arXiv:2509.00649v1 [cs.CV] 31 Aug 2025

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

Aviral Chharia[1], Wenbo Gou[1], Haoye Dong[2,*]
[1]Carnegie Mellon University [2]National University of Singapore
{achharia, wgou}@andrew.cmu.edu, haoye@nus.edu.sg

Abstract

*While significant progress has been made in single-view 3D human pose estimation, multi-view 3D human pose estimation remains challenging, particularly in terms of generalizing to new camera configurations. Existing attention-based transformers often struggle to accurately model the spatial arrangement of keypoints, especially in occluded scenarios. Additionally, they tend to overfit specific camera arrangements and visual scenes from training data, resulting in substantial performance drops in new settings. In this study, we introduce a novel **M**ulti-**V**iew **S**tate **S**pace **M**odeling framework, named **MV-SSM**, for robustly estimating 3D human keypoints. We explicitly model the joint spatial sequence at two distinct levels: the feature level from multi-view images and the person keypoint level. We propose a Projective State Space (PSS) block to learn a generalized representation of joint spatial arrangements using state space modeling. Moreover, we modify Mamba's traditional scanning into an effective Grid Token-guided Bidirectional Scanning (GTBS), which is integral to the PSS block. Multiple experiments demonstrate that MV-SSM achieves strong generalization, outperforming state-of-the-art methods: $+\mathbf{10.8}$ on AP_{25} ($+24\%\uparrow$) on the challenging three-camera setting in CMU Panoptic, $+\mathbf{7.0}$ on AP_{25} ($+13\%\uparrow$) on varying camera arrangements, and $+\mathbf{15.3}$ PCP ($+38\%\uparrow$) on Campus A1 in cross-dataset evaluations. Project Website: https://aviralchharia.github.io/MV-SSM.*

1. Introduction

Estimating 3D human pose from multiple views is a crucial task in computer vision, with diverse applications in sports, industrial monitoring, and surveillance. However, it remains challenging due to its inherent complexity. Unlike single-view methods [1, 9, 14, 26, 31, 34, 39, 42], estimating 3D human pose from multi-view inputs involves several challenges, including—keypoint-level association to find cross-view correspondences while fusing multi-view information, multi-person association, and high occlusion.

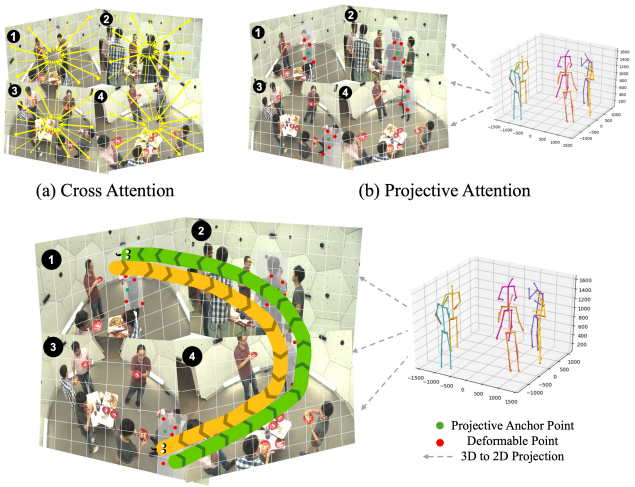


Figure 1. **Motivation.** Comparison of different token scanning methods. (a) Cross Attention acts on all image tokens. (b) Projective Attention [27, 44] obtains anchors with perspective projection and selectively attends to sample tokens surrounding the anchor points. (c) The proposed Grid Token-guided Bidirectional Scanning (GTBS) encodes the local context and the joint spatial sequence at the visual feature and person-keypoint levels.

Previous works [4, 6–8, 12, 13, 33, 45] have tried to address this task using a multi-stage approach. They first estimated the 2D keypoints in all views using off-the-self models such as OpenPose [5] and YOLOv10 [38]. Next, they associate these keypoints with their respective human subjects before finally triangulating the matched 2D keypoints to obtain 3D estimates. Consequently, recent works have focused on refining individual steps in these multi-stage frameworks. However, the multi-stage paradigm has limitations, as the matching algorithm is susceptible to complications in scenes with multiple humans, and errors tend to accumulate at each stage of the pipeline. To address this, a few recent works [20, 27, 35, 36, 43, 44] have focused on end-to-end supervised learning approaches. These models bypass the explicit geometry modeling required in multi-stage models. However, they suffer from a lack of gener-

*Corresponding author

alizability to new scenes and camera arrangements due to overfitting to training camera setups.

To tackle this challenge, we introduce *MV-SSM*, a novel Multi-View State Space Modeling framework. *MV-SSM* is the *first* framework that integrates state space models [15] into 3D multi-view geometry modeling for 3D human pose estimation. By explicitly modeling the joint spatial sequences at both the feature level from multi-view images and the person keypoint level, *MV-SSM* harnesses the sequence-modeling strengths of state space models to accurately predict 3D human keypoints, even in scenarios with significant occlusions.

MV-SSM is designed as an end-to-end differentiable model that progressively refines 3D keypoint predictions using state space modeling blocks. It first extracts multi-scale features from multi-view inputs using a ResNet-50 [19] backbone. These features are then processed through a carefully designed Projective State Space (PSS) block, which models the spatial relationships between keypoints across different views. The PSS block utilizes projective attention to effectively fuse multi-view information and incorporates SSMs to learn the joint spatial sequence. Through this, *MV-SSM* can handle occlusions and improve the accuracy of 3D keypoint estimation. *MV-SSM* further enhances its predictions by progressively refining the keypoint estimates through successive layers of PSS blocks, which update the tokens and adjust the keypoint positions with residual offsets. Finally, *MV-SSM* predicts the 3D human poses via differentiable algebraic triangulation, ensuring robust and accurate estimation. We summarize our contributions as:

- The present study is the *first* to adapt visual mamba for the 3D multi-view human pose estimation task. It presents *MV-SSM*, a novel **M**ulti-**V**iew Geometry **S**tate **S**pace **M**odel, designed to improve the robustness and generalizability of localized 3D skeleton keypoints.
- Our Projective State Space (PSS) block integrates state space modeling and projection attention to effectively capture joint spatial sequences and fuse multi-view information.
- We introduce Grid-Token guided Bidirectional Scanning (GTBS) to enhance performance further and validate its effectiveness through ablations.
- Experiments demonstrate that our *MV-SSM* significantly outperforms SOTA methods on in-domain 3D keypoints estimation and multiple generalizability tests, validating its adaptability to new camera settings and scenarios.

2. Related works

2.1. 3D Human Pose Estimation

Single-view Approaches. Estimating 3D human pose from a single image is inherently ill-posed due to challenges of occlusion and depth ambiguity. Current approaches primarily apply two paradigms: *(i)* 2D-to-3D lifting architectures, and *(ii)* direct regression frameworks. Lifting-based schemes [1, 9, 14, 26, 31, 34, 39, 42] first detect robust 2D keypoints, and then lifts them into corresponding 3D space by depth estimation through geometric constraints and learned priors. Recent advances exploited 3D CNNs and transformers [37], to enable end-to-end 3D regression directly from the image. Li *et al*. [24] introduced dual-task learning combining keypoint detection and regression. Volumetric approaches like [32] employed stacked ConvNets to estimate the likelihoods of keypoints. Few recent methods [30] integrated GNNs [10] with attention mechanisms to constraint limb length priors and contextual relationships.

Multi-view Approaches. Multi-view methods offer relatively accurate pose estimation, as occluded keypoints in one view are often visible in others. However, developing generalizable learning-based models that effectively utilize multi-view inputs remains a significant challenge. Given multiple camera positions around a scene, 3D keypoints can be estimated via triangulation once 2D keypoints are detected in each view. However, a key challenge in this approach is matching subjects across views. Traditional triangulation-based methods, such as one by Dong *et al.* [12] achieve high precision with robust pose estimates but suffer from cross-view matching fragility and non-differentiable pipelines. Learning-based alternatives address these issues by integrating differential geometry [20]. One of the first works by Iskakov *et al.* [20] introduced volumetric feature unprojection using 3D CNNs for keypoint position estimation. Another widely adopted method, VoxelPose [36], directly aggregated multi-view features in 3D voxel space. Recent approaches, such as MvP [44] and MVGFormer [27], leveraged hierarchical querying mechanisms with attention-based fusion, demonstrating improved robustness to occlusions. MVGFormer [27] introduced a queue to filter out improbable subject candidates and iteratively refine keypoint locations. Due to the additional visual information from multi-view inputs, these methods achieve higher accuracy compared to single-view alternatives.

2.2. State Space Models

State Space Models (SSMs) have evolved from the linear dynamic systems used in Kalman filtering [22] to advanced deep learning implementations. Gu *et al.* introduced Structured State Space Sequence (S4) models [16, 17] which achieved SOTA performance in capturing long-range dependencies. Mamba [15] further extended the S4 model by replacing the fixed hidden space projection matrices with a selective projection matrix, dynamically adjusted to match the input sequence. This enhancement demonstrated promising capabilities for processing long-range text sequences. Emerging works have adapted Mamba for visual tasks. Yang *et al.* [47] and Liu *et al.* [29] modified Mamba's

scanning mechanism for vision tasks by replacing the original forward scan with bidirectional horizontal scanning of image patches to better capture spatial relationships. Further improvements, such as VMamba [29], introduced enhanced vertical scans, enabling a cross-scan mechanism. Mamba-ND [25] expanded Mamba to video understanding by scanning video frames across height, width, and temporal dimensions from both directions. VideoMamba [23] followed a vision Transformer (ViT)-inspired design, demonstrating the scalability of SSM models for video processing. Additionally, hybrid Mamba-Transformer approaches have emerged [18], achieving SOTA performance and high image throughput on ImageNet-1K [11] classification. Our proposed MV-SSM is the first to leverage SSMs to the multi-view 3D human pose estimation task, integrating their strengths to improve accuracy and robustness.

3. Proposed Methodology

This section first introduces the preliminaries on state space models, followed by an in-depth description of the proposed model architecture, which includes the Projective State Space block and the Grid Token-guided Bidirectional Scanning.

3.1. Preliminaries

SSMs. Recently, Structured State Space Sequence (S4) models [16] have demonstrated high potential in handling long sequential contexts. State space models transform an input vector $x(t) \in \mathbb{R}$ into an output sequence $y(t) \in \mathbb{R}$ using latent state $h(t) \in \mathbb{R}^N$. This process follows linear differential equations (as shown in Eq. 1), where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1\times N}$ and $D \in \mathbb{R}^1$ form the weight parameters.

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t) + Dx(t), \end{aligned} \tag{1}$$

$$\begin{aligned} h_t &= \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t, \\ y_t &= \boldsymbol{C}h(t). \end{aligned} \tag{2}$$

The zero-order hold (ZOH) rule is often used to discretize these continuous systems (Eq. 3).

$$\begin{aligned} \overline{\boldsymbol{A}} &= \exp(\Delta\boldsymbol{A}), \\ \overline{\boldsymbol{B}} &= (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}. \end{aligned} \tag{3}$$

Here, Δ stands for the discrete step size. The S4 model is a Linear Time Invariant (LTI) system, as the weighting parameters and discretization rules remain fixed over time step t. The S6 models [15] further extend S4 models by adding selective scan, which dynamically adapts the projection matrices along the time dimension of the input sequence. This enables Mamba to project the input into the latent space at each step by remembering and discarding different tokens inside to sequence, compared to the fixed projection matrix of the S4 models.

SSMs for Visual tasks. Mamba [15] was initially designed for 1D data such as text or audio, which makes its direct application to visual tasks challenging since images contain both local context and global spatial information. To adapt Mamba for visual information, most works first chunk the 2D/ 3D input into small patches and encode them as vectors. Vim [47] designed a bidirectional horizontal scan, which goes through the image patches column by column from two directions. VMamba [29] proposed an SS2D block with additional vertical scans and introduced it into the VSS block (See the Figure 3 (a-b)).

SSMs for Joint Spatial Sequence. It is important to note that MV-SSM differs significantly from video-based Mamba, which focuses on learning temporal relationships from consecutive frame sequences. Instead, we model the joint spatial sequence within a series of static multi-view frames. By utilizing Mamba to capture intrinsic keypoint relations in a multi-view context, our method markedly enhances 3D pose estimation accuracy. Specifically in Eq. 1, $x(t)$ corresponds to the j^{th} joint token of the joint spatial sequence $\mathbf{E}_n$, and our PSS block scans along joint dimension rather than the original time dimension. $y(t) = \text{PSS}(x(t))$ corresponds to the updated token from the PSS block.

3.2. MV-SSM

Problem Formulation. Given a set of multi-view images $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ for $t = 1, ..., T$, captured by T cameras, the goal is to estimate the 3D keypoints for each human present in the captured scene. We formulate this problem as a progressive regression task, where our model *MV-SSM* progressively estimates and refines its predictions through PSS blocks, ultimately achieving accurate 3D human pose based on multi-view images.

Model Architecture. As shown in Figure 2, *MV-SSM* first passes the multi-view images into a ResNet-50 backbone [19] to extract multi-scale features $\{\mathbf{F}_t\}_{t=1}^{T}$. The extracted features are subsequently processed by our proposed Projective State Space (PSS) block, which leverages the joint spatial sequence by capturing their relationships within state space through GTBS. Finally, for estimating the 3D human keypoints, it employs a learnable triangulation adopted from [20]. We additionally process keypoint-level information through stacked PSS blocks, with each layer outputting joint offsets to iteratively refine the predictions from the previous layer. The final model consists of multiple PSS blocks stacked layer-wise, which enables *MV-SSM* to learn a progressive regression yielding accurate 3D human pose estimation.

Joint Token Representation. Following prior works [27,

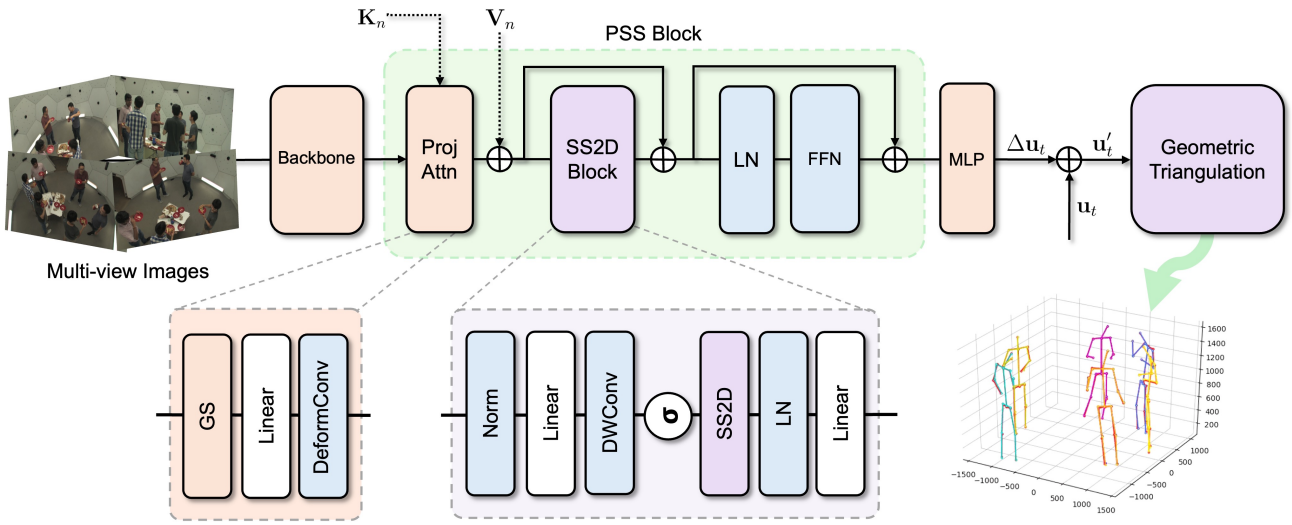


Figure 2. The architecture of the proposed Multi-view State Space Model (*MV-SSM*). It processes multi-view images through a ResNet-50 backbone to extract multi-scale features, which are refined by stacked Projective State Space (PSS) Blocks. These blocks leverage projective attention and state space modeling to progressively refine the keypoints, with final 3D keypoints estimated via geometric triangulation.

44], we use a joint token to represent each human in the scene. Each joint token is denoted as $\{\mathbf{E}_n\}_{n=1}^{N}$, where N is the initial set of humans, and represented as a combination of its visual feature term $\mathbf{V}_n \in \mathbb{R}^{J\times L}$, and its geometry term $\mathbf{K}_n \in \mathbb{R}^{J\times 3}$, where J stands for the number of keypoints and L denotes the feature dimension. These two terms form the joint token denoted by $\mathbf{E}_n = (\mathbf{V}_n, \mathbf{K}_n)$. This representation is beneficial as it explicitly encodes both visual and geometric information.

While learning the visual feature term, we model two sub-tokens, i.e., person-level $\{\mathbf{s}_n\}_{n=1}^{N} \subset \mathbb{R}^L$ and keypoint-level $\{\mathbf{g}_j\}_{j=1}^{J} \subset \mathbb{R}^L$ sub-tokens. The visual feature for the j^{th} keypoint of the n^{th} token is represented by $\mathbf{v}_n^j = \mathbf{s}_n + \mathbf{g}_j$. We represent the combined visual feature term of all keypoints as $\mathbf{V}_n = [\mathbf{v}_n^1, \mathbf{v}_n^2, ..., \mathbf{v}_n^J]^\intercal$. The geometry term directly stores each keypoint position, i.e., $\mathbf{k}_n^j = [\mathbf{x}_n^j, \mathbf{y}_n^j, \mathbf{z}_n^j]^\intercal$, where keypoint j of the n^{th} token is stored. The geometry terms of keypoints' the n^{th} token are denoted as $\mathbf{K}_n = [\mathbf{k}_n^1, \mathbf{k}_n^2, ..., \mathbf{k}_n^J]^\intercal$. We follow the hierarchical token scheme [44] to reduce the state search space, thus resulting in fewer learnable parameters.

Projective State Space (PSS) Block. A simple cross-attention on extracted tokens would extensively interact with the features across all spatial locations from multiple views. However, this leads to high computation costs and inefficiency due to multi-view features being involved. Comparatively, the projective attention [44] projects keypoints to anchor points onto different views and sample deformable points surrounding them to aggregate the local context. Thus, it encodes the corresponding 2D spatial positions across views and the local context extracted from sampling deformable points surrounding the anchors. This enables an efficient fusion of multi-view visual information and is adopted by several studies [27, 44]. However, projective attention alone may not fully exploit inter-joint relationships from the joint spatial sequence $\mathbf{E}_n$. The PSS block addresses this by considering features from both projective attention and state-space modeling.

The PSS block operates on multi-scale tokens and focuses on learning the joint spatial sequence and inherent inter-joint relations. The architecture of the PSS block is shown in Figure 2. Each block is composed of an SS2D block, a projective attention, followed by a normalization, a feed-forward neural network, with a residual connection. The features in each view obtained from the projective attention are passed forward. *MV-SSM* contains M stacked PSS blocks. The PSS block not only leverages projection attention but also considers features from state space modeling. This enables *MV-SSM* to enhance performance with a few crucial tokens by learning effective features, shown in our ablation study 4.2. We further highlight the architectural difference between the Mamba [15], VSS [29], and our proposed PSS block in Figure 3.

As shown in Figure 2, we feed the visual feature term through the PSS block for feature update. For the subsequent layers, the PSS blocks update the tokens progressively and refine the prediction with offsets. The feature $\mathbf{F}_t$ is obtained from the sampling 2D positions $\mathbf{u}_t$. The MLP then learns the 2D residual offsets $\Delta\mathbf{u}_t$. The MLP also produces a confidence score $\mathbf{c}_t$. We obtain the new 3D position $\mathbf{k}'$ using differentiable algebraic triangulation [20], formu-

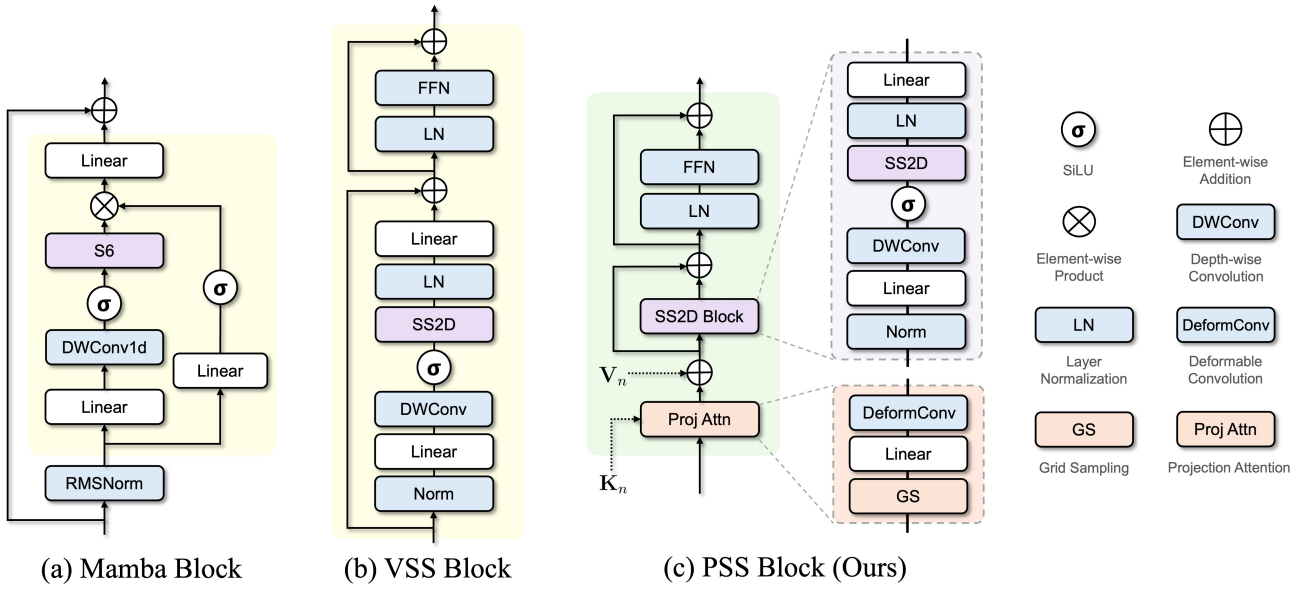


Figure 3. The architecture of Mamba block [15], VSS block [29] and the proposed Projective State Space block. The PSS block captures joint spatial relationships through projective attention and state-space modeling, progressively refining results.

lated as follows:

$$\mathbf{k}' = \text{AlgTriangulation}(\mathbf{u}'_t, \mathbf{c}_t, \mathbf{\Pi}_t), \tag{4}$$

where $\mathbf{u}'_t$ is the 2D positions in T views, $\mathbf{c}_t$ is corresponding confidence scores, and $\mathbf{\Pi}_t$ is the projection matrix for view t. As the model is end-to-end differentiable, the estimated confidence helps reduce 3D pose errors. Following [44], the 2D position is refined by adding the residual to the projection, with $\Delta\mathbf{u}_t$ supervised by projected ground-truth 2D poses. This process is formulated as shown below:

$$\mathbf{u}'_t = \mathbf{u}_t + \Delta\mathbf{u}_t. \tag{5}$$

The updated token can better encode the joint spatial sequence of the corresponding humans in the scene. Passing the feature through the PSS block further improves 3D pose estimation, as the 2D pose prediction benefits from re-projecting the triangulated 3D pose, allowing it to extract more relevant features. For successive layers, the geometry term is directly updated with the estimated 3D keypoints.

Grid Token-guided Bidirectional Scanning (GTBS). To achieve robust 3D keypoint estimation, we design the Grid Token-guided Bi-directional Scanning for multi-view human 3D pose estimation. GTBS acts on the sampled projection tokens from each view. As the layers progressively refine keypoints, it scans more relevant features, boosting performance. A naive approach scans all image patches from multi-views (as shown in (a) of Figure 1), but it is computationally expensive due to redundant background tokens and ignores the spatial relationships of body keypoints. To address this, we perform token-level bidirectional scanning on sampled features instead of scanning all tokens. We adapt SS2D [29] for grid token-guided bidirectional scanning, tailoring it to the joint spatial sequence.

Following [27], we initialize the visual feature term from a standard normal distribution. For the geometry term, rather than relying on 3D predictions from a pre-trained model, we sample N human centers on the ground plane and initialize the keypoints as a conventional T-pose. Since the initial number of assumed tokens is kept large, we train a simple linear layer to predict a score for each token based on the visual feature term. This layer maps an L-dimensional feature to two scores, which are fed via a sigmoid function to output the positive and negative probabilities, respectively. Overall score is the average of scores across all keypoints, filtering tokens whose scores are smaller than ϵ.

We train MV-SSM with a combined loss: a pose loss that supervises both the 2D/ 3D keypoints from positive tokens, and a cross-entropy loss for the MLP classifier distinguishing positive and negative tokens. This loss is applied at each layer of the progressive regression. For token-level ground truth assignment, we use the anchor-based matching strategy [46]. Following [27], we assign ground-truth poses to the nearest token as "positive," while unassigned tokens are considered "negative." We denote initial geometry term as $\{\mathbf{K}_n^{(0)}\}_{n=1}^{N}$, serves as the anchor. The ground-truth keypoints for Z humans are $\{\mathbf{H}_z\}_{z=1}^{Z} \subset \mathbb{R}^{J\times 3}$. The W nearest tokens $\{\mathbf{K}_{z(w)}\}_{w=1}^{W}$ are assigned to each $\mathbf{K}_z$ based on their

geometry term. The loss function is formulated as follows:

$$\mathcal{L}_{\text{pose}} = \sum_{w=1}^{W} \left(\mathcal{L}_1\left(\mathbf{K}_{z(w)}, \mathbf{H}_z\right) + \sum_{t=1}^{T} \mathcal{L}_1\left(\hat{\mathbf{U}}_{z(w),t}, \mathbf{U}_{z,t}\right) \right), \tag{6}$$

where we denote $\hat{\mathbf{U}}_{z(w),t}$ as the estimated 2D pose of the $z(w)^{\text{th}}$ token in the t^{th} view, and $\mathbf{U}_{z,t}$ as the ground truth.

4. Experiments and Results

Datasets. We train MV-SSM on the popular CMU Panoptic [21] dataset. To facilitate a direct comparison, same experimental settings as prior works [27, 36, 44] was used in our experiments. For generalizability evaluation, we employ two additional well-established benchmarks, Campus [3] and Shelf [3]. We do not fine-tune on these datasets, which enables us to assess realistic cross-dataset generalization performance.

Evaluation Metrics. Average Precision (AP) and Recall, along with Mean Per Joint Position Error (MPJPE) is used as the evaluation metric. For Campus and Shelf benchmarks, the Percentage of Correct Parts (PCP) is reported. The camera pose is assumed to be known and utilized, following common practice in the literature for this task. All baselines assume the camera poses are known.

Implementation Details. For extracting image features from multi-view RGB input, ResNet-50 backbone pretrained on the COCO dataset was used as in previous works [27, 36, 44]. The learning rate was $4e^{-4}$ and the model was trained for 40 epochs with early stopping. Additionally, a confidence threshold of $\epsilon = 0.1$ is applied layer-wise to filter redundant predictions [44], followed by Non-Maximum Suppression (NMS) to refine the process further [27]. For additional implementation details, refer to the supplementary materials.

4.1. Main Results

3D Keypoints Estimation Evaluation. First, we evaluate MV-SSM's robustness in predicting 3D keypoints on the CMU Panoptic [21] benchmark. The quantitative comparison with SOTAs is shown in Table 1, demonstrating that MV-SSM significantly outperforms existing models. We use the same camera configurations and train-test splits as VoxelPose [36], MvP [44], and MVGFormer [27] for a fair comparison. MV-SSM surpasses other learning-based methods like VoxelPose by a large margin ($+\mathbf{9.5}\%$) and further outperforms MvP and MVGFormer ($+\mathbf{1.2}\%$). MvP tends to overfit training camera views and arrangements, thus lacking generalizability. We later demonstrate that MV-SSM surpasses both MvP and MVGFormer by a significant margin on generalizability. Qualitative visual comparisons with current SOTA [27] are shown in Figure 4.

Table 1. **In-domain Results** on CMU Panoptic with conventional train-test split. '-' indicates information not reported by model.

Method	Venue	CMU Panoptic	
		AP_{25} ↑	MPJPE (mm) ↓
VoxelPose [36]	ECCV 20	84.0	17.7
Lin *et al.* [28]	CVPR 21	92.1	16.8
Graph3D [40]	ICCV 21	-	15.8
MvP [44]	NeurIPS 21	92.3	15.8
Faster VoxelPose [43]	ECCV 22	85.2	18.3
MVGFormer [27]	CVPR 24	92.3	16.0
MV-SSM	**Ours**	**93.5**	**15.7**

Cross-dataset Generalizability Evaluation. Next, we conduct a cross-scene evaluation by testing MV-SSM model trained on CMU-Panoptic directly on the Campus and Shelf benchmarks, without any additional fine-tuning. The comparative results are reported in Table 2. As to Campus, MV-SSM demonstrates significant improvement compared to MVGFormer ($+\mathbf{15.3}$ on PCP for Actor 1 and $+\mathbf{9.2}$ PCP Average improvement), showcasing superior generalization capabilities. On the Shelf dataset, performance scores are almost saturated, and it is not surprising that the margin of improvement for our method is smaller. The lower PCP scores on Campus can be attributed to its more challenging nature–larger field-of-view and outdoor setting–compared to Shelf's controlled indoor environment. Note that geometry-based models generally achieve higher raw PCP scores than learning-based models. However, they struggle in complex occlusion scenarios due to reliance on 2D keypoints while neglecting feature-level visual information, leading to failures in matching humans across views. We reported the results without fine-tuning because fine-tuning does not provide a clear assessment of a model's transferability to new datasets.

Cross-camera Generalizability Evaluation. Following the protocol in [2], we evaluate MV-SSM's generalizability across varying numbers of cameras. When training the model we used the CMU0 camera setup with 5 cameras (camera IDs: 3, 6, 12, 13, 23). During testing, we systematically vary the number of cameras by adding or removing them from the CMU0 test set, following previous paradigms [27]. This configuration is denoted as "CMU0(K)," where K represents the number of cameras used during inference, as reported in Table 3. We notice a consistent trend: as the number of cameras decreases to 4 and then 3, AP_{25} and mAP scores drop. This performance drop is due to the loss of multi-view visual information from the removed cameras. Dong *et al.* [12] fails to generate accurate 3D human keypoints, likely due to challenges in geometry-based cross-view matching. VoxelPose [36] exhibits a performance drop as the number of cameras increases, indicating overfitting to specific training camera

Table 2. **Cross-scene Generalization Evaluation.** All models were trained on CMU0 configuration of the CMU Panoptic benchmark [21] and evaluated on Campus and Shelf [3] w/o fine-tuning.

Method	Venue	Campus (PCP↑)				Shelf (PCP↑)			
		A1	A2	A3	Average	A1	A2	A3	Average
Dong *et al.* [12]	CVPR 19	97.6	93.3	98.0	96.3	98.8	94.1	97.8	96.9
MvP [44]	NeurIPS 21	0.0	0.0	0.0	0.0	5.7	16.0	4.6	8.7
VoxelPose [36]	ECCV 20	16.1	12.4	5.1	11.2	77.9	63.2	68.1	69.8
MVGFormer [27]	CVPR 24	40.2	61.0	73.1	58.1	89.9	**85.4**	88.6	87.9
MV-SSM	**Ours**	**55.5**	**65.5**	**79.9**	**67.3**	**89.9**	85.2	**88.9**	**88.0**

Table 3. **Cross-camera and Cross-arrangements Generalization Evaluation**. Cross-camera inference dataset CMU0 (K) indicates a total of K cameras. In cross-arrangements, dataset CMU1-4 has varying numbers of cameras and camera IDs compared with CMU0. The CMU0 configuration with 5 cameras is used for training all models, and results are reported w/o fine-tuning. Average is estimated over CMU0 (K) for K=1 to 7.

	Method	Venue	CMU0 (3)		CMU0 (4)		CMU0 (6)		CMU0 (7)		Average	
			AP_{25} ↑	mAP↑	AP_{25}	mAP↑	AP_{25} ↑	mAP↑	AP_{25} ↑	mAP↑	AP_{25} ↑	mAP↑
Cross-Camera	Dong *et al.* [12]	CVPR 19	0.0	37.7	0.0	46.6	0.1	41.8	0.2	40.1	13.3	41.6
	VoxelPose [36]	ECCV 20	13.7	73.0	65.4	93.1	84.5	96.8	81.5	96.3	69.1	89.8
	MvP [44]	NeurIPS 21	12.3	57.1	59.8	84.1	0.0	0.3	0.0	0.2	21.6	35.4
	MVGFormer [27]	CVPR 24	44.6	83.4	83.6	96.0	**94.7**	98.5	95.1	98.6	83.3	94.1
	MV-SSM	**Ours**	**55.4**	**90.3**	**84.9**	**97.0**	94.6	**98.8**	**95.6**	**98.8**	**84.8**	**96.7**
	Method	Venue	CMU1		CMU2		CMU3		CMU4		Average	
Cross-arrange	Dong *et al.* [12]	CVPR 19	0.1	39.9	0.0	29.2	0.0	29.2	0.0	32.8	0.0	32.8
	VoxelPose [36]	ECCV 20	28.8	86.9	26.9	69.6	21.0	69.6	40.9	87.9	29.4	78.5
	MvP [44]	NeurIPS 21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MVGFormer [27]	CVPR 24	**86.8**	96.1	67.9	90.3	52.5	78.8	91.5	97.2	74.7	90.6
	MV-SSM	**Ours**	86.0	**97.0**	**74.6**	**93.5**	**59.5**	**84.8**	**94.2**	**98.5**	**78.6**	**93.5**

configurations. Similarly, MvP [44] shows poor generalizability, as it overfits training cameras and fails on unseen camera numbers. Compared to MVGFormer [27], our proposed model shows substantial improvement (+10.8 on AP_{25} and +6.9 on mAP) in the challenging scenario when just using 3 cameras. This suggests superior generalizability compared to MVGFormer. Furthermore, as the number of camera increases to 6 and 7, MV-SSM continues to surpass MVGFormer [27], demonstrating its ability to effectively leverage additional visual information even for unseen camera views. In contrast, both MvP and VoxelPose demonstrate a performance drop as more cameras are added: MvP due to overfitting to specific camera arrangements during training, and VoxelPose due to not fully exploiting the visual information.

Cross-arrangement Generalizability Evaluation. We further increase the difficulty by conducting a cross-arrangement generalizability evaluation. In this setting, we change the camera arrangement, using different numbers of cameras and different camera IDs compared to the training set. The results are summarized in the lower half of Table 3. The results demonstrate that MV-SSM consistently outperforms SOTA models, including the transformer-based MVGFormer. On the challenging CMU3, it achieves a high improvement of +7.0 on AP_{25} and +6.0 on mAP compared to the current SOTA. Additionally, it significantly outperforms MvP, VoxelPose, and Dong *et al.* [12] thereby offering higher practical value in real-world scenarios where camera setups vary drastically. Its ability to generalize effectively across different camera arrangements w/o fine-tuning highlights its robustness and adaptability.

4.2. Ablation Studies

Effect of Proposed Modules. The PSS block is a key contribution to our work, and we evaluate the effectiveness of each module within it. The results are presented in Table 4. PSS learns the joint spatial sequence and inter-joint relations using the state space model. In the first experiment, we replace the PSS block with a simple Mean block (Row 1), which is used to update the tokens by combining multi-view features. Since the mean operation will discard information, it is not surprising that it has a very low AP_{25} and a high MPJPE. This result highlights the importance of encoding spatial and relational information. Next, we replace the PSS block with cross-attention (Row 2). The cross-attention alone cannot fully capture the complex dependencies modeled by our PSS block. A common objection is that the input features and Projective attention alone suffice for multi-view 3D pose estimation, with little bene-

Table 4. **Ablation study** to verify proposed components. Results are reported on CMU Panoptic [21]. Without is abbreviated as 'w/o' which refers to the removed component / branch.

	Ablation	AP_{25} ↑	AP_{100} ↑	MPJPE ↓
Module-wise Ablations				
1	w Mean	36.2	54.0	71.8
2	w Cross-attention	90.4	99.1	16.8
3	w/o Mamba (SS2D+LN+FFN)	92.3	99.3	16.0
4	w/o GTBS + Mamba	87.7	98.8	18.6
Branch-wise Ablations				
5	w/o PSS_$\mathbf{K}_n$_Branch	91.8	99.7	15.7
6	w/o PSS_$\mathbf{V}_n$_Branch	92.5	99.6	15.9
MV-SSM (Full)		**93.5**	**99.8**	**15.7**

fit from the Mamba Blocks. To address this, we remove the Mamba (SS2D + LN + FFN) block from the model (Row 3), and the PSS block degenerates into a simple Projective attention. This results in a noticeable performance drop compared to the full model, confirming that the Mamba block is essential for capturing critical dependencies within PSS. A larger performance drop occurs when the proposed GTBS and Mamba blocks are removed (Row 4). Specifically, the AP_{25} (↑) drops to $93.5 \rightarrow 87.7$ while the MPJPE (↓) error increases from $15.7 \rightarrow 18.6$. This demonstrates that both components are integral to achieving optimal performance. Table 4 confirms the effectiveness of proposed modules.

Effect of Branch-wise Features. As reported in Table 4, we validate each branch feature's contribution by systematically excluding them. First, we remove the $\mathbf{K}_n$ branch (Row 5) and replace it with a simple MLP. This results in a significant performance drop, with AP_{25} dropping from $93.5 \rightarrow 91.8$ and AP_{100} decreasing from $99.8 \rightarrow 99.7$. This highlights the importance of explicitly modeling 3D keypoints to learn local context. Removing the $\mathbf{V}_n$ branch (Row 6) causes a performance drop, though less important than without the 3D keypoints branch. This is because the 3D joints branch specifically learns strong local context. Here, AP_{25} (↑) value drops from $93.5 \rightarrow 92.5$ while the MPJPE (↓) error increases from $15.7 \rightarrow 15.9$.

5. Conclusion and Future Work

The present work proposes *MV-SSM*, a Multi-view State Space Model that enhances 3D human pose estimation by leveraging joint spatial sequences from multi-view inputs. *MV-SSM* uses Projective State Space blocks to model joint spatial sequences and inter-joint relationships, achieving robust pose estimates. We introduce the progressive regression with Grid Token-guided Bidirectional Scan, enhancing both in- and out-of-domain performance and demonstrating strong generalizability to unseen cameras and geometries. Extensive experiments demonstrate that *MV-SSM* outperforms previous state-of-the-art methods in generalizability,

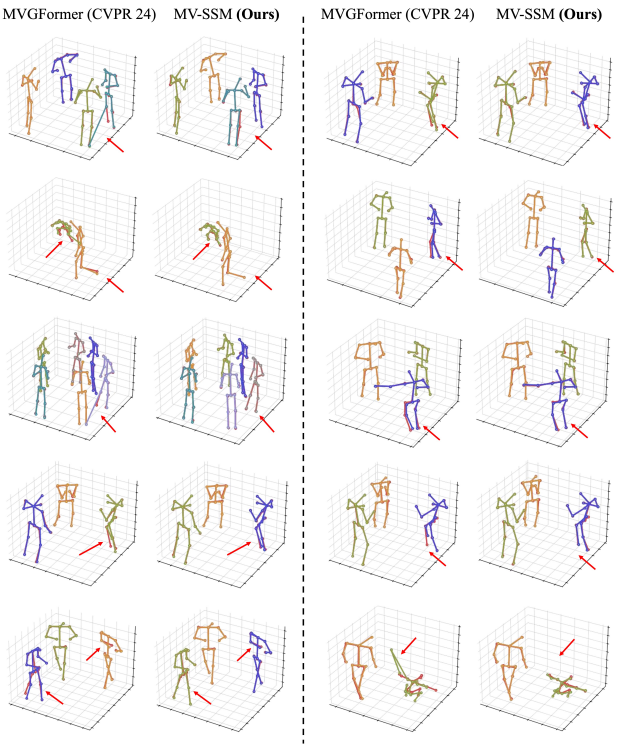


Figure 4. **Qualitative Comparisons.** We present a visual comparison against MVGFormer [27] on CMU Panoptic [21] benchmark. The Ground truth human poses are shown in "red" and the predicted pose is overlapped on it to show an accurate comparison. MV-SSM achieves accurate poses, especially in difficult scenarios, demonstrating superior performance. As illustrated in the first row, MV-SSM is better able to predict, for example, the person's left foot. Note that the colors of persons are different since we do not perform ID-matching.

advancing multi-view 3D keypoint estimation for applications in AR/VR, animation, surveillance, etc.

Like previous models, *MV-SSM* fails when a human is visible in only one camera, leading to missing keypoints from other views, especially when fewer cameras are involved. Future work could focus on extending the model to perform better in the single-view scenarios when the person is hidden from other views. Another potential avenue for improvement is enhancing robustness with generative models that can synthesize the missing keypoints or estimate the occluded keypoints, rather than randomly extracting them from a normal distribution. Future studies could also explore semi-supervised and self-supervised training strategies to leverage unlabeled multi-view images for augmenting the training datasets.

Acknowledgments. Aviral Chharia was supported in part by the ATK-Nick G. Vlahakis Graduate Fellowship from Carnegie Mellon University, USA.

References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015. 1, 2

[2] Kristijan Bartol, David Bojanić, Tomislav Petković, and Tomislav Pribanić. Generalizable human pose triangulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11028–11037, 2022. 6

[3] 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. 6, 7, 11

[4] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 1

[6] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 541–557. Springer, 2020. 1

[7] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3279–3288, 2020.

[8] Hau Chu, Jia-Hong Lee, Yao-Chih Lee, Ching-Hsien Hsu, Jia-Da Li, and Chu-Song Chen. Part-aware measurement for robust multi-view multi-human 3d pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1481, 2021. 1

[9] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European conference on computer vision (ECCV)*, pages 668–683, 2018. 1, 2

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801, 2019. 1, 2, 6, 7

[13] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6981–6992, 2021. 1

[14] Xiaochuan Fan, Kang Zheng, Youjie Zhou, and Song Wang. Pose locality constrained representation for 3d human pose reconstruction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 174–188. Springer, 2014. 1, 2

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. 2023. 2, 3, 4, 5

[16] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. 2021. 2, 3

[17] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 2

[18] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone, 2024. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[20] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 1, 2, 3, 4

[21] 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. 6, 7, 8, 11

[22] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 2

[23] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2025. 3

[24] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12*, pages 332–347. Springer, 2015. 2

[25] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pages 75–92. Springer, 2025. 3

[26] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d

human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, 2022. 1, 2

[27] 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, 3, 4, 5, 6, 7, 8, 11, 12

[28] Jiahao Lin and Gim Hee Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11886–11895, 2021. 6

[29] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. 2024. 2, 3, 4, 5

[30] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6238–6247, 2021. 2

[31] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 1, 2

[32] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2

[33] Alejandro Perez-Yus and Antonio Agudo. Matching and recovering 3d people from multiple views. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3622–3631, 2022. 1

[34] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 573–586. Springer, 2012. 1, 2

[35] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15190–15200, 2021. 1

[36] 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, 2, 6, 7, 11

[37] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[38] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 1

[39] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2361–2368, 2014. 1, 2

[40] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11148–11157, 2021. 6

[41] 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. 11

[42] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016. 1, 2

[43] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision*, pages 142–159. Springer, 2022. 1, 6

[44] 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, 4, 5, 6, 7, 11, 12

[45] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1324–1333, 2020. 1

[46] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5

[47] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2, 3

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 [41]. We used the same backbone weights as previous works [27, 36, 44] 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 [44], followed by NMS to remove redundant tokens [27]. 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 [21], Shelf [3], and Campus [3] datasets. Note that only CMU Panoptic was used for training.

- **CMU Panoptic** [21] 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** [3] 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** [3] 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 7,

$$AP = \sum_{k=0}^{k=n-1} [r(k) - r(k+1)] * p(k) \quad (7)$$

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 [27, 44] 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).