MV-SSM: Multi-View State Space Modeling for 3D Human Pose Estimation

Supplementary Material

The supplementary includes additional model architecture and training details in Section 1) and Section 2) respectively. We include in-depth details of the datasets (Section 2.1), training schemes (Section 2.2), and various evaluation metric definitions (Section 2.3) used in the study. Additional details of the ablation studies are included in Section 3.

1. Model Architecture Details

The training images, sized 960×512, from the CMU-Panoptic dataset were fed into a ResNet-50 backbone pre-trained on the COCO dataset for 2D pose estimation task [5]. We used the same backbone weights as previous works [3, 4, 6] for a fair comparison. The model utilized 4 PSS/ decoder layers. The state space (SS2D + LN + FFN) blocks in the proposed PSS blocks had a 256-dimensional token size with a depth of 1 and 2 respectively. The token dimension is 256. The gate control was deprecated, and downsampling was set to 'none,' ensuring that the output shape of the state space blocks is the same as the input. It is important to note that the decoder layers did not share the parameters. Tokens with scores below $\epsilon = 0.1$ were filtered out at each decoder layer [6], followed by NMS to remove redundant tokens [3]. During initialization, the token number was approximately set to 1024, based on the motion capture space of the dataset.

2. Training and other details

2.1. Datasets

We outline the details of the datasets employed in this section. These include the CMU Panoptic [2], Shelf [1], and Campus [1] datasets. Note that only CMU Panoptic was used for training.

- **CMU Panoptic** [2] is a 3D multi-view dataset that contains multiple persons. The videos are collected in a spherical dome, i.e., an indoor scenario, where 480 RGB cameras and 10 RGBD cameras are installed. The Panoptic dataset contains over 30 videos and 65 sequences, including a variety of subjects wearing casual clothes, and performing a wide range of activities like dancing, playing musical instruments, eating, and so forth. The dataset is widely used in single and multi-view pose estimation tasks. Besides the 3D skeleton and point cloud labels, CMU Panoptic also provides facial landmarks, transcripts, and speaker ID, making it also suitable for whole-body or multi-modal tasks.
- **Campus** [1] dataset includes videos taken on a campus, for up to 3 subjects performing various actions, from 3 different views. The key points of subjects are annotated manually in the dataset.
- **Shelf** [1] dataset, as it is named, contains videos from 5 views of up to 4 subjects disassembling a shelf and interacting with each other. The dataset comes with manually labeled keypoints in all views.

Table 1. The details of the camera IDs, arrangements, and the camera numbers used in various experiments.

Cam Arrangements	Cam IDs	Numbers
CMU1	1, 2, 3, 4, 6, 7, 10	7
CMU2	12, 16, 18, 19, 22, 23, 30	7
CMU3	10, 12, 16, 18	4
CMU4	6, 7, 10, 12, 16, 18, 19, 22, 23, 30	10
CMU0	3, 6, 12, 13, 23	5
CMU0 w/ 2 Extra	3, 6, 12, 13, 23, 10, 16	7
CMU0(K)	First K cameras in CMU0 w/ 2	K

2.2. Training Schemes

MV-SSM was trained on eight NVIDIA TITAN RTX GPUs with a batch size of 1 for 40 epochs using a learning rate of `4e-4`. The training process required around 1.5 days. Early stopping was employed to prevent overfitting, and the backbone was kept frozen throughout training. For training on the CMU Panoptic dataset, the space size was set to $[8000, 8000, 2000]$, with the space center positioned at $[0.0, -500, 800]$. The initial cube size was set to $[80, 80, 20]$. For both the Campus and Shelf datasets, the space size remained constant, while the space centers was set at $[2000, 5000, 1000]$ for the Campus, and $[0, 500, 800]$ for the Shelf dataset.

2.3. Evaluation Metrics

In this section, we provide a detailed definition of the evaluation metrics used in this study.

- **MPJPE** or the Mean Per Joint Position Error is the average error in keypoint positions between the model's predictions and the corresponding ground truth. It is calculated as the Euclidean distance between the predicted keypoints and the ground truth keypoints. MPJPE is expressed in millimeters (mm) and is commonly used as the primary evaluation metric in 3D human pose estimation tasks.
- **AP and mAP** Average Precision (AP) is a widely used metric in various tasks, including classification and object detection. Average Precision (AP) is calculated as shown

in Equation 1,

$$AP = \sum_{k=0}^{k=n-1} [r(k) - r(k+1)] * p(k) \quad (1)$$

where $r(k)$ and $p(k)$ denote the recall and precision at the k^{th} sample, respectively. AP is used to summarize predictions in a binary manner, i.e., whether the predicted 3D keypoints have an MPJPE below a certain threshold or not, to measure the overall prediction's accuracy within a given margin. The threshold is indicated after AP, e.g. AP_{25} refers to the average precision value with an MPJPE threshold of 25mm. Mean Average Precision (mAP) is the mean value of AP across multiple thresholds. We follow the previous works [3, 6] and calculate mAP across MPJPE thresholds of $[25, 50, 75, 100, 125, 150]$ mm.

- **PCP** stands for the Percentage of Correct Parts. This metric measures the Euclidean distance between the predicted endpoints of each limb and the ground truth. If the average error between the two endpoints is less than half of the limb's length, that part of the body is considered correctly predicted. PCP is expressed as the percentage of correctly predicted parts out of the total number of parts.

3. Ablation Details

We provide an in-depth description of the modifications made to MV-SSM to perform the ablation experiments. Since the results of the ablations have already been discussed in detail in the main paper in Rows 1-6 of Table 4, we focus on an experiment-wise detailed description. The first four sets of experiments involve component-wise ablations, which help to test the how effective each of the proposed PSS block components were, while the remaining two discuss the branch-wise ablations by systematically excluding them from the model. We include both the component-wise and the branch-wise ablations to help establish the contribution of each proposed design improvement at both component and branch levels.

- **w Mean.** In the first ablation, we remove the PSS Block from MV-SSM and replace it with a simple Mean Block (Row 1). Note that we still keep the Projective attention in the first block. Therefore, instead of the PSS Block, the Mean is used to update the tokens by averaging the multi-view features. Note that when the PSS block is removed, the multi-scale features from the backbone are directly input in the projective attention, and the resulting token is input into the mean block (instead of the PSS block). Note that the model still learns the projective attention features from the first layer, however, when it fuses the multi-view information, it discards this information leading to a drop in performance. However, the mean operation acts as a very naive baseline, and we simply use it to confirm the importance of encoding spatial and relational information, which were explicitly modeled by the PSS block but were discarded by the introduced mean block.
- **w Cross-attention.** Since the mean is a very naive baseline, we replace the PSS Block with cross-attention (Row 2). Note that for a fair comparison, the same architectural setting was followed as the previous ablation, where the model still learns the projective attention tokens. In this way, when using cross-attention, the model retains the feature information over subsequent layers.
- **w/o Mamba (SS2D + LN + FFN).** In the third ablation, to study the contribution of state space modeling, we remove the SS2D + LN + FFN blocks from MV-SSM (See Row 3). In doing so, the PSS block degenerates into a simple Projective attention. Note that this significantly differs from the previous ablations since the multi-view feature fusion is still performed by the degenerate PSS block ('that consists of the Proj Attn'), while in the previous ablations, it was being performed by the 'mean' and 'cross-attention'.
- **w/o GTBS + Mamba.** In the fourth ablation, we remove the GT-Bidirectional Scan and the Mamba blocks. For this, we modified the appearance token to encode only the instance-level information and removed the Mamba blocks. In this way, only the instance-level tokens are scanned. Note that the multi-view feature fusion also degenerates to only fuse the instance level features (Row 4).
- **w/o PSS_$\mathbf{K}_n$_Branch.** In the branch-wise ablation, we remove the geometric token update branch (or simply the 3D keypoints branch) (Row 5) and replace it with a simple MLP.
- **w/o PSS_$\mathbf{V}_n$_Branch.** For removing the visual token update branch (or simply visual feature branch) (Row 6).

References

[1] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1669–1676, 2014. 1

[2] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 1

[3] Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L Waslander. Multiple view geometry transformers for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 708–717, 2024. 1, 2

[4] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *Computer Vision–ECCV 2020: 16th European*

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 197–212. Springer, 2020. 1

[5] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1

[6] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34:13153–13164, 2021. 1, 2