Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

An Effective 3D Human Pose Estimation Method Based on Dilated Convolutions for Videos

Hong Liu
Shenzhen Graduate School
Peking University
Shenzhen, China
hongliu@pku.edu.cn

Congyaxu Ren
Shenzhen Graduate School
Peking University
Shenzhen, China
rencyx@pku.edu.cn

***Abstract*—3D human pose estimation is a challenging problem due to the diversity of poses, human appearance, clothing, occlusion, etc. In this work, we split the problem into two stages, 2D human pose estimation and 3D pose recovery, and address it by a network based on dilated convolution for videos. We introduce pruning layer to prevent overfitting, which performs better than dropout, because the strategy of pruning is not to drop nodes randomly, but to choose the lower-weight ones. We also employ quantization to accelerate it in the smart shop environment, which gains a trade off between performance and computational complexity, and the accuracy loss is in an acceptable range. In addition, labeled human pose datasets are so limited and expensive especially for 3D poses, we propose a 3D human pose dataset named HRI-I in a smart shop environment, which contains more than 16k poses, 26 people and 6 scenarios of walking, hunkering, fetching objects, etc. We train and test our model on the HumanEva-I, Human3.6M and our proposed HRI-I, it demonstrates that the proposed method is efficient and effective.**

***Index Terms*—3D human pose estimation, smart shop, HRI-I dataset**

I. Introduction

Human pose estimation is a basic topic for computer vision, and can be applied to many scenarios, such as monitoring, intelligent mobile robots, virtual reality and so on [1] [2]. Many efforts have been devoted to tackle this task from different perspectives, including 2D and 3D human pose estimation, inferring poses by a single image or videos, considering one camera or cross view fusion [3] [4] [5] [6].

Our work focus on 3D human pose estimation and its application on smart shop. We split this task into two stages, firstly, leverage advanced 2D human pose estimator [7] [8] to gain the accurate 2D joints and skeletons, and then frames and their corresponding joints are fed into the network based on dilated convolution to gain the predicted 3D human pose results.

For 3D human pose estimation problem, researchers have to be faced with ambiguity and incoherency [9]. A 2D pattern has infinite corresponding 3D structures, as Fig. 1 shows [10]. Our humans are able to infer the 3D structure of a subject easily because of our massive prior experience and knowledge, however, it is hard for models to predict the "correct" result– the 3D pose which owns the largest probability. The difficulty exists in the loss of the third dimension that machines cannot predict the depth and size of a subject in a single picture or videos captured by a common monocular camera, because the subject could be large and far, or small and close. Some previous researches attempted to solve the problem by leveraging temperal information, which enables models to infer 3D structures by context. Recurrent neural networks(RNN) is an effective architecture to utilize temperal information, however, it is not paralleled and occupies a large computation complexity [11]. Thus, we employ the network based on dilated convolutions to exact long-term temporal information and reduce the ambiguity and incoherency.

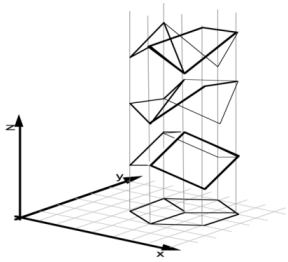

Fig. 1. Any planar line-drawing is geometrically consistent with infinitely many 3D structures. [10]

For human pose estimation task, labeled data is so limited and expensive especially for 3D poses. Thus, we tackle the issue by semi-supervised learning method and proposing a 3D human pose dataset. Inspired by [11], we employ 2D-3D consistency to be the weak supervision, which means 2D pose inferred by the predicted 3D pose should be the same with the original 2D poses. In addition, we propose a 3D human pose dataset named HRI-I to accommodate the smart shop environment in the wild.

In our work, to prevent overfitting, we introduce pruning layer to enable the model to be more sparse. Compared with dropout [12], it does not drop nodes randomly, but set lower weights to zeros. It performs better as a result of higher values usually playing a more important role in the process of exacting features. What is more, to accelerate and gain the

This work is supported by National Natural Science Foundation of China (NSFC U1613209), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No. ZDSYS201703031405467), National Engineering Laboratory for Video Technology - Shenzhen Division.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

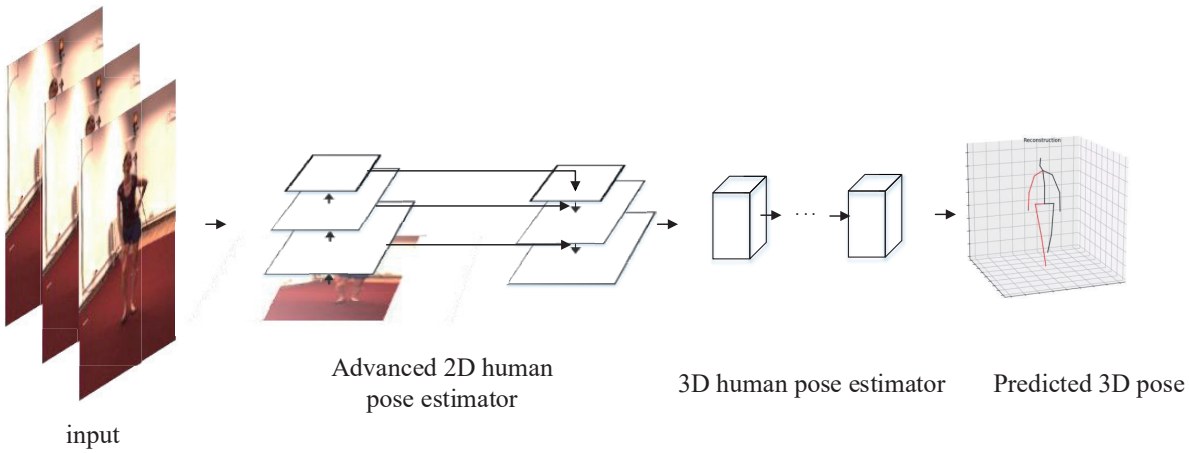

Fig. 2. Pipeline of our model. We set the estimated 2D joints as our input, which are predicted by the advanced 2D human pose estimator. Our model employs ResNet architecture, which is comprised of blocks.

trade off between performance and computational complexity, we employ quantization [13] for float variables, which resize the model to be approximately a quarter of the original one. Finally, we propose a 3D human pose dataset named HRI-I to evaluate its performance under the circumstance of a smart shop, which contains more than 16K poses, 26 people and 6 scenarios of hunkering, walking, etc.

II. Related Work

A. Deep learning based method for pose estimation

As the development of deep learning, many researchers concentrate on deep learning based methods to estimate 2D poses. Toshev et al [14] formulated this task as a deep neural networks based regression problem towards body joints at the first time, they presented a cascade of such regressors to capture context and reason about pose in a holistic manner. Newell et al [15] proposed a stacked hourglass network for human pose estimation, in which features were captured and processed across all scales and spatial information was inferred in different levels. In recent years, 3D human pose estimation is focused by more researchers. Pavlakos et al [16] extended the stacked hourglass architecture from 2D to 3D poses for single image, they employed a coarse-to-fine scheme and predicted per voxel likelihoods for each joint. Zhou et al [4] leveraged a deep fully convolutional network to predict the uncertainty maps of the 2D joint locations over image sequences and the Expectation-Maximization algorithm to estimate 3D poses. Bogo et al proposed the first method to automatically estimate the 3D human pose as well as the 3D shape from a single unconstrained image. Besides, single image or videos, single camera or cross view fusion and some other topics are also focused on recently. Qiu et al [5] estimated 3D pose from multi-view images by incorporating multi-view geometric priors.

B. Estimating 3D pose from 2D joints

3D human pose estimation is a challenging problem due to the diversity of poses, such as human appearance, clothing, occlusion, viewpoints, etc. Thus, we usually attempt to split this task into easier ones, 2D pose detection and 3D pose recovery [17]. Agarwal et al [18] recovered poses by direct nonlinear regression instead of shape descriptor vectors. As the trend of deep learning, many researchers leverage deep learning based methods to tackle the problem. Jiang [19] used estimated 2D keypoints to predict 3D human pose from single images based on exemplar method, it was able to query large pose datasets and effective for 3D reconstruction for complex pose. Akhter and Black [20] defined a parametrization of body pose and estimated 3D human pose from 2D joint locations under the constraints of joint-limits. Mehta et al [21] extracted bounding box from 2D, employed CNN-based model to regress and recovered global 3D position in non-cropped images.

C. Using temporal information

Since recovering 3D poses from a single image cannot take advantage of temporal information to decrease the ambiguity and incoherency, many approaches tend to leverage it and predict 3D poses from a video to improve the performance. LSTM sequence to sequence models are widely used in natural language processing and video retrieval, which require context and temporal information. To solve incoherent and jittery predictions, Lin et al [8] presented a Recurrent 3D Pose Sequence Machine(RPSM) to learn the temporal context information and estimate 3D poses. Instead of manually designed elaborate prior terms and constraints, it automatically learn the image-dependent structural constraints. Hossain and Little [20] designed a sequence-to-sequence network composed of layer-normalized LSTM units to utilize the temporal information. They encoded a sequence of 2D poses from a video and decoded into 3D poses, and imposed the constraints of temporal smoothness.

III. Our approach

We split this task into two stages, firstly, leverage advanced 2D human pose estimator [7] [8]to gain an accurate 2D

skeleton map, and then feed them into our dilated convolution based network and generate the final result, as Fig. 2 shows. The dilated convolution pipeline assist us to utilize long term information and be paralleled and real-time compared with RNN architecture.

Pruning Layer. Pruning Layer is employed to make the entire model be more sparse and prevent overfitting, which is different from traditional dropout operations. Dropout [12] is a regularization technique and sets values to zeros randomly, whose goal is the same with pruning operation. However, pruning does not change values randomly, but keeps larger ones and remove nodes with lower weights.

Fig. 3. HRI-I dataset. We propose a 3D human pose dataset named HRI-I, which is captured in the smart shop environment and contains more than 16k poses, 26 people and 6 scenarios. The scenarios are ”fetch(stand)”, ”fetch(hunker)”, ”hold(stand)”, ”hold(hunker)”, ”hunker” and ”walk”.

Quantization and HRI-I dataset. To make the mode be smaller and faster, after trained and tested on HumanEva-I [24] and Human3.6M [25], we employ quantization for float variables when it runs to estimate customers' poses in our established smart shop environment. At the same time, we propose a customer 3D pose dataset HRI-I, which contains 26 people, more than 16k poses and 6 categories, including ”walk”, ”hunker”, ”fetch(stand)”, ”fetch(hunker)”, ”hold(stand)” and ”hold(hunker)”, as Fig. 3 shows. The poses are captured by 4 calibrated cameras with high-resolution 50Hz. It is trained on our proposed customer dataset, tested in the real environment and it occurs accuracy loss within an acceptable range because of quantization.

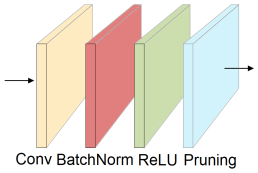

Fig. 4. The architecture of block. We introduce pruning layer and quantization in the block.

We employ a ResNet architecture to enable our network be deeper and exact more high-level information, as Fig. 4 shows. Each block contains convolution, batch normalization, relu and pruning layers, and every $\boldsymbol{N}$ blocks are surrounded by a skip connection. First, we define $\boldsymbol{J}$ joints to represent human pose, which are 12 symmetrical body ones and 5 facial ones and each frame contains one person with his detected 2D joints. It is conducted 1D convolution by a kernel, whose size is $\boldsymbol{W}$ and dilation factor $D = W^B$, and then followed by a convolution by a kernel whose size is 1. Convolutions apply zero-paddings to make sure the number of outputs as the same as inputs. Batch normalization follows convolutions, and then the result is fed into rectified linear units and pruning layer. Receptive field increases exponentially with respect to the factor $\boldsymbol{W}$. Compared with it, the number of parameters grows linearly. Finally, the prediction results of 3D poses are given based on the sequences of past, current and future frames in the video, which is modified and leverages only past and current frames when it applies in the smart shop to run in the real time.

However, even if some human pose estimation datasets were proposed, the labeled data are still so limited and expensive. So we have to figure out how to deal with unlabeled videos and its application in the wild such as smart shop, monitoring, video retrieval etc. Inspired by [11], we focus on semi-supervised learning and take advantage of 2D keypoint detector as the extended supervision. The auto-encoding problem on unlabeled data is solved by an encoder-decoder architecture, which contains an encoder to estimate pose and perform 3D pose estimation from 2D, and a decoder to project 3D pose back to 2D. To avoid the influence of the distance between subject and camera, we also apply the weighted mean per-joint position error loss function for the human trajectory:

$$L = \frac{1}{y_z} \left\| f\left(x\right) - y \right\| \tag{1}$$

where y_z is the inverse of ground truth depth in the coordinate of camera.

Our approach needs off-line camera calibration and distortion correction to reduce focal and principal influence as the pre-processing and relies on advanced 2D pose estimator. In total, we trained and tested in Human3.6M, HumanEva-I and our proposed HRI-I dataset, which is captured under the circumstances of smart shop, and it performs well, both efficiently and effectively.

As all temporal models are sensitive to samples' correlation, we reduce the correlation in the training samples, which means clips are from different and low-correlation videos and it also assists to improve the generalization performance.

IV. Experimatal Evaluation

A. Datasets

The HumanEva-I. The HumanEva-I dataset contains 7 calibrated video sequences, including 4 grayscale and 3 color ones, which are synchronized with 3D body poses obtained from a motion capture system. It includes 6 common actions, they are walking, jogging, gesturing, throwing, boxing, and cambo.

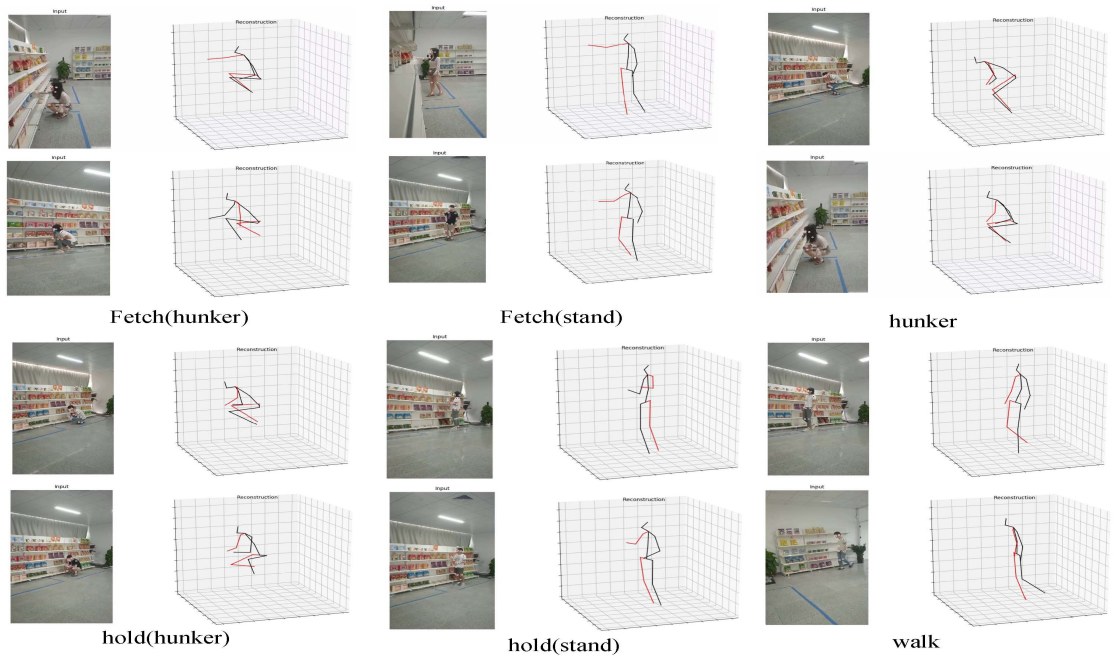

Fig. 5. The 3D human pose estimation results on HRI-I dataset. We classify customers' poses into 6 categories to predict their actions while shopping.

TABLE I
CAMPARISION OF MPJPE ON HUMAN3.6M

	Dir.	Dis	Eat	Greet	Phone	Photo	Posing	Purch
Pavlakos et al. [16]CVPR'17	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3
Tekin et al. [26]ICCV'17	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6
Martinez et al. [27]ICCV'17	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Fang et al. [17]AAAI'18	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7
Yang et al. [28]CVPR'18	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7
Luvizon et al. [6]CVPR'18	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7
Lee et al. [22]ECCV'18	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	43.0
Pavllo et al. [11]CVPR'19	45.2	46.7	43.3	45.6	**48.1**	55.1	44.6	44.3
Ours	45.3	**46.2**	**42.2**	**44.7**	48.3	**54.7**	**43.2**	**42.5**
Ours(with Quantization)	64.7	67.2	61.3	64.2	69.2	78.2	61.8	61.3
	Sit	**SitD**	**Smoke**	**Wait**	**WalkD**	**Walk**	**WalkT**	**Average**
Pavlakos et al. [16]CVPR'17	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Tekin et al. [26]ICCV'17	70.1	107.3	69.3	70.3	74.3	51.8	63.2	69.7
Martinez et al. [27]ICCV'17	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang et al. [17]AAAI'18	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Yang et al. [28]CVPR'18	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Luvizon et al. [6]CVPR'18	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
Lee et al. [22]ECCV'18	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Pavllo et al. [11]CVPR'19	57.3	65.8	47.1	**44.0**	49.0	**32.8**	**33.9**	46.8
Ours	**55.6**	**63.2**	**46.5**	45.6	**48.7**	34.7	35.8	**46.5**
Ours(with Quantization)	79.5	90.1	66.2	65.1	67.5	49.6	51.1	**66.5**

The Human3.6M. The Human3.6M dataset contains 3.6 million 3D human poses and corresponding images, which is conducted by 11 professional actors and actresses(6 males, 5 females). It consists of 17 scenarios, including discussion, smoking, taking photo, talking on the phone, etc.

The HRI-I. To adapt to the smart shop environment, we propose a 3D human pose dataset named HRI-I that comprises more than 16k poses, 26 people and 6 scenarios, including "walk", "hunker", "fetch(stand)", "fetch(hunker)", "hold(stand)" and "hold(hunker)".

B. Evaluation

Inspired by the work of detectron [29], we apply Mask R-CNN, whose architecture is ResNet-101-FPN and cascaded pyramid network to gain the results of 2D human pose estimation as our input. For 3D pose estimation, we train for 90 epoches, adopt a decaying learning rate schedule, which starts at η =0.001 and its shrink factor α=0.95. To avoid fluctuations because of the hyperparameters belonging to batch normalization, we adopt a schedule for the batch normalization, which starts from 0.1 and decays exponentially

until 0.001 in the final epoch.

Metric. In our experiments, the 3D pose estimation is evaluated by the mean per-joint position error(MPJPE) millimeters, which represents the Euclidean distance between predicted joint positions and ground-truth ones.

We set that the ground truth 3D pose

$$y = \left[p_1^3, ..., p_M^3\right] \tag{2}$$

and the predicted 3D pose is $\overline{y}$,

$$\overline{y} = \left[\overline{p}_1^3, ..., \overline{p}_M^3\right] \tag{3}$$

MPJPE is calculated as follow,

$$MPJPE = \frac{1}{M}\sum_{i=1}^{M}\left\|p_i^3 - \overline{p}_i^3\right\|_2 \tag{4}$$

C. Results

Table. I shows results for our model with 4 blocks, 17 joints per frame, and 243 input frames. The model has lower average MPJPE compared with other approaches.

To decrease the size of model and run in real time, we employ quantization to accelerate and its accuracy loss is in an acceptable range. Fig. 5 shows the results on our proposed dataset HRI-I, it is able to estimate single person's pose in an effective and efficient manner.

V. Conclusions

We introduce an effective 3D human pose estimation based on dilated convolution to leverage temporal information and propose a 3D human pose dataset named HRI-I in a smart shop environment, which contains more than 16k poses, 26 people and 6 scenarios. To prevent overfitting, we employ pruning layer to set lower weights to zeros and it performs better than dropout, which drops nodes randomly. To accelerate it in the smart shop environment and gain a trade off between performance and computational complexity, we conduct quantization for float variables to decrease the size of model, which occurs accuracy loss in an acceptable range.

References

[1] W. Gong et al., "Human Pose Estimation from Monocular Images: A Comprehensive Survey," *Sensors*, vol. 16(12), pp. 1–39, 2016.

[2] X. Perez-Sala, S. Escalera, C. Angulo, and J. Gonzàlez, "A survey on model based approaches for 2D and 3D visual human pose recovery," *Sensors*, vol. 14(3), pp. 4189–4210, 2014.

[3] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," *In Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, pp. 1–8, 2010.

[4] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 4966–4975, 2016.

[5] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross View Fusion for 3D Human Pose Estimation," *arXiv preprint arXiv:1909.01203*, 2019.

[6] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 5137–5146, 2018.

[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *In International conference on computer vision(ICCV)*, pp. 2980–2988, 2017.

[8] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 936–944, 2017.

[9] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng, "Recurrent 3D pose sequence machines," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 5543–5552, 2017.

[10] P. Sinha and E. Adelson, "Recovering reflectance and illumination in a world of painted polyhedra," *In International Conference on Computer Vision(ICCV)*, pp. 156–163, 1993.

[11] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.7753-7762, 2018.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, S. Ruslan, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15(1), pp. 1929-1958, 2014.

[13] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44(6), pp. 2325–2383, 1998.

[14] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1653–1660, 2014.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *In European Conference on Computer Vision(ECCV)*, pp. 483–499, 2016.

[16] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 7025–7034, 2017.

[17] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," *In AAAI Conference on Artificial Intelligence(AAAI)*, pp. 6821–6828, 2018.

[18] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, vol. 28(1), pp. 44–58, 2006.

[19] H. Jiang, "3D human pose reconstruction using millions of exemplars," *International Conference on Pattern Recognition(ICPR)*, pp. 1674–1677, 2010.

[20] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1446–1455, 2015.

[21] D. Mehta et al., "Monocular 3D human pose estimation in the wild using improved CNN supervision," *International Conference on 3D Vision (3DV)*, pp. 506–516, 2018.

[22] K. Lee, I. Lee, and S. Lee, "Propagating LSTM: 3D pose estimation based on joint interdependency," *In European Conference on Computer Vision(ECCV)*, pp. 119–135, 2018.

[23] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3D human pose estimation," *In European Conference on Computer Vision(ECCV)*, pp. 68–84, 2018.

[24] L. Sigal and M. J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion", *Brown Univertsity TR*, pp. 1-16, 2006.

[25] 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 Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, vol. 36(7), pp. 1325–1339, 2014.

[26] B. Tekin, P. Marquez-Neila, M. Salzmann, and P. Fua, "Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation," *In International Conference on Computer Vision(ICCV)*, pp. 3961–3970, 2017.

[27] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A Simple Yet Effective Baseline for 3d Human Pose Estimation," *In International Conference on Computer Vision(ICCV)*, pp. 2640–2649, 2017.

[28] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3D Human Pose Estimation in the Wild by Adversarial Learning," *In Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 5255–5264, 2018.

[29] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron." https://github.com/facebookresearch/detectron, 2018.