up and poly decay strategy. The training period was set to 210 epochs and the batch size was dynamically adjusted according to the memory.

4.2 Comparative analysis of overall structural performance

To validate the effectiveness of the proposed framework, we conducted a comprehensive comparison between DTMPose and representative pose estimation models (e.g., HRNet, ViTPose) on the COCO 2017 and MPII datasets.

Table 1. Comparison of representative pose estimation models on the COCO val2017 dataset in terms of accuracy, model size, and computational cost.

Model	Input Size	Backbone	PT	Params	GFLOPs	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR
SimpleBaseline [13]	384×288	ResNet-152	Y	60.0	15.7	73.7	91.9	81.1	70.3	80.0	79.0
TokenPose-L/D24 [10]	256×192	HRNet-W48	Y	27.5	11.0	75.8	90.3	82.5	72.3	82.7	80.9
TransPose-H-A6 [14]	256×192	HRNet-W48	Y	17.5	21.8	75.0	92.2	82.2	71.3	81.1	80.8
HRNet-W48 [2]	256×192	HRNet-W48	Y	63.6	14.6	75.1	90.6	82.2	71.5	81.8	80.4
HRNet-W48 [2]	384×288	HRNet-W48	Y	63.6	32.9	76.3	90.8	82.9	72.3	83.4	81.2
I^2R-Net [7]	256×192	HRFormer-B	Y	43.7	12.8	76.4	90.8	83.2	72.3	83.7	81.4
Lite-HRNet [23]	384×288	Lite-HRNet-30	N	**1.8**	**0.7**	70.4	88.7	77.7	67.5	76.3	76.2
ViTPose-B [4]	256×192	ViT-B	Y	80.0	17.1	75.8	90.7	83.2	68.7	78.4	81.1
ViTPose-L [4]	256×192	ViT-B	Y	307.0	59.8	**78.3**	91.4	**85.2**	71.0	81.1	83.5
SwinPose [24]	384×384	Swin-L	Y	196.4	202.6	76.3	**93.5**	83.4	72.5	81.7	**84.7**
PRTR [8]	512×384	HRNet-W32	Y	57.2	37.8	73.3	89.2	79.9	69.0	80.9	80.2
RTMO [11]	256×192	CSPDarknet	Y	44.8	8.0	71.6	91.1	79.0	66.8	79.1	75.6
ED-Pose* [22]	256×192	ResNet-50	N	42.5	33.5	71.6	89.7	78.3	79.3	**94.3**	79.3
RTMPose-L* [25]	256×192	CSPNeXt-L	N	27.7	4.2	75.8	90.6	82.6	80.6	94.2	79.5
YOLOPose [12]	256×192	CSPDarknet	Y	61.3	11.7	70.2	91.1	77.8	65.3	78.2	74.3
SimCC [9]	256×192	HRNet-W48	Y	66.3	14.6	75.4	92.4	82.7	71.9	81.3	80.5
DTMPose-B*(Ours)	256×192	Mamba	N	15.1	10.0	75.8	90.3	82.6	72.4	82.3	80.8
DTMPose-L*(Ours)	256×192	Mamba	N	21.6	12.2	76.5	90.5	83.3	72.8	83.2	81.4

Models marked with "" are trained and evaluated using the official MMPose [21] framework.*

As shown in Table 1, on the COCO dataset, DTMPose-B and DTMPose-L achieve 75.8 and 76.5 AP, respectively, matching or surpassing ViTPose-B (75.8 AP) while using significantly fewer resources—15.1M/21.6M parameters and 10.0/12.2 GFLOPs, compared to ViTPose-B's 80M parameters and 17.1 GFLOPs.

These results demonstrate that DTMPose delivers comparable accuracy with much higher computational efficiency, highlighting its advantage in model compactness and practical applicability.

On the MPII dataset (Table 2), DTMPose-L achieves a PCK of 89.0, closely matching ViTPose-B (90.9) across keypoints such as the hip, knee, and ankle, and clearly outperforming LiteHRNet and the Mamba baseline.

Notably, DTMPose-B attains 88.3 PCK with only 15.1M parameters, demonstrating a strong trade-off between accuracy and efficiency.

Table 2. Keypoint-wise PCK comparison on the MPII validation set

Method	**Head**	**Sho.**	**Elb.**	**Wri.**	**Hip**	**Knee**	**Ank.**	**Mean**
LiteHRNet-18* [23]	96.1	93.7	85.5	79.2	87.0	80.0	75.1	85.9
LiteHRNet-30* [23]	96.3	94.7	87.0	80.6	87.1	82.0	77.0	87.0
HRNet-W32 [2]	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
PRTR [8]	**97.3**	96.0	90.6	84.5	89.7	85.5	79.0	89.5
TokenPose-L [10]	97.1	95.9	**91.0**	**88.5**	89.8	86.1	82.7	90.2
ViTPose-S [4]	96.4	94.7	88.1	83.2	88.4	84.3	80.0	88.4
ViTPose-B [4]	97.0	**96.2**	90.7	86.7	**90.4**	**88.2**	**84.2**	**90.9**
DTMPose-B*(Ours)	96.5	95.0	88.1	81.6	88.6	84.4	80.2	88.3
DTMPose-L*(Ours)	96.8	95.4	88.7	83.1	89.1	85.4	81.4	89.0

In summary, DTMPose performs well on both COCO and MPII benchmark datasets, especially in terms of the number of parameters and computational complexity, which is significantly reduced, compared to ViTPose-B, verifying the potential of our proposed framework for practical deployment.

4.3 Ablation Study of Local-Perception Modules

To quantify the effect of the proposed Depthwise Partial Convolution (DPConv) and Decoder Path Unit (DPU), we ablated them on the COCO and MPII validation sets. Six variants are compared: ViTPose-S/B baselines, the lightweight SS2D/Mamba backbone, and the same backbone equipped with DPConv, with DPU, or with both modules; see Tables 3 and 4.

On COCO, the plain Mamba backbone already surpasses ViTPose-S with only 9.4 M parameters and 3.5 GFLOPs. Adding DPConv lifts the score to 75.3 AP/80.3 AR at minimal cost, while the full model (DPConv + DPU) attains 75.8 AP and 80.8 AR—matching ViTPose-B yet using roughly 20 % of its computational budget.

Table 3. Ablation study of DPConv and DPU on the **COCO** validation set

Model	**DPConv**	**DPU**	**AP**	**AR**	**Params (M)**	**GFLOPs**
ViTPose-S			73.8	79.2	22.0	5.3
ViTPose-B			75.8	81.1	80.0	17.1
Mamba			74.1	79.2	9.4	3.5
Mamba	√		75.3	80.3	13.2	5.4
Mamba		√	74.3	79.4	11.4	8.1
Mamba	√	√	75.8	80.8	15.1	10.0

Table 4. Ablation study of DPConv and DPU on the **MPII** validation set

Model	**DPConv**	**DPU**	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**PCK**
ViTPose-S			96.4	94.7	88.1	83.2	88.4	84.3	80.0	88.4
ViTPose-B			97.0	96.2	90.7	86.7	90.4	88.2	84.2	90.9
Mamba			96.8	94.9	86.4	80.0	87.2	81.9	77.1	86.9
Mamba	√		96.6	95.1	87.6	81.9	88.2	83.7	79.7	88.1
Mamba		√	96.9	95.1	87.2	80.7	87.8	83.2	78.6	87.6
Mamba	√	√	96.5	95.0	88.1	81.6	88.6	84.4	80.2	88.3

A similar pattern appears on MPII: DPConv boosts the base Mamba from 86.9 to 88.1 PCK, particularly improving elbow and ankle localisation. With both modules, the model achieves 88.3 PCK—again comparable to ViTPose-B yet at a fraction of its size and FLOPs.

These results demonstrate that the SS2D/Mamba backbone already offers a strong efficiency–accuracy balance, and the addition of DPConv and DPU further narrows the accuracy gap to heavier Transformer backbones while preserving its lightweight nature.

4.4 Experimental analysis of module-level ablation

To further verify the effectiveness of the proposed module in feature modeling and keypoint localization, we compare the pose heatmaps produced by the baseline Mamba and the enhanced DTMPose across different stages. The heatmaps visualize the model's response to keypoints in the input image, intuitively reflecting its spatial sensitivity, localization accuracy, and robustness to background noise.

As shown in Figure 5, we present the heatmap responses of Mamba and DTMPose across four stages (Stem, Stage 1–3): As shown in Figure 5, in the Stem phase, Mamba exhibits weak and diffuse heatmap responses, often affected by background noise. In contrast, DTMPose already produces more concentrated and focused activations at this early stage. As the network deepens, DTMPose progressively narrows the response range and sharpens the peak activations at keypoints, maintaining localization accuracy even under motion blur or occlusion (e.g., at the arms and feet). By Stage 3, DTMPose is able to highlight nearly all keypoints with high-confidence predictions while effectively suppressing back-

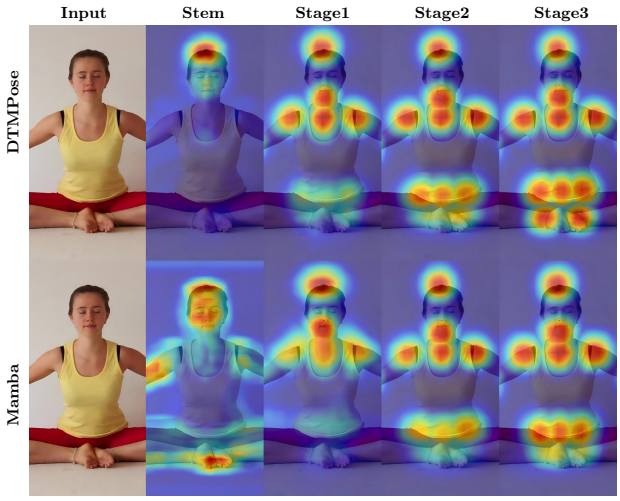

Fig. 5. Comparative heatmap visualizations of DTMPose and Mamba across different stages.

ground interference. In comparison, the baseline Mamba continues to struggle with blurry or drifting heatmap responses in several regions.

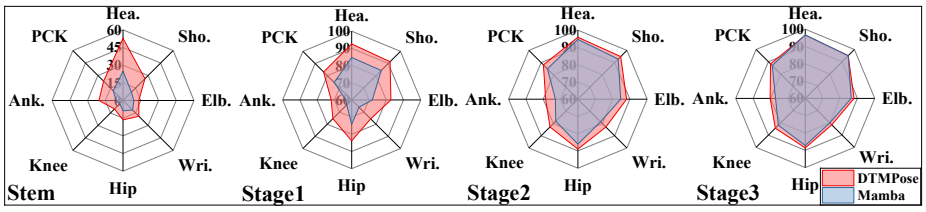

Fig. 6. Radar chart comparison of DTMPose and Mamba models at different stages on the MPII dataset, evaluated by PCK of major keypoints.

To assess the stage-wise effectiveness of DTMamba, we compare it with baseline Mamba on the MPII validation set (Figure 6, Table 5). DTMamba consistently outperforms Mamba across all stages, with notable gains at peripheral keypoints like wrists and ankles. At the Stem and Stage 1 levels, it improves PCK by +11.52/+12.97 and PCK@0.1 by +1.19/+8.26, respectively. These improvements are attributed to the DPConv and DPU modules, which enhance early-stage spatial perception and keypoint separability.

Overall, DTMamba delivers stronger multi-stage representations with higher accuracy and efficiency, validating its suitability for lightweight pose estimation.

Table 5. Stage-wise comparison of PCK and PCK@0.1 between DTMamba and Mamba

Model	Stem	Stage 1	Stage 2	Stage 3
PCK				
DTMamba(Ours)	23.20	74.51	85.18	88.26
Mamba	11.68	61.54	81.06	86.90
PCK@0.1				
DTMamba(Ours)	2.00	17.41	27.20	30.14
Mamba	0.81	9.15	21.94	27.53

4.5 Pose Visualization

(a) Visualization results on the COCO test-dev2017 dataset

(b) Visualization results on the MPII dataset

Fig. 7. Qualitative comparison of DTMPose on two datasets. (a) COCO dataset. (b) MPII dataset.

We present qualitative results of DTMPose on the COCO test-dev2017 and MPII datasets (Figure 7). The model demonstrates reliable keypoint localization under challenging conditions, including crowded scenes, occlusions, diverse actions, and varying environments.

On COCO, DTMPose maintains accurate structural predictions despite dense multi-person settings and background clutter. On MPII, it consistently localizes joints across a wide range of real-world motions. These examples highlight the model's robustness and generalization across diverse application scenarios.

5 Conclusions and outlook for the future

To address the challenges of high computational cost, limited local detail perception, and poor adaptability in 2D pose estimation, this paper proposed **DTM-Pose**—an efficient framework that integrates Mamba-based state-space modeling with sensory field-enhanced convolution. By replacing self-attention with SS2D and incorporating local enhancement modules such as DPConv, DTMPose achieves a strong balance between global dependency modeling and fine-grained feature extraction. Experimental results on the COCO dataset show that DTM-Pose achieves comparable or superior accuracy to Transformer-based baselines, with improved robustness and generalization in complex, high-density scenarios.

In the future, we plan to extend DTMPose to multi-frame or video-based pose estimation tasks, explore lightweight deployment on mobile devices, and further enhance the temporal modeling capabilities by combining dynamic state-space designs.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
2. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703 (2019)
3. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. Adv. Neural Inf. Process. Syst. **35**, 38571–38584 (2022)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Xu, R., Yang, S., Wang, Y., Cai, Y., Du, B., Chen, H.: Visual Mamba: A survey and new outlooks. arXiv preprint arXiv:2404.18861 (2024)
7. Ding, Y., et al.: I^2R-Net: Intra- and inter-human relation network for multi-person pose estimation. arXiv preprint arXiv:2206.10892 (2022)
8. Li, K., et al.: Pose recognition with cascade transformers. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1944–1953 (2021)
9. Li, Y., et al.: SimCC: A simple coordinate classification perspective for human pose estimation. In: European Conf. on Computer Vision (ECCV), pp. 89–106. Springer, Cham (2022)
10. Li, Y., et al.: TokenPose: Learning keypoint tokens for human pose estimation. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 11313–11322 (2021)
11. Lu, P., et al.: RTMO: Towards high-performance one-stage real-time multi-person pose estimation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1491–1500 (2024)
12. Maji, D., et al.: YOLO-Pose: Enhancing YOLO for multi-person pose estimation using object keypoint similarity loss. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2637–2646 (2022)

13. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proc. European Conf. on Computer Vision (ECCV), pp. 466–481 (2018)
14. Yang, S., Quan, Z., Nie, M., Yang, W.: TransPose: Keypoint localization via transformer. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 11802–11812 (2021)
15. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)
16. Zhu, L., et al.: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: Proc. 41st Int. Conf. on Machine Learning (ICML) (2024)
17. Liu, Y., et al.: VMamba: Visual state space model. Adv. Neural Inf. Process. Syst. **37**, 103031–103063 (2024)
18. Yang, J., et al.: Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In: Proc. AAAI Conf. on Artificial Intelligence **39**(9), 9202–9210 (2025)
19. Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014, LNCS **8693**, pp. 740–755. Springer, Cham (2014)
20. Andriluka, M., et al.: 2D human pose estimation: New benchmark and state of the art analysis. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3686–3693 (2014)
21. MMPose Contributors: OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020). Accessed 18 May 2025
22. Yang, J., et al.: Explicit box detection unifies end-to-end multi-person pose estimation. arXiv preprint arXiv:2302.01593 (2023)
23. Yu, C., et al.: Lite-HRNet: A lightweight high-resolution network. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10440–10450 (2021)
24. Xiong, Z., et al.: SwinPose: Swin transformer based human pose estimation. In: Proc. IEEE Int. Conf. on Multimedia Information Processing and Retrieval (MIPR), pp. 228–233 (2022)
25. Jiang, T., et al.: RTMPose: Real-time multi-person pose estimation based on MMPose. arXiv preprint arXiv:2303.07399 (2023)