3DSA :Multi-View 3D Human Pose Estimation With 3D Space Attention Mechanisms

Pohan Chen and Chiachi Tsi *

National Cheng Kung University
n28111089,cctsai@gs.ncku.edu.tw

Abstract. In this study, we introduce the 3D space attention module (3DSA) as a novel approach to address the drawback of multi-view 3D human pose estimation methods, which fail to recognize the object's significance from diverse viewpoints. Specifically, we utilize the 3D space subdivision algorithm to divide the feature volume into multiple regions. Predicted 3D space attention scores are assigned to the different regions to construct the feature volume with space attention. The purpose of the 3D space attention module is to distinguish the significance of individual regions within the feature volume by applying weighted attention adjustments derived from corresponding viewpoints. We conduct experiments on existing voxel-based methods, VoxelPose and Faster VoxelPose. By incorporating the space attention module, both achieve state-of-the-art performance on the CMU Panoptic Studio dataset.

Keywords: : 3D Human Pose Estimation · 3D space attention

1 Introduction

Estimating multiple 3D human poses simultaneously from multiple camera views is an enduring challenge in computer vision. The aim is to determine the 3D locations of the body joints for all people present in a scene. It is a task that offers benefits to numerous real-world applications, including intelligent sports analysis [5] and retail monitoring [35].

In 2D-3D lifting approaches [9, 10, 42], a monocular pose estimator identifies 2D bounding boxes and 2D poses for individuals in each view. A multi-view matching algorithm then establishes consistent correspondences between the 2D poses across different views. Finally, the matched 2D poses are lifted to 3D using geometry models such as triangulation [15] or Pictorial Structure Models [2,3,14].

As shown in Fig. 1(a), the voxel-based method [35] constructs the discretized feature volume from the detected heatmaps through 2D-3D projection. Based on the identified per-person proposal, the 3D pose for each person is individually estimated by feeding the feature volume into 3D-CNNs. To reduce the computational cost, another voxel-based approach [38] re-projects the feature volume to three two-dimensional coordinate planes and replaces the 3D-CNNs

* Corresponding author: Chiachi Tsi, cctsai@gs.ncku.edu.tw

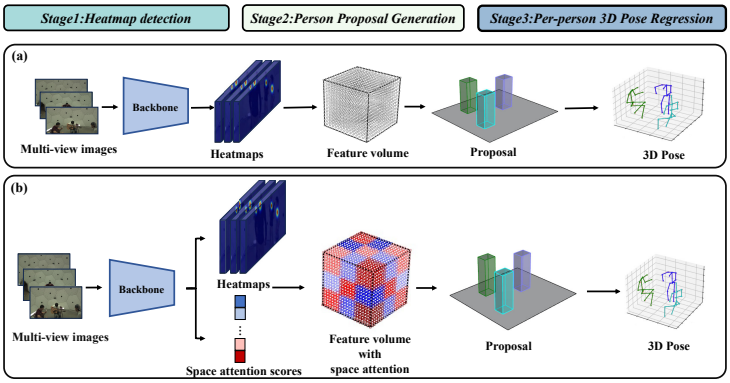

Fig. 1: Comparison between our method and the existing voxel-based method. The primary distinction is that: (a) The existing method simply projects heatmaps into 3D space. (b) We enhance feature volumes using space attention, maintaining 3D information, and emphasizing critical regions within the feature volume.

with 2D-CNNs, which increases inference speed. The existing direct prediction method [40] uses the Transformer architecture to regress multi-person 3D poses directly, bypassing the need for intermediate tasks. However, owing to the constraints of the transformer architecture, the inference speed of the method still falls behind the Voxel-based 2D-CNNs method [38].

Existing multi-view approaches often fail to consider an important visual phenomenon: the visible parts of the same object should differ when observed from various angles. As depicted in Fig. 2, the four people in the scene are visible in Views 1 to 4. However, in View 5, only three are visible in the image due to obstruction by one of the people. To address this issue, we propose the 3D space attention module (referred to as 3DSA) and apply it to two open-source voxel-based methods [35,38]. Fig. 1(b) shows the overview of our proposed method. We added the space attention layers to the end of backbone network to predict the space attention scores. Directly estimating the importance of each voxel within the feature volume could lead to excessive computational demands. Therefore, we employed the 3D space subdivision algorithm to divide the feature volume into multiple regions. The voxels within each region were treated as a group, and the space attention scores were assigned to the group, representing the importance of the region. Finally, the feature volume with space attention was constructed, retaining the 3D information and paying more attention to crucial regions.

We have conducted extensive experiments on the 3D human pose benchmark, Panoptic [21], to evaluate the efficacy of our space attention module. By applying the space attention module into the VoxelPose [35] and Faster VoxelPose [38] methods, our models show significant improvements of 20.93% and 20.32% in *MPJPE* respectively, both models achieve the state-of-the-art results. The voxel-based methods undermine their performance on the AP_{25} metric when compared to other multi-view methods. Our space attention module addresses

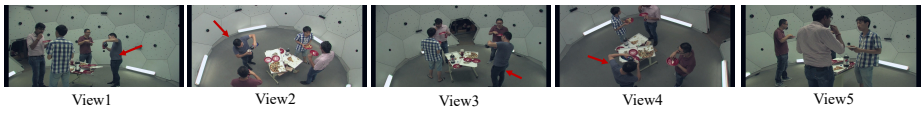

Fig. 2: The visual phenomenon in the 3D space. Due to differences in camera viewing angles and obstruction issues, the visibility of the same person changes across different viewpoints. As shown by the red arrow in the figure, we can clearly observe the person in the images from Views 1 to 4, but they are not visible in View 5. This visual phenomenon is commonly encountered in multi-view human pose estimation tasks.

this weakness, resulting in our model achieving 94.2% and 94.22% on the metric. Compared to the baseline model [35, 38], these scores demonstrate a significant improvement, surpassing all existing multi-view approaches.

Our contributions are summarized as follows: (1)We proposed the 3D space attention module (3DSA), which addresses the drawbacks of the existing multi-view 3D human pose estimation methods and validates its effectiveness on existing voxel-based methods [35, 38]. (2)We introduced a 3D space subdivision algorithm to reduce the computational complexity of the module. (3)By incorporating our space attention module into existing voxel-based methods [35, 38], both models achieve state-of-the-art results on the Panoptic benchmark, demonstrating the effectiveness of this attention mechanism.

2 Related work

2.1 Multi-view 3D human pose estimation

Unlike monocular 3D human pose estimation [8, 13, 33, 34], multi-view human pose estimation leverages image information from different viewpoints. This approach not only effectively overcomes challenges such as occlusion and depth ambiguity but also ensures a richer and more accurate depiction of the 3D pose. Existing methods can be categorized into three types: (1)2D to 3D lifting methods [1–3,5,9,10,18,25,42] (2)Voxel-based methods [6,7,19,20,27,30,32,35,38,41] (3)Direct regression method [40].

2D to 3D lifting method Firstly, a monocular pose estimator is utilized to estimate the 2D joints of each person in each view, through triangulation [15] and a 3D pictorial model [14], the 3D pose of each person is reconstructed from the associated 2D poses. Dong *et al*. [9,10] propose MvPose. MvPose utilizes a human pose detector to generate and cluster 2D bounding boxes and associated poses for each view. Each cluster represents the same person from different views. The 3D pose of each person is then reconstructed from these clusters using triangulation and a 3D pictorial model. The drawback of this 2D to 3D pose lifting method is its significant dependence on the preceding steps of 2D pose estimation and cross-view matching, as their quality directly influences the results.

Voxel-based method In contrast to the 2D to 3D lifting methods, which require establishing cross-view correspondence based on noisy and incomplete 2D

pose estimates, the voxel-based method directly operates in the 3D space and therefore avoids making incorrect decisions in each camera view. Tu *et al.* [35] propose VoxelPose, the method that discretizes the 3D space into voxels and uses 2D heatmaps to construct a 3D feature volume. 3D-CNNs process this volume to locate human proposals and regress the 3D joint. Since the voxel-based method heavily relies on 3D convolutions, it requires higher computational cost and inference time to predict 3D joints. To enhance the model efficiency, Ye *et al.* [38] proposed Faster VoxelPose, an optimization method based on orthographic projection. This method projects the 3D feature volume to three mutually perpendicular planes and then utilizes 2D-CNNs to locate the center proposal and regress the 3D joint. By doing this, it eliminates the need for time-consuming 3D convolutions. Choudhury *et al.* [7] proposed TEMPO, which utilizes temporal context to enhance pose estimation, delivering smoother and more accurate human poses by integrating spatiotemporal information.

It has been observed that the voxel-based methods generally yield lower scores on the AP_{25} metric in Panoptic datasets when compared to other methods. In this paper, we introduce a novel 3D space attention module, which applies weighted attention adjustments to the feature volume from corresponding viewpoints. This attention mechanism guides the network to focus more effectively on crucial feature regions and yields significant improvements in the AP_{25} metric.

Direct regression method In contrast to previous methods, Zhang *et al.* [40] proposed MvP, which leverages the Transformer architecture to regress multi-person 3D poses directly, thus eliminating the need for intermediate tasks. MvP achieved impressive results on the Panoptic [21] datasets. It showed significant progress (8%) on the most stringent AP_{25} compared to the Voxel-based methods [35, 38] and is more robust and accurate than previous models. However, due to the limitations of the transformer architecture, the inference speed of MvP still can't compete with 2D CNN-based voxel method [38], which is not conducive to its deployment in practical applications.

2.2 Multi-view 3D body mesh estimation

Multi-view 3D body mesh estimation [11, 20, 24, 31, 39] is a task closely related to 3D pose estimation. Instead of directly estimating joint positions, this task involves predicting the parameters of SMPL [26] or employing a fitting method [4] to align the SMPL model with detected joint positions. Yu *et al.* [39]use neural networks to directly predict local attention, assigning importance to visual features across views. Our method focuses on using space subdivision and space attention to address the varying importance of different viewpoints in the same 3D space. Directly predicting the space attention and projecting to the 3D space will result in equal attention values along the projection ray, which prevents the model from accurately identifying depth information.

2.3 Attention mechanisms

The methodology of predicting attention scores from input features and then using these scores to enhance discriminative feature learning has been adopted by numerous studies [17, 23, 36]. The most famous is SENet proposed by Hu *et al*. [17], which employs attention mechanisms to adaptively recalibrate channel-wise features by modeling inter-channel dependencies. Ma *et al*. proposed global attention in ContextPose [28], which focus on features within each voxel by estimating confidence scores for each joint, effectively reducing interference from non-human body voxels and improving joint estimation accuracy. Regarding merging 3D features extracted from different 2D viewpoints, the inherent physical characteristics of imaging result in varying importance of different viewpoints for the same 3D space. Therefore, we introduced the space attention module to solve this problem in a voxel-represented 3D space.

3 Method

3.1 The drawback of existing multi-view 3D human pose method

Despite the impressive achievements of the existing multi-view 3D human pose methods, they ignore an important visual phenomenon: the visible parts of an object could vary when observed from different viewpoints. Specifically, an object's visibility can differ dramatically across various viewpoints, for instance, an object may be distinctly visible from viewpoint A, yet as we transition to viewpoint B, its visibility may diminish or even vanish due to interposing obstacles or occluded persons.

In this work, we introduce the space attention module to address the drawback, and we validate its effectiveness on existing voxel-based methods [35, 38]. The existing methods merely project heatmaps into 3D space. As depicted in Fig. 1(b), our approach leverages the space attention module to enhance feature volumes. This not only preserves 3D information but also emphasizes crucial regions inside the feature volume. The objective of this attention mechanism is to focus on significant regions within the feature volume, by applying weighted attention adjustments to the feature volume from corresponding viewpoints.

3.2 Network architecture

Heatmap and space attention prediction. As shown in Fig. 3 (a), our model adopts a simple multi-layer design with a backbone and two additional layers. In the heatmap layer, the probability of a 2D pose heatmaps for the corresponding view is predicted. Meanwhile, in the space attention layer, the attention scores of the feature volume are determined. The attention scores are dynamically adjusted based on the input image, emphasizing regions with higher visibility in the 3D space.

Space attention with person proposal generation. As shown in Fig. 3 (b), by projecting the output heatmaps to the 3D space, the discretized feature volume $\left\{\mathbf{G} \in \mathbb{R}^{80\times80\times20}\right\}$ is constructed. Following [35], the 3D space is discretized

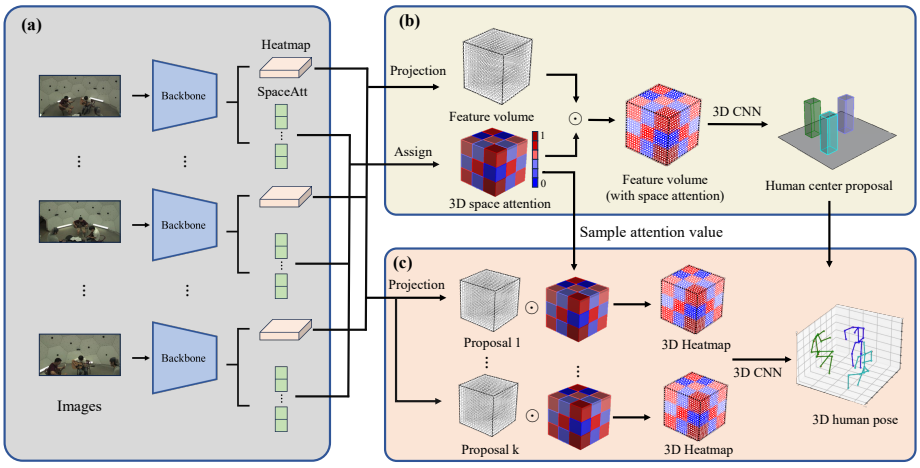

Fig. 3: Overview of network architecture. (a) Given the multi-view image as input, the backbone network predicted both the heatmaps and the space attention scores for each view. Each heatmap is projected to a 3D space, which is physically shared but independent for each view, constructing the feature volume. The space attention scores for each view are assigned to the different regions in the shared 3D space. (b) By performing an element-wise multiplication of the raw feature volume with the space attention scores, we produce a feature volume infused with spatial attention. Subsequently, this attention-enhanced feature volume is fed into 3D-CNNs to locate the per-person proposal. (c) A more detailed feature volume corresponding to the proposal was generated. By calculating the spatial relationship between the proposal and the feature volume, space attention scores for the proposal were sampled from the attention in 3D space. Finally, the human pose was estimated.

into $X \times Y \times Z$ locations. Based on observations from the space [2,21],X, Y and Z are set to be 80, 80, and 20 respectively to maintain a good balance between speed and precision. Let the 2D heatmap of a view be denoted as $\boldsymbol{M_v} \in \mathbb{R}^{K \times w \times h}$, where K is the number of person's joints. For each voxel location $\boldsymbol{G}^{X,Y,Z}$, the projected location in 2D view $\boldsymbol{V}$ is represented as $\boldsymbol{P}_v^{X,Y,Z}$. The heatmap values at $\boldsymbol{P}_v^{X,Y,Z}$ is denoted as $\boldsymbol{M}_v^{X,Y,Z} \in \mathbb{R}^K$.$\boldsymbol{v} \in \mathbb{R}^V$ represents one view from total $\boldsymbol{V}$ views.

Directly predicting the importance of each voxel in the feature volume would result in an overwhelming computational burden(Given that the output dimension of the model equals the number of voxels in the feature volume, which is 128,000). To reduce computational complexity, we use a 3D space subdivision algorithm to divide the feature volume $\left\{\boldsymbol{G_V} \in \mathbb{R}^{80\times80\times20}\right\}$ from each view, $\boldsymbol{V}$,into several regions $\left\{Div\overline{\boldsymbol{G}}_{\boldsymbol{V}} \in \mathbb{R}^{80\times80\times20}\right\}$. Subsequently, the space attention scores predicted from the model are assigned to each region in the divided feature volume to compute the attention of the feature volume $\left\{\boldsymbol{V}_v^{X,Y,Z} \in \mathbb{R}^{80\times80\times20}\right\}$, which represents the attention scores for view $\boldsymbol{v}$.

After that, an element-wise multiplication is performed between the space attention $\{\boldsymbol{V_v}\}$ and the raw feature volume $\{\boldsymbol{M_v}\}$ to obtain a feature volume with space attention on view $\boldsymbol{v}$, denoted as $\boldsymbol{MV_v}$. Following this, the feature volumes (with space attention) constructed from multi-view images are fused on average to obtain the aggregated feature volume $\left\{\boldsymbol{F} \in \mathbb{R}^{80\times80\times20}\right\}$:

$$F = \frac{1}{V}\sum_{v=1}^{V} M_v \tag{1}$$

where $\boldsymbol{V}$ represents the number of cameras. F represents the likelihood of K joints in G. Through applying space attention to the feature volume, 3D information is retained while emphasizing important voxels. Finally, the aggregated feature volume $\boldsymbol{F}$ is input into the 3D convolutional network to determine the per-person likelihood in the 3D discretized feature volume.

Space attention with per-person pose regression. In the final step, the completed 3D human pose corresponding to the proposal is predicted, as illustrated in Fig. 3 (c). For a fair evaluation of the effect of the space attention module, [35] is adopted to build an individual fine-grained feature volume centered at each predicted proposal. The size of the fine-grained feature volume is set to be $2000\,\text{mm} \times 2000\,\text{mm} \times 2000\,\text{mm}$, and the feature volume is divided into a discrete grid with $X' \times Y' \times Z'$ voxel where X', Y', Z' equal to 64. Each feature volume under a particular perspective will only have one space attention score to indicate its importance. In this work, we sample the attention score for each proposal by analyzing the spatial relationship between the proposal and the feature volume. Then, we employ a nearest neighbor sampling method to precisely calculate the attention scores for each proposal. The aggregated fine-grained feature volume is computed based on the descriptions from the previous stage. Finally, the 3D heatmap is estimated and the complete 3D human poses of the persons in the space are regressed.

3.3 3D space subdivision algorithm

As mentioned in Sec. 3.2, the 3D space subdivision algorithm is crucial to the implementation of our space attention module. Computational challenges arise when directly predicting the significance of each voxel in the feature volume. Inspired by Lai *et al.* [22] utilizing the cell subdivision search algorithm to reduce the computational complexity associated with searching through a large amount of data points, we employ a 3D space subdivision algorithm to divide the feature volume into distinct regions. Specifically, the voxels within each region are considered as a group, and attention scores are assigned to these groups to signify the importance of each region. Through the backbone network, the weight of each region is predicted, representing the importance of corresponding areas within the same viewpoint in the feature volume. If voxels within a specific region exhibit higher confidence levels, this indicates their relative importance. Conversely, lower confidence levels in voxels, caused by obstructions, occlusion, or other factors, suggest that they are less significant within that region. As

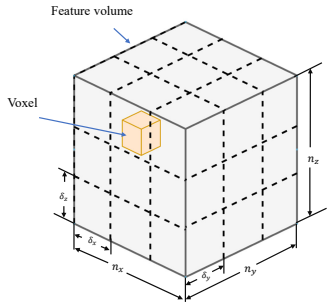

Fig. 4: Subdivision of the voxel within the feature volume. We utilize a 3D space subdivision algorithm to partition the feature volume into separate regions, with the voxels in each region being treated as a group.

Fig. 4 depicts, the feature volume in 3D space $\boldsymbol{G}$ is divided into several cells along the x, y, and z axes. Assume l_i, $i \in \{x, y, z\}$ represents the length, width, and height of the feature volume, while δ_i, $i \in \{x, y, z\}$ represents the cell length along a particular axis. The relationship between l_i and δ_i can be expressed as follows:

$$\delta_i = \text{int}\left(\frac{l_i}{n_i}\right) + 1 \quad i \in \{x, y, z\} \tag{2}$$

where n_i represents the number of regions divided along the i-axis. The total number of regions n_{total} in the 3D space is given by the product of

$$n_{\text{Total}} = n_x \times n_y \times n_z \tag{3}$$

Let the position vector of a voxel be $\mathbf{V} = [v_x, v_y, v_z]^\top$. Then, the region that $\mathbf{V}$ resides in can be computed using the following equation:

$$i_j = \text{floor}\left(\frac{(v_j - j_{\min})}{\delta_j}\right) + 1 \quad j \in \{x, y, z\} \tag{4}$$

where i_j represents the indices of the voxel in x, y, z directions, floor() is used to round down to integer representation, and $j_{\min}$ represents the minimum coordinates in x, y and z directions of the voxel within the feature volume. Finally, the region id of the voxel ($Voxel_{id}$) within the feature volume can be calculated by the following formula:

$$Voxel_{id} = i_z \times (n_x \times n_y) + i_y \times n_x + i_x \tag{5}$$

The ID of each voxel can be calculated according to the formulas, however, in practical applications, the total number of voxels in the feature volume is substantial, which could lead to excessive computation times. To tackle this

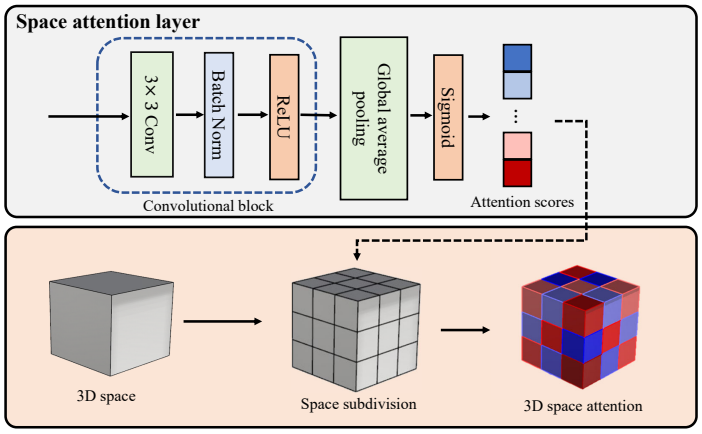

Fig. 5: Detailed architecture of space attention module.

challenge, we have optimized the weight assignment process within the space attention module, adopting the following Python code (Algorithm 1). Compared Eq. (2) to Eq. (5), our approach is better adapted to practical applications, achieving the same objectives and results more efficiently.

Algorithm 1 Weight assignment algorithm

```
# Suppose we have 3 intervals along x, y, and z axis
x, y, z = [0, 27, 54, 80], [0, 27, 54, 80], [0, 7, 14, 20]
# Assign space attention value to the tensor for one view
subdivision num = 0
for i in range (3):
    for j in range (3):
        for k in range (3):
            space attention[ x [i]: x [i+1],y [j]: y [j+1],z [k]: z [k+1]]
            = attention value[subdivision num]
            subdivision num = subdivision num+1
```

3.4 Implementation of space attention module

In implementations, only the following adjustments were made: (1) A simple branch was derived from the backbone network [16] to predict the space attention scores. (2) We executed an element-wise multiplication of the raw feature volume with the space attention scores calculated by Algorithm 1. (3)The attention scores of the proposal are computed by analyzing the positional relationship between the proposal and the feature volume.

The space attention module can be easily applied to existing multi-person voxel-based human pose methods [7, 35, 38, 41]. However, since some of these methods are not open-sourced, it prevents us from performing validation. Consequently, we chose to validate our method using the two open-sourced voxel-based methods [35, 38].

It is important to emphasize that for a fair evaluation of the impact of the space attention module on existing voxel-based methods [35,38], the network architecture [29] used for locating the person proposal and regressing the 3D pose remained unaltered. For the model's loss function and hyperparameter configuration, the original design proposed by [35, 38] has remained.

The architecture of the space attention layer is presented in Fig. 5. It is a straightforward and lightweight design, which uses a simple convolutional block followed by global average pooling and the sigmoid activation function to estimate the space attention scores of the corresponding image. The purpose of the global average pooling is to replace the traditional fully connected layers, thereby reducing the number of parameters. The output dimensions of the space attention layer are equal to the number of regions in the feature volume. The space attention scores $\boldsymbol{S} \in \mathbb{R}^n$ represent the n space attention values, indicating that the feature volume is divided into n regions.

4 Experiments

4.1 Implementation detail

Training and evaluation datasets. CMU Panoptic [21] is a 3D dataset with multi-view images. To evaluate and analyze our approach, we conducted extensive experiments on the Panoptic dataset. Following VoxelPose [35], the same data sequences were used for both training and evaluating our model. Our experiments were conducted using five HD cameras with camera IDs 3, 6, 12, 13, 23. Shelf and Campus [2] are two datasets that are commonly used in multi-view and multi-person research. We evaluated our method using the same data setup as in [35].

Evaluation metric. For the Panoptic datasets [21], we adopt the Average Precision (AP^K) and Mean Per Joint Position Error ($MPJPE$) as metrics that demonstrate the robustness and accuracy of multi-person 3D pose estimation. To assess the influence of the space attention module on model size and computational complexity, we consider key metrics such as $MACs$ and model parameters. For both the Campus and Shelf datasets, we present the results in terms of the Percentage of Correct Parts (PCP).

Training details. For the Panoptic datasets, we use an off-the-shelf pose estimation model constructed based on ResNet-50 [16] to extract features from multi-view images. The difference from VoxelPose [35] is that since our backbone network needs to predict the space attention scores, the parameters of the model are updated throughout the training iteration.

Due to the incomplete data annotation in the Campus and Shelf datasets [2], Tu *et al.* [35] use synthetic 3D poses to train the network. To implement the

space attention module, we use the synthetic heatmap as the input feature to predict the space attention scores. In summary, the space attention module has two modes: the first predicts the space attention scores from the ground truth multi-view image, referred to as Image-based input; the second predicts the space attention scores from the synthetic heatmaps, referred to as Heatmap-based input.

4.2 Comparisons to Existing Methods

Panoptic. We first evaluate our model on the Panoptic dataset [21] and compare it with the state-of-the-art model. As illustrated in Tab. 1, by incorporating the space attention module($10 \times 10 \times 3$ configuration) with two voxel-based methods, VoxelPose [35] and Faster VoxelPose [38], our model achieves 94.2% and 94.22% on the most strict evaluation metric AP_{25}, outperforming the transformer model MvP [40]. Our proposed method shows inferior performance in terms of AP@50,100,150 when compared to VoxelPose, and this 0.5% performance gap is generally attributed to model variation. It particularly emphasizes that in terms of the AP_{25} metric, our method has significantly improved, outperforming VoxelPose by 12.69% and Faster VoxelPose by 10.56%. Remarkably, both methods achieved much lower *MPJPE* with values of 13.98 and 14.55, outperforming the TEMPO [7] and achieving the SOTA records. This demonstrates the effectiveness of our space attention module.

Table 1: Comparison with existing methods on the Panoptic datasets.

Method	AP_{25}	AP_{50}	AP_{100}	AP_{150}	MPJPE
VoxelPose [35]	83.59	98.33	**99.76**	**99.91**	17.68mm
Faster VoxelPose [38]	85.22	98.08	99.32	99.48	18.26mm
PlaneSweep Pose [25]	92.12	98.96	99.81	99.84	16.75mm
RPGN [37]	–	–	–	–	15.84mm
MvP [40]	92.28	96.6	97.45	97.69	15.76mm
TEMPO [7]	89.01	**99.08**	99.76	99.93	14.68mm
VoxelPose + 3DSA	94.2	98.49	99.21	99.31	**13.98**mm
Faster VoxelPose + 3DSA	**94.22**	98.65	99.49	99.75	14.55mm

Campus and Shelf. The quantitative evaluation results on Shelf and Campus datasets [2] are presented in Tab. 2. Our proposed method (VoxelPose [35] with space attention,$10 \times 10 \times 3$ configuration) remains competitive on both datasets. The performance of space attention is not as outstanding on Panoptic datasets [21], and we believe this is related to the Heatmap-based input. Since the heatmap lacks image information, the model is hard to determine the importance of different regions in 3D space from the heatmap. We will detail our research on this issue in the subsequent ablation study.

Table 2: Quantitative results on Shelf and Campus datasets.

Method	Shelf				Campus			
	Actor1	Actor2	Actor3	Average	Actor1	Actor2	Actor3	Average
Ershadi et al. [12]	93.3	75.9	94.8	88	94.2	92.9	84.6	90.6
Dong et al. [10]	98.8	94.1	97.8	96.9	97.6	93.3	98	96.3
MvP [40]	99.3	94.1	97.8	97.4	98.2	94.1	97.4	96.6
TEMPO [7]	99.3	95.1	97.8	97.4	97.7	95.5	97.9	97.3
Faster VoxelPose. [38]	99.4	96	97.5	97.6	96.5	94.1	97.9	96.2
VoxelPose [35]	99.3	94.1	97.6	97	97.6	93.8	98.8	96.7
Ours	99.4	95.4	97.6	97.5	98	93.4	98.6	96.7

Table 3: Space subdivision and efficiency analysis on the Panoptic dataset

VoxelPose incorporate with space attention					
Space subdivision	AP_{25}	AP_{100}	MPJPE	MACs(G)	Parameter(M)
Tu *et al.* [35]	83.59	99.76	17.68	178.88	40.62
$3 \times 3 \times 3$	92.73	99.58	14.78	179.09	40.64
$7 \times 7 \times 3$	93.71	99.33	14.41	180.04	40.77
$10 \times 10 \times 3$	94.2	99.21	13.98	181.24	40.92
$15 \times 15 \times 6$	94.33	99.1	13.97	193.24	42.47
$20 \times 20 \times 9$	94.44	99.44	13.94	221.58	46.15
Faster VoxelPose incorporate with space attention					
Space subdivision	AP_{25}	AP_{100}	MPJPE	MACs(G)	Parameter(M)
Ye *et al.* [38]	85.22	99.32	18.26	106.87	36.37
$3 \times 3 \times 3$	92.57	99.61	15.54	107.08	36.39
$7 \times 7 \times 3$	93.75	99.54	14.88	108.03	36.52
$10 \times 10 \times 3$	94.22	99.49	14.55	109.23	36.67

4.3 Ablation studies

In this section, we conduct ablative experiments to analyze a variety of factors within our approach.

Individual contributions of the space attention module and the 3D space subdivision algorithm. By comparing the results in Tab. 3, we can see that the finer the subdivision of the 3D space, the model's accuracy and precision improve correspondingly. However, the model's performance tends to converge after subdividing into $10 \times 10 \times 3$ regions. The result demonstrates the critical importance of the space subdivision algorithm within the space attention module. The direct prediction of all voxels does not result in significant improvements in performance.

Efficiency analysis. In this work, we focus on comparing our method with existing voxel-based methods [35, 38]. Tab. 3 demonstrates that incorporating the space attention module into the voxel-based approach resulted in a slight

increase in the model's complexity. Regarding the model we eventually selected(VoxelPose with $10 \times 10 \times 3$ space attention module), *MACs* increased by 1.32% and parameters by 0.74% when compared to the VoxelPose method. As previously mentioned, excessively increasing the number of spatial subdivisions does not enhance performance but significantly increases the model's complexity. For instance, subdividing the space into $20 \times 20 \times 9$ regions resulted in a 23.8% increase in the model's MACs and a 13.6% increase in parameters. This further demonstrates the importance of the space subdivision algorithm in improving the efficiency of the space attention module. To strike a balance between performance and efficiency, we adopt the $10 \times 10 \times 3$ space attention configuration on VoxelPose [35] to study the impact of the individual factors.

Number of cameras. We compared our method with existing 3D Pose methods [7, 35, 38, 40] . Tab. 4 shows that the feature volume representation is diminished with fewer camera views, leading to a drop in accuracy. The improvement in both *AP* and *MPJPE* metrics over other models, as the number of cameras increases, underscores the significance of multi-view images for enhancing the space attention module's performance.

Image-based input /Heatmap-based input. To further validate the impact of different inputs on the space attention module, we conducted experiments on the Panoptic dataset [2]. As shown in Tab. 5, although the space attention with Heatmap-based input shows an improvement compared to the baseline model [35], it is noticeably inferior to the space attention with Image-based input. We consider that this disparity occurs because heatmaps lack spatial and depth information in comparison to images.

Table 4: Number of cameras analysis on the Panoptic dataset

Method	Cam	AP_{25}	AP_{50}	AP_{100}	AP_{150}	MPJPE
Faster VoxelPose [38]	4	73.95	97.02	99.21	99.35	21.12
MvP [40]		84.1	–	96.7	–	19.3
TEMPO [7]		–	–	–	–	17.34
ours		88.4	98.1	99.59	99.7	16.78
VoxelPose [35]	3	58.94	93.88	98.45	99.32	24.29
Faster VoxelPose [38]		53.68	91.89	97.4	98.3	26.13
MvP [40]		71.8	–	95.1	–	21.1
TEMPO [7]		–	–	–	–	19.22
ours		73.06	95.23	98.64	99.25	19.03
MvP [40]	2	37.7	–	93	–	34.8
TEMPO [7]		–	–	–	–	32.13
ours		47.95	88.74	97.84	98.8	27.35

Table 5: Effect of different inputs on space attention module

Image-based input / Heatmap-based input					
Input	AP_{25}	AP_{50}	AP_{100}	AP_{150}	MPJPE
Image	94.2	98.49	99.21	99.31	13.98
Heatmap	86.97	98.3	99.29	9 9.38	17.21

4.4 3D space attention visualization

In Fig. 6, we provide the space attention visualization results on Panoptic datasets. Red regions indicate attention scores above 0.8, while blue for below 0.8. Observing the spatial distribution of attention in 3D space (1st row), most key attention areas are focused where people are present. In view 5, an obscured person is not visible from that angle, resulting in lower attention scores in that area. This result aligns with our hypothesis, confirming that the space attention mechanism discriminates the importance of different regions in the feature volume based on visibility. More visualization results are provided in the supplementary material.

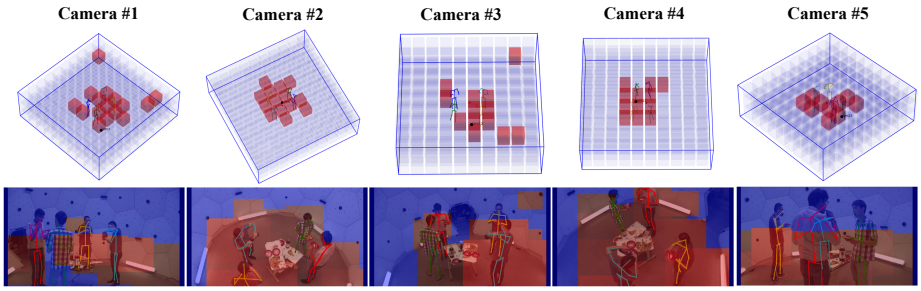

Fig. 6: 3D space attention visualization. We marked areas with scores above 0.8 (red regions) in 3D space (1st row) and projected them onto the corresponding 2D image (2nd row).

5 Conclusion

In this paper, we present the novel space attention module for the voxel-based multi-view 3D pose estimation method. We learn the space attention scores from the input image and utilize the 3D space subdivision algorithm to divide the feature volume, finally constructing the feature volumes with space attention. By integrating our space attention module into two existing voxel-based methods, both models achieve the state-of-the-art results on the panoptic benchmarks.

Acknowledgements

This work is supported and National Science and Technology Council (NSTC), Taiwan R.O.C. projects with grants 112-2222-E-006-009-, 113-2218-E-035-001-, 113-2425-H-006-007- and NSTC 113-2627-M-006-005 -.

References

1. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3d human pose estimation. In: Bmvc. vol. 1. Bristol, UK (2013)
2. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1669–1676 (2014)
3. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures revisited: Multiple human pose estimation. IEEE transactions on pattern analysis and machine intelligence **38**(10), 1929–1942 (2015)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016)
5. Bridgeman, L., Volino, M., Guillemaut, J.Y., Hilton, A.: Multi-person 3d pose estimation and tracking in sports. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
6. Chen, Y., Gu, R., Huang, O., Jia, G.: Vtp: volumetric transformer for multi-view multi-person 3d pose estimation. Applied Intelligence **53**(22), 26568–26579 (2023)
7. Choudhury, R., Kitani, K.M., Jeni, L.A.: Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14704–14714 (2023). https://doi.org/10.1109/ICCV51070.2023.01355
8. Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., Jain, A.: Learning 3d human pose from structure and motion. In: Proceedings of the European conference on computer vision (ECCV). pp. 668–683 (2018)
9. Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., Zhou, X.: 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)
10. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7792–7801 (2019)
11. Dong, Z., Song, J., Chen, X., Guo, C., Hilliges, O.: Shape-aware multi-person pose estimation from multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11158–11168 (2021)
12. Ershadi-Nasab, S., Noury, E., Kasaei, S., Sanaei, E.: Multiple human 3d pose estimation from multiview images. Multimedia Tools and Applications **77**, 15573–15601 (2018)
13. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2334–2343 (2017)

14. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on computers **100**(1), 67–92 (1973)
15. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
18. Huang, C., Jiang, S., Li, Y., Zhang, Z., Traish, J., Deng, C., Ferguson, S., Da Xu, R.Y.: End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. pp. 477–493. Springer (2020)
19. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7718–7727 (2019)
20. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7718–7727 (2019)
21. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015)
22. Lai, J.Y., Shu, S.H., Huang, Y.C.: A cell subdivision strategy for r-nearest neighbors computation. Journal of the Chinese Institute of Engineers **29**(6), 953–965 (2006)
23. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 510–519 (2019)
24. Li, Z., Oskarsson, M., Heyden, A.: 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1888–1897 (2021)
25. Lin, J., Lee, G.H.: Multi-view multi-person 3d pose estimation with plane sweep stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11886–11895 (2021)
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
27. Ma, X., Su, J., Wang, C., Ci, H., Wang, Y.: Context modeling in 3d human pose estimation: A unified perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6238–6247 (2021)
28. Ma, X., Su, J., Wang, C., Ci, H., Wang, Y.: Context modeling in 3d human pose estimation: A unified perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6238–6247 (2021)
29. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 5079–5088 (2018)

30. Reddy, N.D., Guigues, L., Pishchulin, L., Eledath, J., Narasimhan, S.G.: Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15190–15200 (2021)
31. Shin, S., Halilaj, E.: Multi-view human pose and shape estimation using learnable volumetric aggregation. arXiv preprint arXiv:2011.13427 (2020)
32. Su, J., Wang, C., Ma, X., Zeng, W., Wang, Y.: Virtualpose: Learning generalizable 3d human pose models from virtual data. In: European Conference on Computer Vision. pp. 55–71. Springer (2022)
33. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11179–11188 (2021)
34. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3d people in depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13243–13252 (2022)
35. Tu, H., Wang, C., Zeng, W.: 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. pp. 197–212. Springer (2020)
36. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
37. Wu, S., Jin, S., Liu, W., Bai, L., Qian, C., Liu, D., Ouyang, W.: Graph-based 3d multi-person pose estimation using multi-view images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11148–11157 (2021)
38. Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y.: Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: European Conference on Computer Vision. pp. 142–159. Springer (2022)
39. Yu, Z., Zhang, L., Xu, Y., Tang, C., Tran, L., Keskin, C., Park, H.S.: Multiview human body reconstruction from uncalibrated cameras. Advances in Neural Information Processing Systems **35**, 7879–7891 (2022)
40. Zhang, J., Cai, Y., Yan, S., Feng, J., et al.: Direct multi-view multi-person 3d pose estimation. Advances in Neural Information Processing Systems **34**, 13153–13164 (2021)
41. Zhang, Y., Wang, C., Wang, X., Liu, W., Zeng, W.: Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2613–2626 (2022)
42. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 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. pp. 1324–1333 (2020)