3D Human Pose Estimation with Dilated Sampled Frames

1st Ge Cao
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
caoge@islab.ulsan.ac.kr

2nd Qing Tang
Data Science Group
INTERX
Ulsan, Korea, Republic of
tangqing@interxlab.com

3rd Tran Tien Dat
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
tdat@islab.ulsan.ac.kr

4th Ashraf Uddin Russo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
russo@islab.ulsan.ac.kr

5th Kanghyun Jo*
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
acejo@ulsan.ac.kr

2023 International Workshop on Intelligent Systems (IWIS) | 979-8-3503-0504-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/IWIS58789.2023.10284636

***Abstract*—Three-dimensional (3D) human pose estimation (HPE) targets to produce the 3D spatial coordinates of the human pose from 2D images. 3D HPE is a basic computer vision task for many intelligent industrial applications. Commonly, the coordinates of the predicted 3D human pose joints are calculated through the 2D keypoint from the ground truth provided by the datasets or generated by a classical and robust 2D human pose estimator. With the development of transformer-based methods, the methods with a sequence of monocular images have achieved great success in 2D-to-3D lifting human pose estimation. In this paper, the sampled frames with a dilated ratio are given as the input of the 3D human pose estimator. Extensive experiments on the public benchmark Human 3.6M demonstrate the significance and effectiveness of the proposed method.**

***Index Terms*—3D human pose estimation, intelligent industry, monocular image, transformer.**

I. Introduction

3D human pose estimation (HPE) targets to generate the coordinates of human joints in the three-dimensional coordinate system. The 3D HPE could provide significant representations of human body motion, which has amounts of industrial applications like action recognition [1], virtual and augmented reality [2], human-robot interaction [3], Etc. The 3D HPE could be applied to reconstruct human motions in augmented reality applications and assist workers with complicated tasks. In some collaborative manufacturing conditions, 3D HPE is helpful in ensuring safe interactions and coordination between humans and machines.

Based on the different modeling approaches, HPE could be divided into the skeleton-based model and skinned multi-person linear (SPML)-based model, of which the latter is also called human body recovery. This paper focuses on the skeleton-based model and the method is developed based on 2D-to-3D lifting approaches [4], [5]. Thanks to the great performance of state-of-the-art 2D human pose estimator, lifting-based research become mainstream in 3D human pose estimation. Among them, some methods major in fusing 2D human pose from multiple views to get the final 3D human pose, the others employ a sequence of frames as the input for generating the target.

Recently, some methods [4], [5] were proposed to deal with the lifting-based issues and surely achieved some progress based on the large-scale dataset [6]. Especially, the poseformer v2 [5] overcomes the unreliable 2D pose estimation results problem and decreases the computing cost for the transformer-based method simultaneously. After the huge success achieved by ViT [7], the transformer-based methods [4] have made remarkable performance by modifying the Transformer to make it become *de facto* methods for 3D human pose estimation. These methods enhance the robustness of the network by providing temporal information from a continuous sequence of frames. However, this is accompanied by higher computational time and resource costs.

The poseformer v2 [5] attempts to take many frames as the input, but only a few continuous joint coordinates are given as input into the network to extract the spatial correlation. The discrete cosine transform (DCT) is utilized to compress the temporal information from the whole input sequence, which greatly reduces the computational cost. In this paper, we proposed to take the dilated sampled frames as the input based on the following observation. Commonly the RGB camera collects the sequence with 25 or 30 frames. The location of the human joints in the continuous 3 or 5 frames basically possesses no differences. In order to provide more spatially variable information, the dilated sampled frames are taken as the input for extracting spatial feature embeddings.

The remaining content is organized as follows. Section II

introduces the related work for the transformer-based methods. The main methodology is shown in section III. Extensive experiments are demonstrated in section IV, and the section V concludes this paper.

II. Related Works

The three-dimensional human body information is very complex and complicated for collecting and processing. Based on the specific consideration of the different types of human body structure, many researchers employed various models for their algorithm in 3D HPE. Generally, the most widely applied method is the skeleton and shape models. The skeleton-based model has been commonly applied in 2D human pose estimation for decades and many classic and efficient models are proposed based on it. Thanks to the dataset creator, the significant evaluation method is naturally extended to 3D human pose estimation.

Thanks to the great performance of the state-of-the-art 2D human pose estimator are the commonly employed methods for generating the 2D keypoints' coordinates of the initial input, sequence, or images. Based on the 2D joints' location, 2D-to-3D lifting human pose estimation methods leverage 2D poses to produce the 3D human pose joints coordinate in the three-dimensional coordinate system. Initially, most researchers focus on applying deep neural networks to regress 3D joint coordinates just from a single frame from a monocular RGB camera. And [8] focused on decreasing the depth ambiguity by fusing the heatmaps of 2D joints and 3D image cues. Nevertheless, based on the depth ambiguity, even within the same 2D joints skeleton, there are many varied 3D human poses. Iskakov et al. fused the images from different capture views and generated the final 3D human pose by algebraic triangulation. Kocabas et al. utilized epipolar geometry to predict 2D human pose from multi-views and gained the 3D human pose finally. Rhodin at al. only trained with multiple views for 2D human poses and finally predicted the 3D human pose and camera pose. So taking only a single image as the input is not enough for providing spatial information, a sequence of images can provide temporal information to enhance the robustness and performance [4], [9]. Among them, [10] applied LSTM cells to extract temporal features for the initial input sequence. There are many methods that employ the deep neural network to extract spatial and temporal relationships simultaneously. But those works only project the 2D joint coordinates into a latent space, which is not learnable and lacks spatial and temporal correlation.

After the great success of vision transformer, the transformer-based methods obtained better performance in almost vision tasks like object recognition, object detection, object segmentation, Etc. PoseFormer [4] firstly adopted and modified the vision transformer as the backbone network to fit the training for 2D-to-3D lifting human pose estimation and achieved state-of-the-art performance. MixSTE proposed a mixed spatio-temporal encoder to separately learn the temporal and spatial information of inter-frames correlation and each joint correlation. And MHFormer proposed a transformer-based method to learn spatial representations of multiple hypotheses for potential 3D human poses, which applied a one-to-many mapping first and then a many-to-one mapping with multi-hypothesis. However, most of the research met the problem that the computational cost is very high with the sequence input. PoseFormer V2 [5] dealt with this problem by applying a discrete cosine transform to represent the temporal information of each joint, which reduces the computational resource by a large step.

III. Methodology

The proposed method is based on the PoseFormer V2 [5]. The section III. A introduces the basic preliminaries of the spatial transformer. Section II. B demonstrates the temporal transformer and applied discrete cosine transform.

A. Spatial Transformer Encoder

Given the input $X \in \mathbb{R}^{f\times(J\cdot 2)}$, where f denotes the number of frames, J denotes the number of human joints, and 2 means the input sequence is 2D joints coordinates in two-dimensional space. So $\left\{x^i \in \mathbb{R}^{1\times(J\cdot 2)}\right\}$ denotes the input joints coordinates of each frame. Following the [5], there is $X' \in \mathbb{R}^{F\times(J\cdot 2)}$ is sampled from the original sequence as the real input to decrease the computational cost in the spatial transformer encoder part, where F denotes the number of sampled frames which are much smaller than f. The joints coordinates of each frame are denoted as $\left\{x^{i'} \in \mathbb{R}^{1\times(J\cdot 2)} | i = 1, 2, ..., F\right\}$. So the F frames' 2D coordinates are taken as the input patch to the patch embedding $E \in \mathbb{R}^{(J\cdot 2)\times C}$ to extract each coordinate as a high dimensional feature, where C denotes the dimension of the coordinates representation. The spatial embedding $E_{spa} \in \mathbb{R}^{1\times J\times C}$. So the process of computing the patch embedding could be calculated by,

$$Z_0 = \left[x^{1'}E; x^{2'}E; ...x^{F'}E\right] + E_{spa} \tag{1}$$

where $Z_0 \in \mathbb{R}^{F\times C}$. The structure of the spatial transformer encoder is the vanilla one following [7],

$$Z'_l = MSA(LN(Z_{l-1})) + Z_{l-1}, \quad l = 1, 2, ..., L_1 \tag{2}$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l, \quad l = 1, 2, ..., L_1 \tag{3}$$

where MSA denotes the multi-head self-attention block, MLP denotes the multi-layer perceptron block, and LN denotes the layer normalization, so the final output of the spatial transformer encoder could the denoted as $Z_{L_1} \in \mathbb{R}^{F\times(J\cdot 2)\cdot C)}$.

When sampling the F frame for the input of the spatial transformer, the baseline will choose the index of $[Central - \left\lfloor \frac{F}{2} \right\rfloor, ..., Central, ..., Central + \left\lfloor \frac{F}{2} \right\rfloor]$, where the $Central$ denotes the index of the central frame of the full input sequence. And the dilated sample frames would be obtained from the real input frame with a dilated frame d.

TABLE I
EXTENSIVE EXPERIMENTS RESULTS IN PROTOCOL 1 (MPJPE) AND PROTOCOL 2 (P-MPJPE) ON HUMAN 3.6M DATASET [6]. "*" MEANS THE RESULTS ARE RE-TRAINED BY THIS PAPER.

Protocol #1	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Fang et al. (AAAI'18) [9]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Dabral *et al.* (ECCV'18) [11]	44.8	50.4	44.7	49.0	52.9	61.4	43.5	45.5	63.1	87.3	51.7	48.5	52.2	37.6	41.9	52.1
GraphH (CVPR'21) [12]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Cai *et al.* (ICCV'19) [13]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
MGCN (ICCV'21) [14]	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4
ST-GCN (ICCV'19) [15]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
PoseformerV2* (CVPR'23) [5]	46.2	43.2	52.8	43.9	47.5	57.9	48.7	43.4	47.7	64.4	47.5	47.6	53.2	34.3	35.0	47.5
Proposed	45.2	42.6	51.9	42.9	46.8	58.3	48.1	42.9	48.8	65.6	46.5	46.9	52.5	30.0	32.9	46.8
Protocol #2	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Pavlakos *et al.* (CVPR'18) [16]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Hossain *et al.* (ECCV'18) [10]	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Cai *et al.* (ICCV'19) [13]	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	32.3	39.0
Lin *et al.* (BMVC'19) [17]	32.5	35.3	34.3	36.2	37.8	43.0	33.0	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Pavllo *et al.* (CVPR'19) [18]	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
PoseformerV2* (CVPR'23) [5]	31.4	33.4	40.2	35.3	35.3	43.6	34.8	32.4	35.3	53.2	37.6	33.9	39.8	25.1	27.6	35.9
Proposed	31.6	33.5	40.3	35.5	35.5	43.9	34.9	32.9	35.5	53.2	37.4	33.7	40.1	23.8	27.2	35.9

B. Temporal Transformer Encoder

Before starting introducing the temporal encoder, the discrete cosine transform should be demonstrated first. Following PoseFormer v2 [5], the discrete cosine transform is employed to extract and compress the temporal-spatial correlation of a sequence of coordinates. From the original full sequence $X \in \mathbb{R}^{f\times(J\cdot 2)}$, where f denotes the length of the full sequence and J is the number of joints. So we could obtain $2 \times J$ sequences of the coordinate values of the two axes (x and y), which could be denoted as $x_j \in \mathbb{R}^f$ and $y_j \in \mathbb{R}^f$, and the coordinate of the v-th frame and the j-th joint of the x-axis is denoted as $x_{j,v}$. So the i-th discrete cosine transform coefficient is formulated as [5],

$$C_{j,i} = \sqrt{\frac{2}{f}} \sum_{v=1}^{f} x_{j,v} \frac{1}{\sqrt{1+\sigma_{i1}}} cos(\frac{\pi}{2f}(2v-1)(i-1)), \quad (4)$$

where σ_{i1}=1, when i is equal to 1, or σ_{i1}=0. The discrete cosine transform coefficients represent the feature embedding in the frequency domain. The original input sequence could be recovered by inverse discrete cosine transform,

$$x_{j,v} = \sqrt{\frac{2}{f}} \sum_{v=1}^{f} C_{j,i} \frac{1}{\sqrt{1+\sigma_{i1}}} cos(\frac{\pi}{2f}(2v-1)(i-1)) \quad (5)$$

So the original full sequence $X \in \mathbb{R}^{f\times(J\cdot 2)}$ is processed by the discrete cosine transform (DCT) and taken the n DCT coefficients t obtain the $X_D \in \mathbb{R}^{n\times(J\cdot 2)}$. Then a frequency embedding $E_{Freq} \in \mathbb{R}^{(J\cdot 2)\times((J\cdot 2)\cdot C)}$ is utilized to make it a learnable embedding to represent the temporal information as,

$$T_0 = [Z_{L_1}; X_D^1 E_{Freq}; X_D^2 E_{Freq}, ..., X_D^n E_{Freq}] + E_{Tpos} \quad (6)$$

Then process the T_0 with L_2 layers of transformer following [5] to predict the 3D human pose $p \in \mathbb{R}^{1\times(J\cdot 3)}$.

The loss for training the whole architecture is the standard MPJPE (Mean Per Joint Position Error). The detail of MPJPE is introduced in section IV.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

Dataset. The extensive experiments and ablation study is trained and tested on the widely-applied public dataset Human3.6M [6].

Evaluation metrics. MPJPE (Mean Per Joint Position Error): This metric is calculated by,

$$E_{MPJPE}(j,S) = \frac{1}{N_S} \sum_{i=1}^{N_S} \left\| P_{p,S}^{(j)}(i) - P_{gt,S}^{(j)}(i) \right\|_2, \quad (7)$$

where j denotes a frame and S denotes the corresponding skeleton model, e.g. in this paper, N_S is set to 17 as the number of joints. And $P_{p,S}^{(j)}(i)$ is the predicted coordinate, $P_{gt,S}^{(j)}(i)$ denotes the ground truth corresponding to the predicted 3D human pose. P-MPJPE is setting following [5].

B. Implementation Details

Model hyper-parameters. The dimension of the embedding feature for each joint's coordinate C is equal to 32. The number of the spatial transformer layer $L_1 = 4$. The number of the temporal transformer layer $L_2 = 4$ following [5]. The length of input sequence f is set to 81, and the real input number of frame F is set to 3. The number of kept discrete cosine transform coefficients n is equal to 3. The dilated frame d is set to 0 to 4 which is shown in Table. II.

Experimental settings. All the extensive experiments are implemented by Pytorch ToolBox. The AdamW optimizer is applied for training the architecture for 200 epochs with a weight decay of 0.99. The batch size of the training process is set to 1024. The initial learning rate is set to 8e-4 with an exponential learning rate decay schedule and the decay factor is 0.99. The final predicted 3D human pose results are based on the 2D pose detection from CPN instead of the ground truth from the dataset [6].

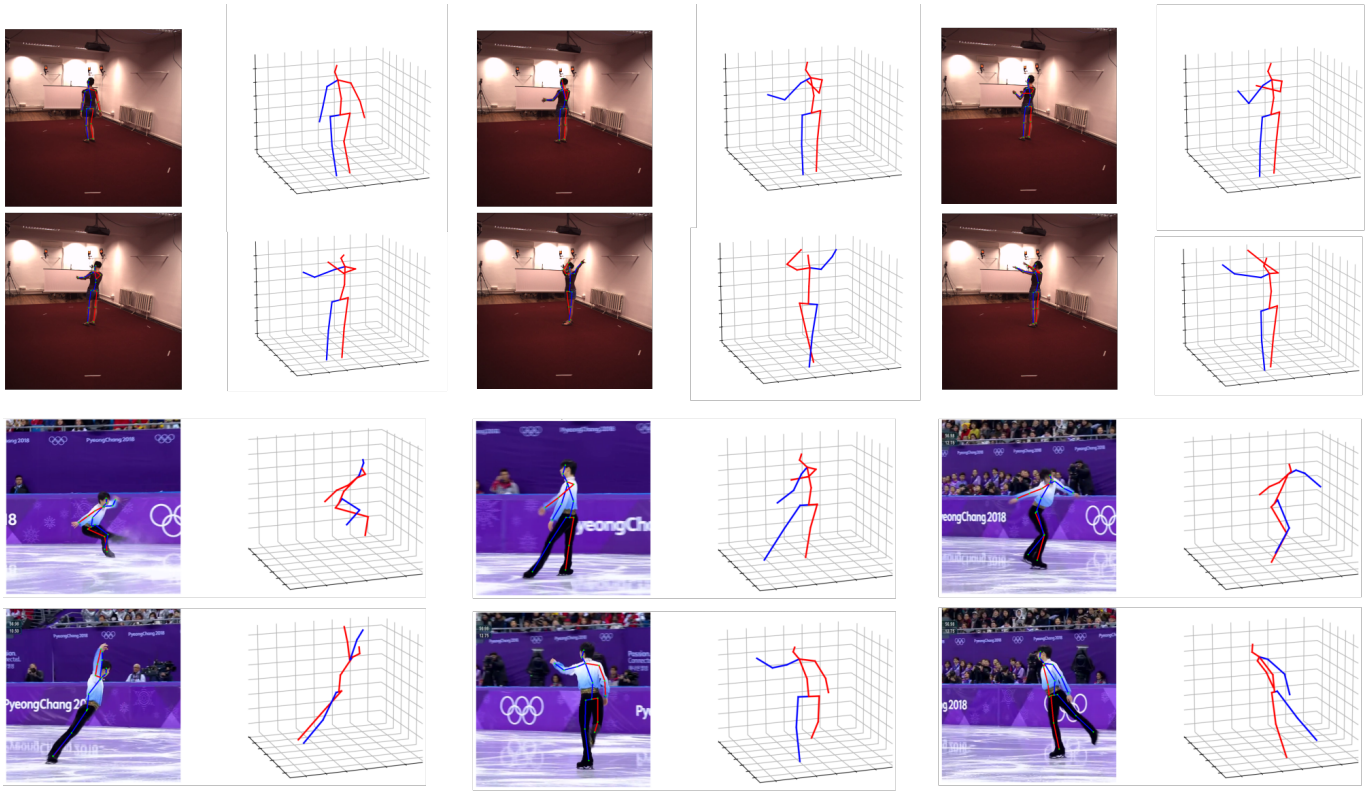

Fig. 1. Quantative results of the proposed method for on eof "Direction" video in S1 of Human3.6M dataset [6] and a video from the Internet.

C. Comparisons with other researches

For the dataset Human3.6M [6], the testing results of all 15 actions are shown in Table. I. The subtable on the top shows the results for protocol #1 MPJPE and subtable on the bottom shows the results for protocol #2 P-MPJPE, where the baseline poseFormer v2 is retrained with this paper (with f=81 and F=3). And the proposed method shown in the table is set with dilated frame d equal to 3. Especially, the poseFormer v2 and the proposed method occupy only 117.3 MFLOPs computing resources, which is much lower than other algorithms. As shown in Table. I, the performance of the proposed method surpasses the baseline on the average score and most subjects of action, which achieves an average of 46.8 mm in MPJPE and an average of 35.9 mm in P-MPJPE. And most of the evaluation value of the detail action is better.

D. Abaltion Study

In this subsection, the comparison experiments of different dilated frames are shown in Table. II, where the PoseFormerV2 is the baseline and d=0 means the original sample method. The proposed method with dilated frame d equal to 3 achieves the best performance. For the different dilated frames d, the performance has a different level of improvement, which proves the effectiveness of the proposed method.

E. Visualization

Some visualization results for training set S1 and test videos from the Internet are shown in Fig. 1 to show the effectiveness of the proposed method. The upper six groups of images show the results of the S1 training set. The bottom groups of images show the results tested on images from the Internet. It is obvious that these displayed results demonstrate the superiority of the 3D HPE model.

TABLE II
COMPARISONS OF DIFFERENT DILATED FRAMES d ON HUMAN3.6M, WHERE THE f DENOTES THE LENGTH OF THE INPUT SEQUENCE AND F DENOTES THE NUMBER OF REAL INPUT FRAMES. THIS TABLE ONLY SHOWS THE AVERAGE VALUE OF EVALUATION METRICS.

Method	f	F	d	MPJPE↓	P-MPJPE↓
PoseFormerV2 [5]	81	3	0	47.5	35.9
Proposed	81	3	1	46.9	**35.8**
Proposed	81	3	2	48.1	35.9
Proposed	81	3	3	**46.8**	35.9
Proposed	81	3	4	46.9	36.1

V. CONCLUSIONS

A method of dilated sampled frames in lifting-based three-dimensional human pose estimation is proposed in this paper. The proposed method exploits the potential solution for choosing discontinuous frames to provide better spatial information. Extensive experiments show the significance of the proposed method, which surpasses the baseline to some content on the public large-scale dataset. In future work, the sample algorithm for choosing the real input frame to keep the low-computing cost and performance simultaneously would be exploited.

ACKNOWLEDGMENT

This result was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

REFERENCES

[1] Can Zhang et al. "Unsupervised pre-training for temporal action localization tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14031–14041.

[2] Jae Shin Yoon et al. "Pose-guided human animation from a single image in the wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15039–15048.

[3] Mang Ye et al. "Collaborative refining for person re-identification with label noise". In: *IEEE Transactions on Image Processing* 31 (2021), pp. 379–391.

[4] Ce Zheng et al. "3d human pose estimation with spatial and temporal transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11656–11665.

[5] Qitao Zhao et al. "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8877–8886.

[6] Catalin Ionescu et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments". In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), pp. 1325–1339.

[7] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[8] Ikhsanul Habibie et al. "In the wild human pose estimation using explicit 2d features and intermediate 3d representations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10905–10914.

[9] Hao-Shu Fang et al. "Learning pose grammar to encode human body configuration for 3d pose estimation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[10] Mir Rayat Imtiaz Hossain and James J Little. "Exploiting temporal information for 3d human pose estimation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 68–84.

[11] Rishabh Dabral et al. "Learning 3d human pose from structure and motion". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 668–683.

[12] Tianhan Xu and Wataru Takano. "Graph stacked hourglass networks for 3d human pose estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 16105–16114.

[13] Yujun Cai et al. "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2272–2281.

[14] Zhiming Zou and Wei Tang. "Modulated graph convolutional network for 3D human pose estimation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11477–11487.

[15] Yujun Cai et al. "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2272–2281.

[16] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. "Ordinal depth supervision for 3d human pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7307–7316.

[17] Jiahao Lin and Gim Hee Lee. "Trajectory space factorization for deep video-based 3d human pose estimation". In: *arXiv preprint arXiv:1908.08289* (2019).

[18] Dario Pavllo et al. "3d human pose estimation in video with temporal convolutions and semi-supervised training". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 7753–7762.