Context-Aware PoseFormer: Single Image Beats Hundreds for 3D Human Pose Estimation – Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

1 Overview

The supplementary material is organized as follows:

- Sec. 2: Broader impacts and limitations.
- Sec. 3: A simple temporal extension of our method.
- Sec. 4: Implementation details.
- Sec. 5: More details and visualization for *Deformable Context Extraction*.
- Sec. 6: Ablation on pre-training tasks for backbones.
- Sec. 7: More visualization (two standard benchmarks & in-the-wild videos).

2 Broader impacts and limitations

Broader impacts. In this paper, we propose a new framework to address the time-intensive issue of existing lifting-based 3D human pose estimation methods. This is done by discovering the long overlooked "free lunch" in the overall lifting-based pipeline – the intermediate visual representations (*i.e.*, multi-scale feature maps) learned by off-the-shelf 2D pose detectors. We show that such representations easily boost pose estimation accuracy (*e.g.*, our single-frame model outperforms 351-frame MHFormer [4]) while bringing no extra costs (no finetuning on 2D detectors is required). We expect our framework to generalize to more research topics, especially other skeleton-sequence-based tasks where long-term temporal modeling may also bring issues (*e.g.*, performance saturation, heavy computation, and the non-causal problem [9]). Retrieving the readily available visual representations from the upstream backbones (that produce the input skeleton) is a promising direction to reduce the temporal reliance of models and further push the performance boundary. Moreover, we hope that our work inspires a wider scope of research – multi-stage tasks in general. Unlike previous lifting-based methods that split the whole lifting pipeline into two independent stages, we engage the intermediate representations from the first stage into the second stage, *i.e.*, two consecutive stages work closer instead of being simply cascaded. Seeking wise collaborations between different stages (in multi-stage tasks) is also a promising research direction.

Limitations. Our single-frame method effectively utilizes the spatial context from 2D pose detectors, achieving comparable or superior precision to multi-frame methods that rely on hundreds of video frames. Furthermore, as demonstrated in Sec. 4.4 of the main paper, we observe that incorporating spatial contextual information improves temporal stability, enhancing consistency and smoothness in the estimated results, even without access to explicit temporal clues. However, for all single-frame methods, including ours, mitigating jitters remains a challenge compared to multi-frame methods that leverage temporal clues. This is primarily due to the non-temporal nature of single-frame methods.

To address this limitation, we naively extend our single-frame approach to a multi-frame variant, allowing capturing temporal dependencies. We present preliminary results in Sec. 3, where we

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

demonstrate that our multi-frame version effectively uses temporal clues to reduce jitters while improving precision. This reveals the potential of our work to further benefit multi-frame approaches. Exploring this direction is part of our future work.

3 A Simple Temporal Extension of Our Method

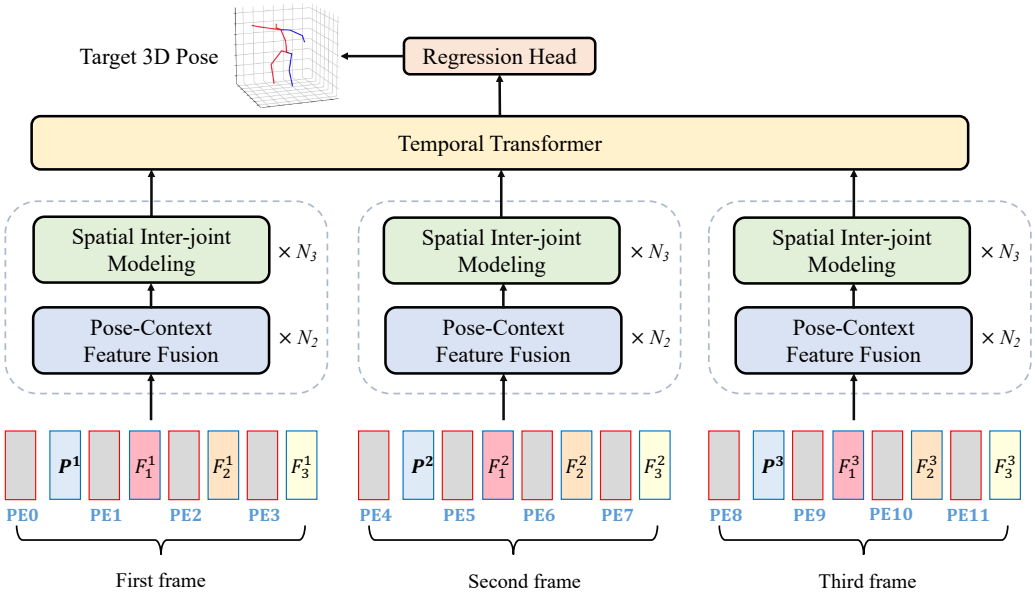

Figure 1: Architecture of the simple temporal extension (3 frames as input) of our method.

In this section, we show that our single-frame method can naturally extend to model temporal dependencies (an overview is in Fig. 1). As illustrated in Sec. 3.2 of the main paper, the *Spatial Inter-joint Modeling* module outputs a feature vector of dimension $(L+1) \times C$ for each joint, where L is the number of multi-scale feature maps and C is a shared projection dimension of the model. We use a *Temporal Transformer* to model temporal correlations of each joint independently. Specifically, for *Temporal Transformer*, the input token number is the total frame number, and each token is of dimension $(L+1) \times C$. Using transformers to build up temporal dependencies is straightforward, and this approach has been adopted by PoseFormer [15], MixSTE [14], etc. The output of the temporal transformer encoder can be denoted by $\boldsymbol{Z}_{Temp} \in \mathbb{R}^{J \times F \times [(L+1) \cdot C]}$, where J is the joint number and F is the frame number. Following PoseFormer, we use 1D convolution to reduce its temporal dimension (gather temporal information) and a linear layer to obtain the final estimated 3D pose $\boldsymbol{y} \in \mathbb{R}^{J \times 3}$.

Table 1: We compare the (short-term) temporal extension of our small model (CA-PF-S) with PoseFormer. MPJPE: Mean Per Joint Position Error, the precision metric. MPJVE: Velocity Error, the temporal smoothness metric. The results are reported on Human3.6M (in millimeters).

Model	Frame	MPJPE ↓	MPJVE ↓
PoseFormer	1	53.2	13.7
	3	51.0	7.1
	9	49.9	4.8
	81	44.3	3.1
CA-PF-S	1	44.7	8.5
	3	44.2$_{\downarrow 0.5}$	4.8$_{\downarrow 3.7}$
	9	43.4$_{\downarrow 1.3}$	3.4$_{\downarrow 5.1}$

Quantitative results. Due to limited computational resources, the experiments in this section are conducted with a small variant of our model, referred to as "CA-PF-S", which has fewer FLOPs compared to our full model "CA-PF" presented in Sec. 4.1 of the main paper. To show the gains in precision and temporal smoothness from temporal modeling, we report two metrics, MPJPE

(Position Error, the precision metric) and MPJVE (Velocity Error, the temporal smoothness metric), on Human3.6M [3]. We compare it with PoseFormer [15] and the results are in Table 1: **First,** increasing the number of input frames brings consistent improvements in both precision and temporal smoothness. For example, by using only 3 video frames, the MPJVE of our method decreases from 8.5 to 4.8mm (a **43.5**% reduction), and the MPJPE is reduced by 1.1%. This indicates that even short-term temporal modeling largely mitigates jitters in estimated results and further improves precision upper bound. **Second,** considering the same number of input frames, our CA-PF-S consistently outperforms PoseFormer in terms of both MPJPE and MPJVE. Moreover, our 3-frame CA-PF-S has already achieved superior MPJPE to 81-frame PoseFormer, and our 9-frame CA-PF-S achieves comparable MPJVE with 81-frame PoseFormer. The results verify the two-fold benefits of spatial contextual clues from 2D pose detectors – accuracy and temporal stability.

Based on the results above, we expect that our method can be successfully extended to model even longer-term temporal dependencies (*e.g.*, 81 video frames) to further boost precision and temporal smoothness. We provide visualization of in-the-wild videos in Sec. 7 to show the advantage of our method in temporal consistency (stability).

4 Implementation Details

2D pose detector settings. The overall pipeline of our method includes two parts: an off-the-self 2D pose detector and a lifting model. For the first stage, the 2D pose detector is pre-trained on the COCO [6] dataset, without finetuning on the 2D poses from 3D pose estimation datasets, *i.e.*, Human3.6M [3] and MPI-INF-3DHP [8]. We use 256×192 resolution for input images. For 2D-to-3D lifting, the weights of pre-trained 2D detectors are frozen, *i.e.*, no finetuning on the 3D task is needed either. This approach makes our method preferably flexible – our method is compatible with a wide range of off-the-shelf (pre-trained) 2D pose detectors. We show in Sec. 4.1 of the main paper that our method gains consistent improvements by increasing the capability of 2D pose detectors. In the future, we may leverage more advanced 2D pose detectors to further improve the performance upper bound.

Lifting model settings. Our lifting model includes three basic modules: *Deformable Context Extraction*, *Pose-Context Feature Fusion* and *Spatial Inter-joint Modeling*. The layer number of each module is set to 4, following PoseFormer [15]. The hidden dimension (a shared projection dimension C) of the model is 128. We use 8 heads in self-attention for transformer layers.

Training details. The experiments are conducted on a single NVIDIA RTX 3090 GPU. Our lifting model is trained using the AdamW optimizer [7] for 50 epochs with a batch size of 512. The initial learning rate is 6.4e-3 with an exponential learning rate decay schedule, and the decay factor is 0.99. *We will release the source code and the trained models of our method upon acceptance of the paper.*

5 More Details and Visualization for Deformable Context Extraction

More details. The *Deformable Context Extraction* module extracts informative joint-centric context features from feature maps using deformable attention [16]. The sampling points of each attention head are initialized in different directions (w.r.t. the reference joint) to promote learning diverse contextual clues from images. To prevent overly aggressive updates on sampling offsets (*e.g.*, they may go outside the image), the learning rate of linear layers that generate sampling offsets is set to 6.4e-4 ($^1/_{10}$ of that for other layers in the lifting model). We use 4 heads for deformable attention.

Visualization of learned sampling points on consecutive video frames. In Fig. 2, a subject in the Human3.6M test set raises his right arm where severe self-occlusion occurs, and the 2D pose detector fails to localize the right wrist (blue dots are detected results, and green dots are ground truth). We find that most sampling points (gray dots) are concentrated on the upper body of the subject. More interestingly, despite the unreliable 2D joint detection (reference points), some learned sampling points attempt to approach the ground truth. We indicate the sampling points that gain larger attention weights with higher brightness (we decrease the brightness of images for better visual effects). Note that we do *not* train sampling points using ground truth. This indicates that our adaptive context extraction strategy can learn informative contextual features based on the visual cues of images despite bad sampling references (*i.e.*, false joint detection), which helps reduce uncertainty brought by imperfect 2D pose detectors.

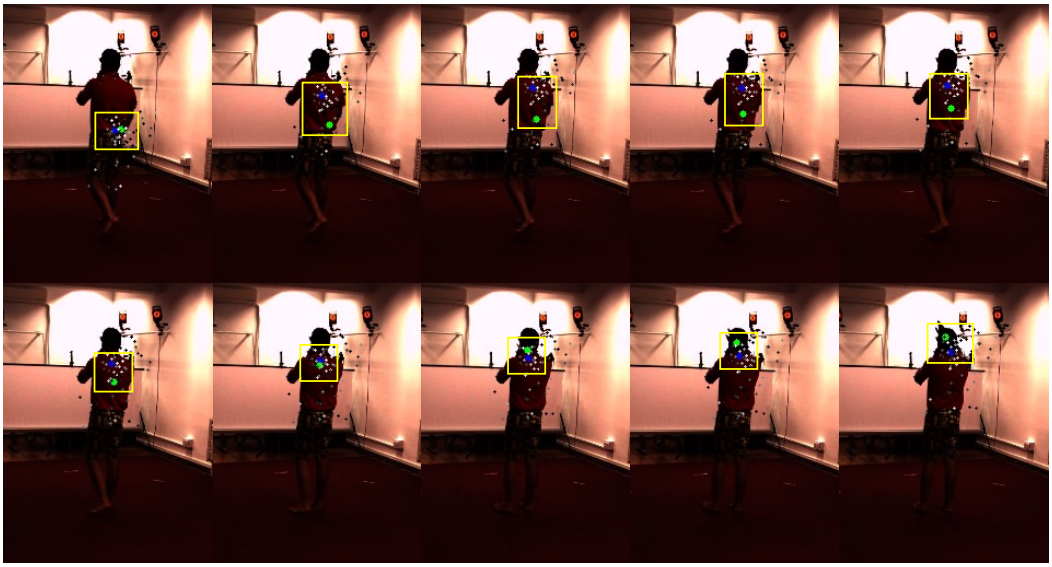

Figure 2: Visualization of consecutive frames on the Human3.6M test set where severe self-occlusion occurs. *Deformable Context Extraction* learns sampling points that attempt to discover ground truth joints given false 2D joint detection as reference.

6 Ablation on pre-training tasks for backbones

ImageNet [10] pre-trained backbones (*e.g.*, ResNet [2]) profit a series of downstream tasks, including object detection [5], segmentation [1], and 2D human pose estimation [12, 13], yet this seems not applicable to 3D human pose estimation. In Table 2, we replace COCO pre-trained backbones in our method with ImageNet classification pre-trained ones, showing a remarkable performance drop. This should be attributed to the large gap between the pre-training task (image classification) and the downstream task (3D human pose estimation). Our method provides a starting point to leverage visual representations from (pre-trained) 2D human pose detectors, while challenges still remain: First, is there any better way to use visual representations from pre-trained backbone networks? Second, as the 2D computer vision community has gained a lot from emerging pre-training methods (from supervised to unsupervised), can we design pre-training tasks that are more appropriate for 3D human pose estimation (more generally, 3D perception)? They are also potential research topics in the future.

Table 2: Ablation study on pre-training tasks with different backbones. MPJPE (mm) is reported on Human3.6M.

Backbone	Pre-training	MPJPE ↓
ResNet-50	2D Pose	45.0
	Image Class.	51.4$_{\uparrow 6.4}$
HRNet-32	2D Pose	41.4
	Image Class.	45.8$_{\uparrow 4.4}$
HRNet-48	2D Pose	39.8
	Image Class.	43.9$_{\uparrow 4.1}$

7 More Visualization

Static results on two standard benchmarks. We provide more qualitative results on Human3.6M (Fig. 3) and MPI-INF-3DHP (Fig. 4). Our single-frame method obtains reliable estimated results in hard cases, *e.g.*, self-occlusion, and rare poses, compared to state-of-the-art multi-frame methods such as 351-frame MHFormer [4] and 81-frame P-STMO [11].

Temporal results on in-the-wild videos. To show the advantage in temporal stability of our method, we provide visualization of in-the-wild videos in Fig. 5. We compare the 9-frame temporal extension of our model (more details are in Sec. 3) with PoseFormer [15]. We choose two sets of consecutive video frames where the 2D joint detection fails due to confusing clothing (the left column) or self-occlusion (the right column). Since PoseFormer only accepts 2D joints as input, its estimated 3D poses are sensitive to the noise of input 2D poses. Therefore, it infers unreliable 3D poses given

Figure 3: Qualitative comparison with MHFormer (351 frames) [4] and our context-agnostic counterpart (please refer to Sec. 4.3 in the main paper for more details) on Human3.6M. Our method obtains reliable results despite severe self-occlusion, which may cause false 2D joint detection. Notable parts are indicated by arrows.

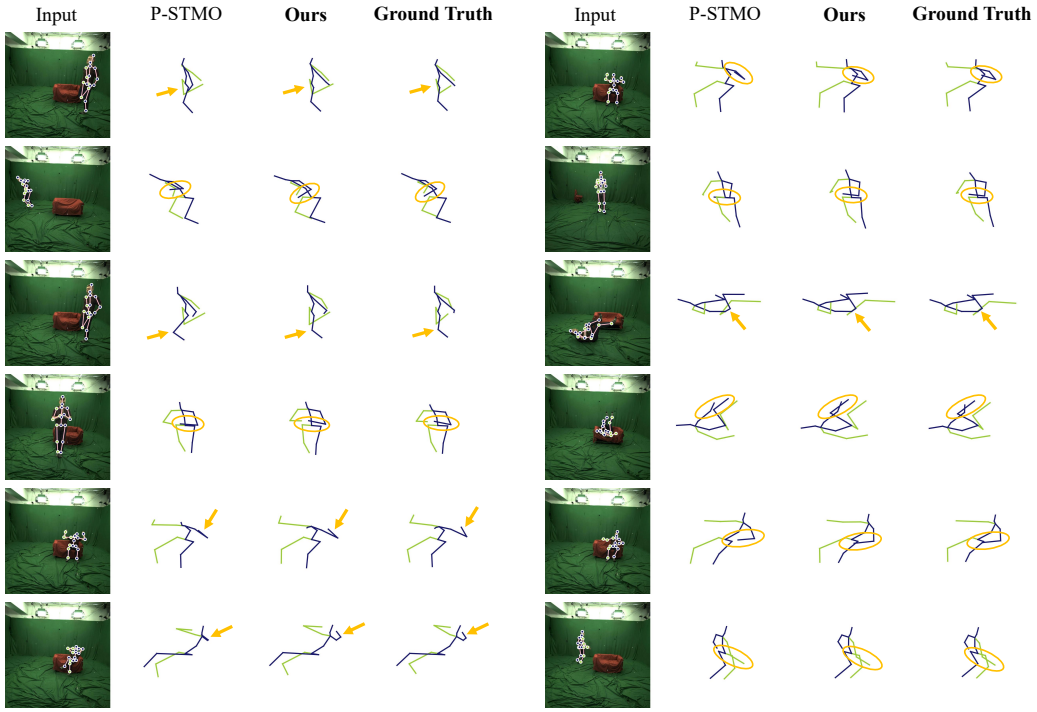

Figure 4: Qualitative comparison with P-STMO (81 frames) [11] on MPI-INF-3DHP. Our method infers correct results given rare poses (*e.g.*, the subject is lying on the ground and relaxing on the couch). Notable parts are indicated by arrows or circles.

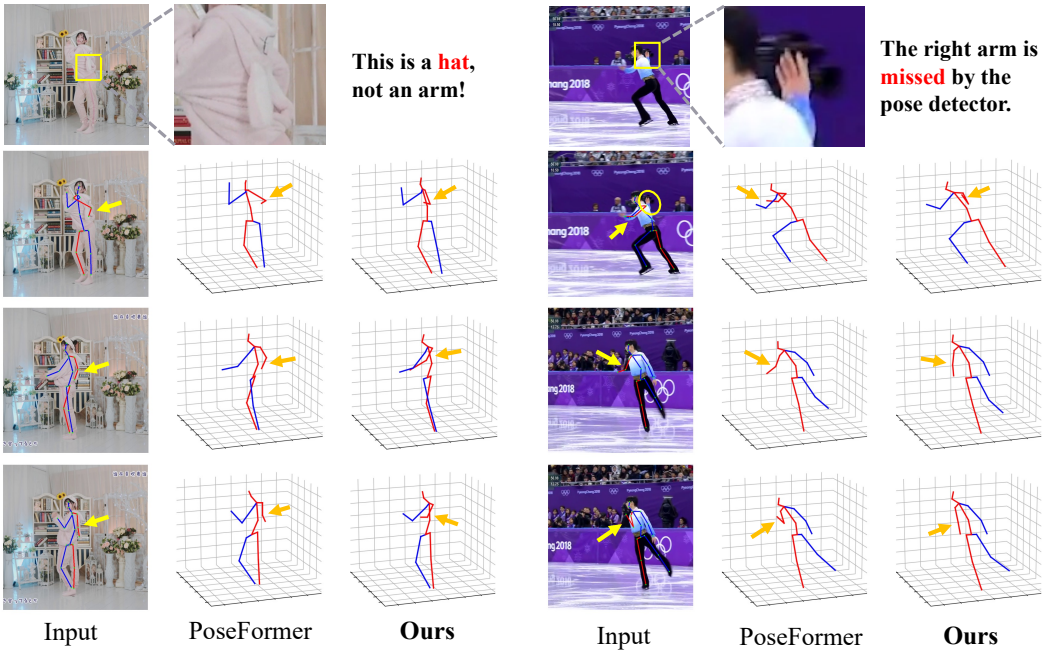

Figure 5: Comparison between PoseFormer and our small model variant on in-the-wild videos. The 2D pose detector fails to localize 2D joints, given *confusing clothing* (left) and severe *self-occlusion* (right). In such hard cases, our method is more robust and enjoys better temporal consistency. False joint detection is indicated by yellow arrows, and the corresponding 3D joint estimation is indicated by orange arrows.

false 2D detection. On the contrary, in addition to the positional information provided by 2D joint locations, we also leverage spatial contextual clues from images to localize joints in 3D. Thus our method shows more robust (stable) and smooth results despite noisy input 2D joints. We provide the source mp4 file of both video clips in the supplementary material.

References

[1] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[2] 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.

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014.

[4] 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, June 2022.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[9] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[11] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022.

[12] 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.

[13] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021.

[14] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022.

[15] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11656–11665, October 2021.

[16] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.