2022 The 22st International Conference on Control, Automation and Systems (ICCAS 2022)
BEXCO, Busan, Korea, Nov. 27∼Dec. 01, 2022

Combination of Deep Learner Network and Transformer for 3D Human Pose Estimation

Tien-Dat Tran[1*], Xuan-Thuy Vo[2], Duy-Linh Nguyen[3], and Kang-Hyun Jo[4*]

[1,2,3,4]Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, 44610, Korea

Email: (tdat,xthuy)@islab.ulsan.ac.kr, ndlinh301@mail.ulsan.ac.kr, acejo@ulsan.ac.kr * Corresponding author

Abstract: Deep neural networks (DNNs) have attained the maximum performance today not just for human pose estimation but also for other machine vision applications (e.g., semantic segmentation, object detection, image classification). Besides, the Transformer shows its good performance for extracting the information in temporal information for video challenges. As a result, the combination of deep learner and transformer gains a better performance than only the utility one, especially for 3D human pose estimation. At the start point, input the 2D key point into the deep learner layer and transformer and then use the additional function to combine the extracted information. Finally, the network collects more data in terms of using the fully connected layer to generate the 3D human pose which makes the result increased precision efficiency. Our research would also reveal the relationship between the use of the deep learner and transformer. When compared to the baseline-DNNs, the suggested architecture outperforms the baseline-DNNs average error under Protocol 1 and Protocol 2 in the Human3.6M dataset, which is now available as a popular dataset for 3D human pose estimation.

Keywords: Deep neural network, Transformer, 3D Human pose estimation.

1. INTRODUCTION

In today's world, 3D human posture estimate plays an essential role in computer vision, fulfilling a variety of goals such as human re-identification [1], [2], activity recognition [3], [4], 2D human pose estimation [5], [6], and 3D human pose estimation [7], [8]. The fundamental purpose of the human posture is to identify bodily portions for human body keypoints. The importance of deep neural network also use spatial information that decrease the error of 3D key point in regression stage. As a result, the focus of this study will be on how to teach the network to pay better 3D accurate pose estimation.

According to recent developments, deep convolutional neural networks have lately achieved outstanding performance. Before raising the resolution, most existing techniques route the input through a network, after that apply the 3D human pose estimation on the 2D result, which show in Fig.1. The 3D network take the series of 2D keypoint as the input and is typically made up of high-to-low resolution subnetworks connected in series. Hourglass [9], for example, uses a symmetric low-to-high technique to recover high resolution. SimpleBaseline [10] uses a few transposed convolution layers to build high-resolution representations. Dilated convolutions are also used to increase the last layers of a high-to-low resolution network (such as VGGNet or ResNet) [11], [12]. In other hand, some network remain the High-resolution network to make the 2D keypoint better[13] or also use attention mechanism[14] to make the network better in AP (Average Precision) for 2D human pose estimation

Deep neural network has now encoded major advancements in human posture [15], [16]. However, these networks face numerous obstacles. To begin, how can the accuracy of various types of networks be improved (For example, a real-time network or a network that measures correctness.) Second, it is common to need to check the

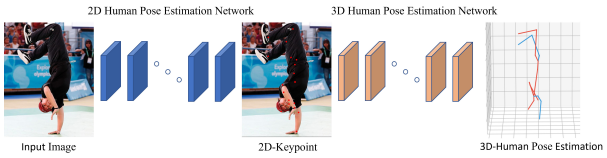

Fig. 1. Summary the process from input image to 3D human pose estimation

speed of a network while updating or modifying it. Finally, the current network must increase accuracy while remaining as fast as possible. This study examines a one-of-a-kind network as well as the speed and accuracy of the combination of transformer and deep learner network. Using and not using the transformer is the subject of the proposed experiment. The experiment also differs from the Simple Baseline [17] that it does not use the transformer mechanism. The Simple Baseline also experiments for only deep learner network, which used a lot of layer of fully-connected layer.

The proposed technique was used to create a 3D pose network, which showed a significant improvement in Mean Per Joint Position Error (MPJPE). The proposed network, which is based on the deep learner network [17], aims to improve the evolution theory for the 2D keypoint by using the gene method. By employing a new transformer inside the deep learner network, the network keeps the MPJPE higher while minimizing the implementation cost. In addition, the number of parameters was reduced, which resulted in a faster network. To further comprehend transformer inside, the suggested network decrease the error 0.3 points in Average MPJPE while decreasing the number parameters. This study presents a novel 3D network that can quickly respond to a wide range of challenges in a variety of applications, including object recognition, picture classification, and human position estimate. The suggested method uses do not

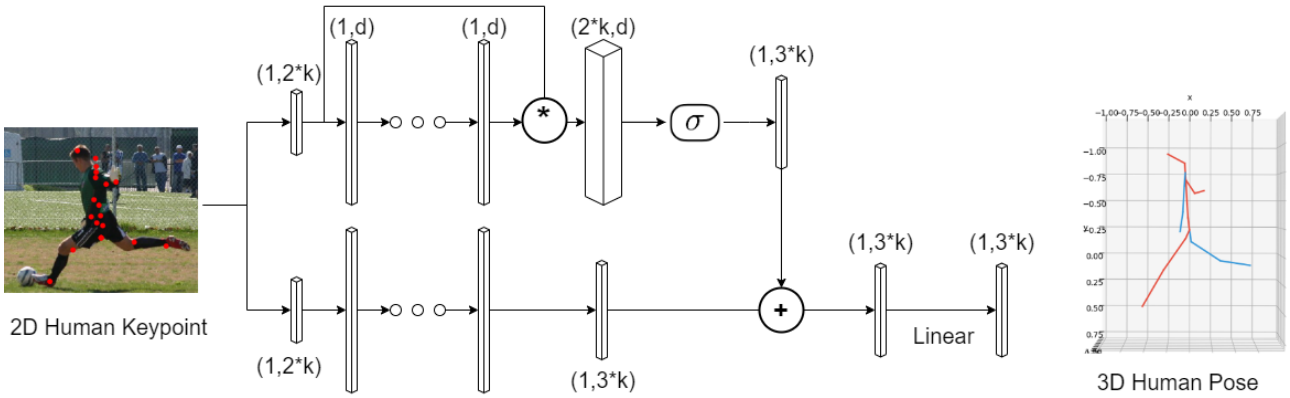

Fig. 2. Illustrate from 2D to 3D human pose estimation. The proposed method separated the network into two stages, the first stage take ideal from transformer and the second for deep layer network.

use 2D ground truth for joint human posture predictions for the fair competition.

2. METHODOLOGY

2.1 Network architecture

Our system utilized a backbone comprised of HRNet-W32 and HRNet-W48 [13], as the main architecture for 2D human pose estimation. After that, a series of key-point for humans will apply for making 3D which can show in Fig.2. First, from $2 \times k$, in which k is the number of key-point, the proposed architecture is divided into two stages. The first stage takes the ideal from transformer [18]. we used the fully connected layer to extract the information of $2 \times k$ keypoint into d-dimension which is set at 1024 and apply 8 layers of the fully connected layer. After that, our network utilizes element multiplication for $2 \times k$ and d by a skip connection. Hence we apply sigmoid to extract the probability and one more time fully connected to make $3 \times k$. In the second stage, the tensor traverses each pillar layer, utilizing only the fully connected layer to extract the information of $2 \times k$ into the d-dimension similar to the first stage. However, in the second stage, the proposed architecture did not apply the sigmoid so the information is just about the key point. Finally, we used an additional function for stage 1 and stage 2 to combine information from the deep learner network and transformer network followed by a final fully connected layer (Linear) to make the 3D information.

2.2 Loss Function

The Heat maps which generate from the last layer of 3D human pose network. We apply the baseline MPJPE loss to minimize the error. By $m = \{mj\}\, J = 1^J$, where $Xj = (xj, yj)$ is the geographical harmonize of the jth body joint for each image. The value of heat map for Ground-truth Hj is then constructed using the Gaussian distribution and the mean aj with variance $\sum$ as shown below.

$$H_j(p) \sim N(a_j, \sum) \tag{1}$$

where $\mathbf{p} \in \mathbf{R^2}$ demonstrate the coordinate, and $\sum$ is experimentally decided as an identity matrix $\mathbf{I}$. The last layer of the neural architecture forecast J heat maps, *i.e.,* $\hat{S} = \left\{\hat{S}j\right\} j = 1^J$ for J body joints. A L2 loss function is defined by the mean of MPJPE, which is calculated as follows:

$$L = \frac{1}{MJ} \sum_{m=1}^{M} \sum_{j=1}^{J} \left\| S_j - \hat{S}_j \right\|^2 \tag{2}$$

M denotes the number of selected in the training process. Using 3D pose data from the last layer or backbone architecture, the trained network generated predict 3D joint maps using ground-truth 3D pose.

3. EXPERIMENTS

3.1 Experiment Setup

3.1.1 Dataset

The biggest 3D human pose estimation benchmark with precise 3D labels is called Human 3.6M (H36M), and it comprises of 3.6 million photos taken by four synchronized cameras at 50 frames per second. Seven professionals are engaged in 15 daily tasks including "waiting", "smoking", and "posing". By adding the subject ID to S, we indicate a group of data. For example, S15 indicates data from subjects 1 and 5. Five individuals (S1, S5, S6, S7, and S8) are utilized for training, and two subjects (S9 and S11) are employed for evaluation, all in accordance with the standard procedure from earlier works [28]. A single model is trained for all actions using the frames from all viewpoints.

With millimeter-based Mean Per Joint Position Error (MPJPE), we assess the model's performance. The use of two common assessment methodologies. While protocol 2 (P2) first aligns the ground-truth 3D poses with the predictions via a rigid transformation, protocol 1 (P1) computes MPJPE immediately, which is the euclidean distance between grouth-truth and predicted keypoint. Protocol P1 eliminates the impact of the first stage model by using inputs of ground truth 2D key points. Moreover,

Table 1. Quantitative comparisons with state-of-the-art on Human3.6M dataset under protocol #1 and protocol #2 for fully-supervised methods. Bold number is the best performance in each case

.

Protocol # 1	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [19] ICCV'17	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang *et al.* [20] AAAI'18	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
Yang *et al.* [21] CVPR'18	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Pavlakos *et al.* [22] CVPR'18	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Moon *et al.* [23] CVPR'19	51.5	56.8	51.2	52.2	55.2	47.7	50.9	63.3	69.9	54.2	57.4	50.4	42.5	57.5	47.7	54.4
Liu *et al.* [24] ECCV'20	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0	60.4	71.1	51.5	50.1	54.5	40.3	**43.7**	52.4
Xu *et al.* [25] CVPR'21	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
Li *et al.* [17] CVPR'20 †	47.0	**47.1**	49.3	50.5	**53.9**	**58.5**	48.8	45.5	**55.2**	68.6	50.8	**47.5**	53.6	42.3	45.6	50.9
Our	**45.0**	48.3	**46.6**	**49.8**	54.0	59.0	48.7	**45.1**	57.7	**68.2**	**49.0**	48.2	**52.9**	41.0	45.1	**50.6**
Protocol # 2	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [19] ICCV'17	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang *et al.* [20] AAAI'18	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Pavlakos *et al.* [22] CVPR'18	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
Yang *et al.* [21] CVPR'18	**26.9**	**30.9**	**36.3**	39.9	43.9	47.4	**28.8**	**29.4**	**36.9**	58.4	41.5	**30.5**	**29.5**	42.5	**32.2**	**37.7**
Sharma *et al.* [26] ICCV'19	35.3	35.9	45.8	42.0	40.9	52.6	36.9	35.8	43.5	51.9	44.3	38.8	45.5	**29.4**	34.3	40.9
Cai *et al.* [27] ICCV'19	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	33.2	39.0
Liu *et al.* [24] ECCV'20	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
Li *et al.* [17] CVPR'20 †	34.5	34.9	37.6	39.6	**38.8**	45.9	34.8	33.0	40.8	51.6	38.0	35.7	40.2	30.2	34.8	38.0
Our	34.1	36.0	36.4	39.9	39.4	**45.0**	35.9	32.8	43.1	52.1	**37.3**	36.6	39.7	30.2	35.8	38.3

the proposed paper also use an industrial dataset that included 4 videos. The video record the people's action in the industrial laboratory with 9980 for a total number of frames. Our quantitative results for Fig.3 and Fig.4 take from this dataset.

3.1.2 Implementation details

. The monocular camera is implemented on training process. We train 2D human pose estimation with the input size is 384×288 and the output heatmap is same. For the 3D human pose network, the batch size was set at 24. The total number of epochs was stick 210. We set the learning-rate at 0.001 the learning decade factor is 0.1 at the 170-th and 200-th epoch. All the experimental research are implemeted by using the Pytorch framework and tested on the H36M datasets. The Adam optimizer [29] and the momentum is 0.9 was employed.

The proposed architecture was trained using CuDNN 7.3 and CUDA 10.2 on a single NVIDIA GTX 1080Ti GPU.

3.1.3 Human3.6M datasets result

Our result was estimated on the Human 3.6M test dataset. The value of error in the proposed perspective which show in Table 1 gets better than other research standards in Protocol #1 in some activities such as Direction - 45.0, Eating - 46.6, Greeting - 49.8, Purchases - 45.1, SittingDown - 68.2, Smoking - 49.0, WalkDog 52.9 and the average error is 50.6 which is the best for the comparison research. However, For Protocol #2 our research only takes third place with an average error is 38.3. Photo and Smoking take the best accuracy with the lowest error of 45.0 and 37.3 respectively. Smoking is the action that our architecture detect well in all case. In Table 1, † is mean we only use the result for backbone network training for the Human 3.6M dataset with out the method to make the dataset more bigger or using multy view for fair compatition.

Human pose estimation, like many other modern designs, has a variety of problems that need to be solved. The pictures' concealed joints, which were challenging to train for and predict, were the first problem. Second, joints in the human body must be accurately eliminated from low-resolution human images. The photographs that follow show crowd situations, when it is usually challenging to pinpoint where each participant's joints are located. Last but not least, there is a dearth of data on photos with missing pieces for assessing human postures.

4. CONCLUSION

This paper presents a novel 3D deep learner network combined with the transformer network is introduced and achieves a better result for the monocular condition when compared with the baseline. From the 2D ground-truth keypoint, the network generate the 3D information by using both fully connected network and transformer, which show that it outperform for the error and the number of parameter compared with the baseline. For the future work, the network can improve performance by using temporal information which affects much in the error for 3D human pose estimation. The other work is to find out how to solve the challenge of human posture estimation, which construct the network hard to gain superior performance. In addition, applications and environments for our research such as those developed in mobile devices, which need small of number parameters, can progress with the proposed ideal.

ACKNOWLEDGEMENT

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] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning

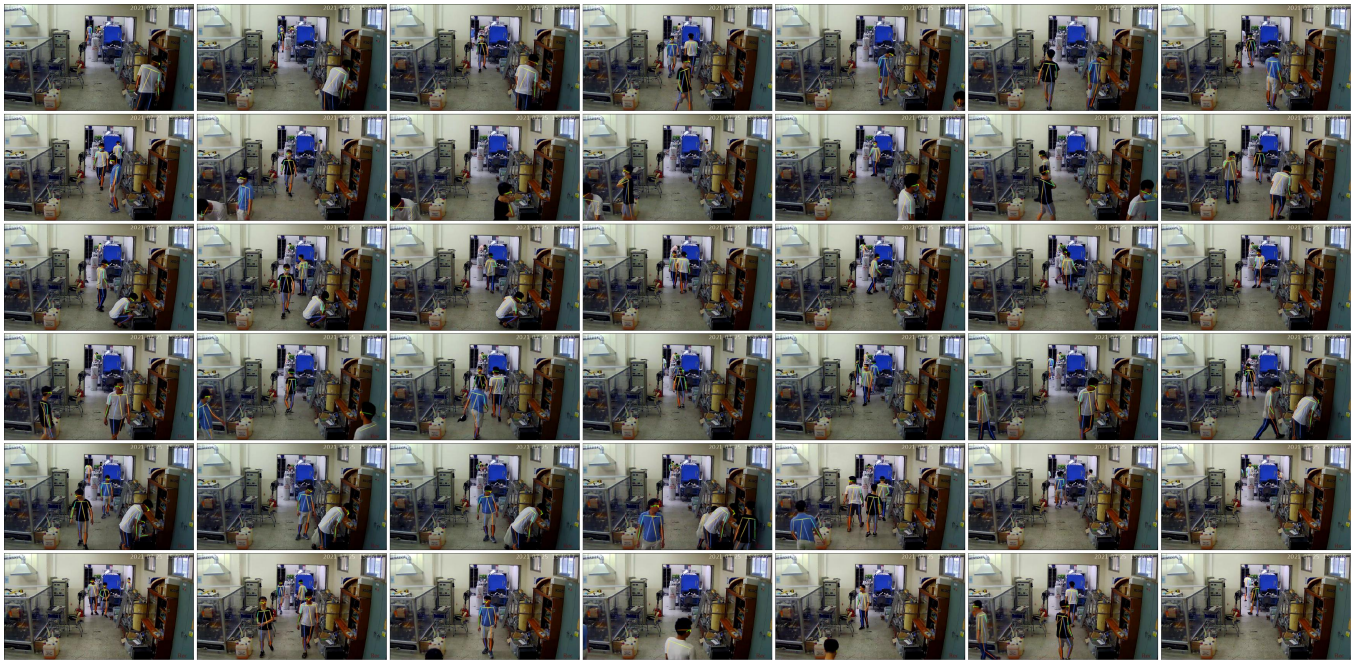

Fig. 3. Qualitative result for 2D human pose tracking in video 1 of industrial dataset

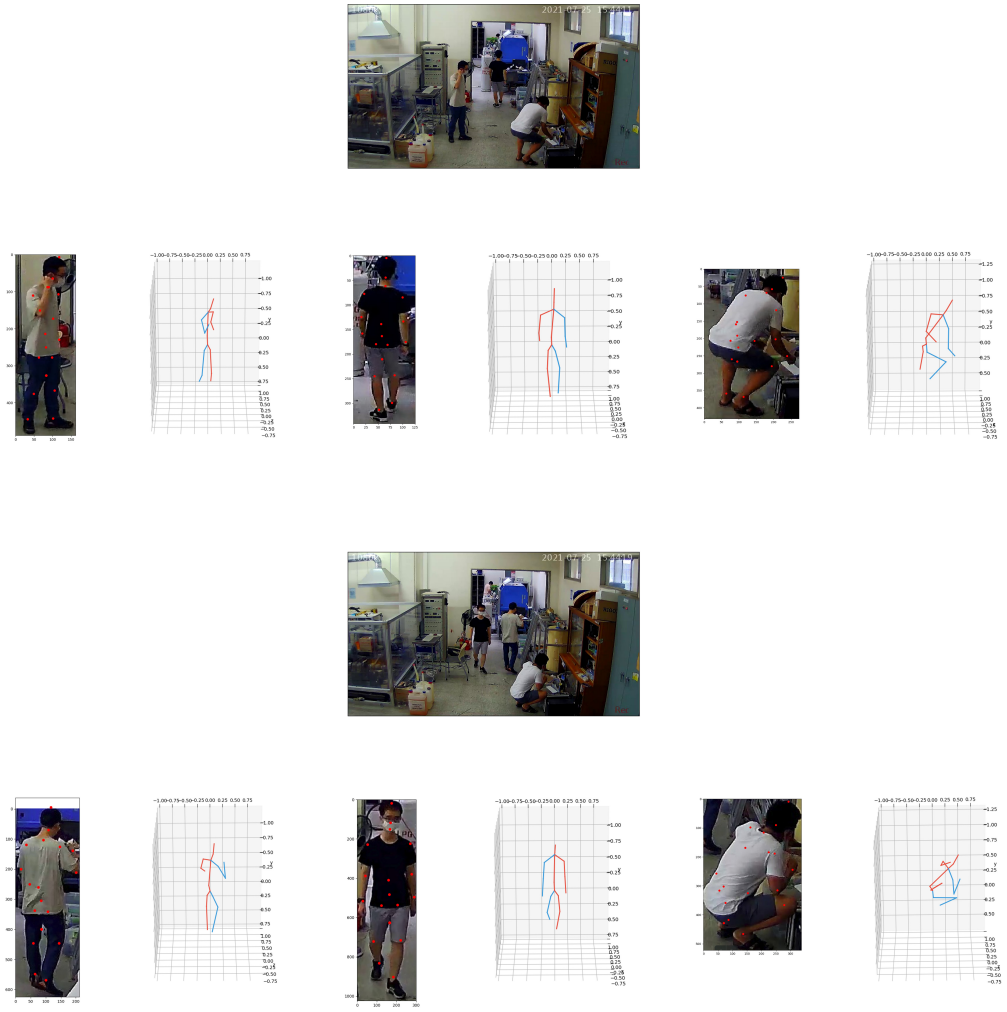

Fig. 4. Qualitative result for 3D human pose estimation in frame number 1308 and frame number 1469 of industrial dataset - video 1

using privileged information," *CoRR*, vol. abs/1904.05005, 2019. [Online]. Available: http://arxiv.org/abs/1904.05005

[2] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision (ACCV)*, 11 2012, pp. 31–44.

[3] Z. Hussain, M. Sheng, and W. E. Zhang, "Different approaches for human activity recognition: A survey," 2019.

[4] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 48–53, Jan 2010.

[5] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," 2016.

[6] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," 2017.

[7] C. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5759–5767.

[8] S. Li, L. Ke, K. Pratama, Y. Tai, C. Tang, and K. Cheng, "Cascaded deep monocular 3d human pose estimation with evolutionary training data," *CoRR*, vol. abs/2006.07778, 2020. [Online]. Available: https://arxiv.org/abs/2006.07778

[9] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016. [Online]. Available: http://arxiv.org/abs/1603.06937

[10] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *CoRR*, vol. abs/1804.06208, 2018. [Online]. Available: http://arxiv.org/abs/1804.06208

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019.

[14] T.-D. Tran, X.-T. Vo, M.-A. Russo, and K.-H. Jo, "Simple fine-tuning attention modules for human pose estimation," in *International Conference on Computational Collective Intelligence*. Springer, 2020, pp. 175–185.

[15] G. Moon, J. Y. Chang, and K. M. Lee, "Posefix: Model-agnostic general human pose refinement network," 2018.

[16] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," 2016.

[17] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3d human pose estimation with evolutionary training data," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6172–6182.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[19] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," 2017. [Online]. Available: https://arxiv.org/abs/1705.03098

[20] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," 2017. [Online]. Available: https://arxiv.org/abs/1710.06513

[21] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5255–5264.

[22] G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal depth supervision for 3d human pose estimation," 2018. [Online]. Available: https://arxiv.org/abs/1805.04095

[23] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10 132–10 141.

[24] K. Liu, R. Ding, Z. Zou, L. Wang, and W. Tang, "A comprehensive study of weight sharing in graph networks for 3d human pose estimation," in *ECCV*, 2020.

[25] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang, "Deep kinematics analysis for monocular 3d human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 896–905.

[26] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, "Monocular 3d human pose estimation by generation and ordinal ranking," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2325–2334.

[27] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2272–2281.

[28] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7745–7754.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.