Pose Sequence Model Using the Encoder-Decoder Structure for 3D Pose Estimation

Jiwei Zhang[1(✉)], Lian Yang[1], Tianbo Ye[1], Jiaen Zhou[1], Wendong Wang[1], and Ying Tan[2]

[1] Beijing University of Posts and Telecommunications, Beijing 100876, China
jwzhang666@bupt.edu.cn
[2] Peking University, Beijing 100091, China

Abstract. Human pose estimation is a hot research problem in computer vision, it has a certain application prospect in the automatic driving industry, security field, film and television industry, and specific action monitoring of special scenes. Because a 2D skeleton usually corresponds to multiple 3D skeletons, the mapping from 2D to 3D in the monocular video has inherent depth ambiguity and is ill-posed, which makes the research on the technology of 3D human pose estimation in monocular video challenging. In this paper, a Pose Sequence Model (PSM) for 3D human pose estimation in the monocular video is proposed, which combines the full convolution neural network based on extended convolution with the Long Short-Term Memory (LSTM) network. We make full use of convolution to extract spatial features and use LSTM to obtain temporal features. With this model, we can predict 3D human posture through 2D sequences. Compared with the previous work on classical data sets, our method has good detection results.

Keywords: 3D human pose estimation · PSM · Monocular video

1 Introduction

The research on 2D pose estimation has 2 main methods: top-down and bottom-up methods. The top-down method [19,25,28,30] takes the result from human detection, generally a bounding box, and performs the single human pose estimation on each human block diagram. The bottom-up method [18], oppositely, starts by detecting the human body key points in the image and then groups the key points into a human body. Toshev et al. [26] transformed the 2D human pose estimation problem from the original image processing and template matching problem into CNN image feature extraction and keypoint coordinate regression problem, and used DNN-based regression criteria to estimate the occludes/missing human joint nodes, which brings great influence. Until now, the 2D pose estimation has reached relatively high accuracy and high resolution

Y. Tan and Y. Shi (Eds.): DMBD 2022, CCIS 1744, pp. 165–177, 2022.
https://doi.org/10.1007/978-981-19-9297-1_13

[23]. Combining the 2D and 3D human pose estimation, Chen et al. [4] conveyed that rather than directly measure 3D pose from images, the procedure of 3D pose estimation can be divided into 2D pose estimating using Mature deep neural networks, and 3D mocap data matching, this has been the main idea of posture estimating.

In recent years, there has been vast research on 3D posture estimating. Some focus on estimating 3D pose from 2D pose of a single image, Martinez et al. [16] conducted an efficient neural network to infer from 2D projections to 3D joints, which focuses on the visual parsing of human bodies in 2d images. To solve the problem of unknown motions and camera positions, Wandt et al. [27] proposed an extra camera network to infer camera parameters, followed by a reprojection layer to reproject the 3D pose back to 2D. Li et al. [14] designed a dataset evolution framework to address the problem of the biased dataset, along with a cascaded network: TAGNet to predict the final 3D skeleton from the enhanced data. Based on the part-guided novel image synthesis, Kundu et al. [10] proposed a self-supervised learning framework to disentangle the inherent factors of variations: shape and appearance. Some research may get 3D pose from explicit middle representations, Pavlakos et al. [20] proposed volumetric representation for 3D human pose(3D heatmap) and coarse-to-fine prediction technique to validate the value of end-to-end learning for the representation of 3D pose, which addresses the challenge of estimating 3D human poses from a single color image. Li et al. [12] introduced the mixture density networks (MDN) [1,32] into the 3D joint estimation to verify the hypotheses that multiple feasible poses can be inferred from a monocular input. Li et al. [13] designed HybrIK reconstructing 3D body mesh by twist-and-swing decomposition to bridge the gap between volume grid estimation and 3D keypoint estimation, which both preserve the accuracy of the 3D pose and the real body structure of the parameterized human body model, to obtain a pixel-aligned 3D body grid and a more accurate 3D pose.

CNN can fully learn images or videos' high-level semantic information and has excellent spatial information extraction capabilities. However, the 3D human pose recognition task based on the human skeleton sequence is a significant time-dependent problem for monocular video. So, balancing and making better use of spatial and temporal information is an extremely difficult task. An additional issue that needs to be addressed is raising the model's generalizability for outdoor datasets. As a result, we propose a multi-stage framework for estimating the 3D human pose that begins by estimating the 2D human pose from the image, then from the result to estimate the 3D human pose. Some 2D outdoor datasets can be used to provide the model with generalization capabilities by making use of the model's revolutionized 2D human pose detector. After that, the mapping relationship sequence from the 2D human pose to the 3D human pose is modeled, transforming the issue into a time-based sequence modeling task. The encoder-decoder structure PSM makes use of the LSTM as an encoder and the fully convolutional neural network as a decoder. The use of LSTM as an encoder makes it possible to first encode the video frame's correlation into a vector of fixed size and then decode it using a fully convolutional neural network. The CNN network's spatial information processing capability and the RNN

network's temporal information acquisition capability can be combined and fully utilized in this manner. Additionally, the jitter of 3D human motion between video frames frequently presents a challenge when estimating the 3D human pose from monocular video. Since the polynomial order can be modeled using motion refinement and used as an optional branch to optimize the prediction results, motion refinement is used to reduce bounce and increase accuracy. Consequently, the performance of 3D human pose estimation can be improved by the proposed framework. In Fig. 1, the first row contains the images, and the second row corresponds to the estimation results of the 3D pose.

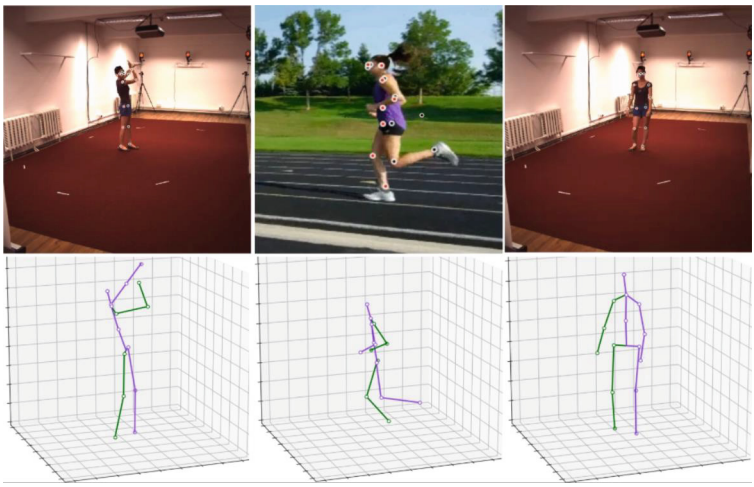

Fig. 1. The effect of the 3D human pose estimation.

The contribution is reflected in the following points:

1. The structure of encode-decode is formed by CNN and RNN, which can utilize spatial and temporal information by encoder-decoder structure.
2. We propose a PSM that combines LSTM with a fully convolutional neural network, which can be used for the estimation of 3D pose.
3. Our framework is implemented in the 3D datasets and good results can also be obtained for wild web videos.

2 Related Work

In recent years, the field of 3D pose recognition has developed rapidly, mainly in two directions: picture recognition and video recognition. A method for multi-person 3D pose recognition using a single image is proposed in [7], which allows the recognition of single or multi-person poses using constant time. They designed a simple and effective compression method using high-resolution body

heat maps and decoded them using an auto-encoder. [6] addressed the problem of 3D pose estimation for multiple people in a few calibrated camera views. A multi-way matching algorithm is used to cluster the detected 2D poses in all views. Each cluster encodes the correspondence between the pose and key points of the same person in different 2D views to efficiently infer the 3D pose of the person. A feature-enhanced network is proposed in [15] to estimate 3D hand pose and 3D body pose using a single RGB image. To address the effects arising from texture, illumination changes, and occlusion in real applications, a long and short-term dependent perception module is used for enhancement. A contextual consistency gate is also introduced to modulate based on contextual consistency. A graph-based approach is proposed in [3] for the problems of depth ambiguity and severe self-obscuration, considering spatial dependence and temporal consistency. A local-to-global network structure is also implemented to solve the 3D human pose estimation problem from short sequence 2D joint detection.

Although it is possible to divide the video into multiple frames for pose recognition, there are often different problems in the video. Graph convolutional networks are often built on fixed human-joint affinities, which can reduce the adaptive ability of GCNs to handle complex Spatio-temporal pose changes in videos. A 3D pose estimation neural network that can adaptively learn video Spatio-temporal relations is proposed in [22]. And [2] proposed a method for multi-person 3D pose estimation and tracking from multi-point video, where each point undergoes independent pose detection followed by correction and correlation, thus generating and tracking 3D skeletons using the associated pose. Multiplayer full-body 3D pose estimation and tracking in dynamic motion scenes are achieved. In exception to joint position prediction, a prediction based on skeletal orientation and skeletal length is proposed in [5], and since the human skeletal length is constant, a full convolutional propagation architecture with long jump connections that can effectively use the information in the video for prediction is proposed. To address the accurate recognition of depth ambiguity, self-obscuration, or other uncommon poses, [33] proposed a new skeletal GNN solution using a hop-count-aware hierarchical channel squeezing fusion layer that effectively extracts information from neighboring nodes while suppressing undesired noise in the GNN, thus effectively improving the prediction accuracy.

3 Method

3D human pose detection methods which are end-to-end must simultaneously complete feature extraction and 3D joint prediction. In addition, since 2D human posture corresponds to multiple 3D human postures, there are inherent fuzziness and discomfort in using end-to-end methods to estimate 3D human posture using monocular images. In this paper, the framework we propose is multi-stage. First, we convert the image into 2D human pose through a 2D detector, and then establish the mapping relationship between 2D and 3D human pose through depth learning method. The prediction framework is shown in Fig. 2.

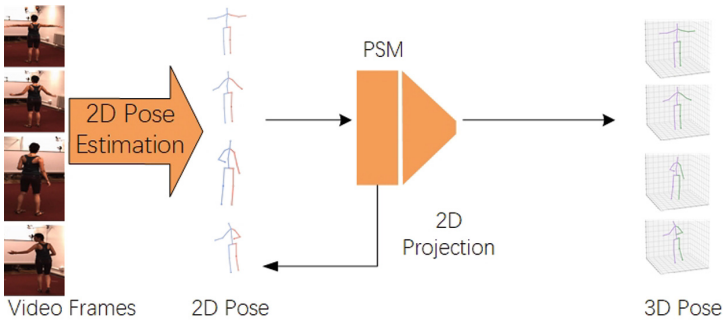

Fig. 2. The prediction framework for the estimation of 3D human pose.

It can be seen from Fig. 2 that our framework is a multi-stage 3D human posture, and different 2D human posture detectors can be used to improve the generalization performance of the model. At the same time, a PSM composed of LSTM and a complete convolution model is proposed to predict 3D human pose through 2D human pose sequence.

3.1 Time-Series Modeling

The input sequence can supposed as $x_0 \ldots, x_t \ldots, x_T$, where x_t is the 2D human poses. Then, we estimate the output $y_0, \ldots, y_t \ldots, y_T$, where y_t is the 3D human poses. In the case of non-causality, for the given time t, the output y_t can be got by passing any subset of x_T. For the causal cases, the data $x_0, \ldots, x_t$ is observed before the t state can only be used. So the time series modeling can be as a function $f : X^\tau \rightarrow Y^\tau$ that can produce a mapping relationship:

$$y'_0, \ldots, y'_T = f\left(x_0, \ldots, x_T\right) \tag{1}$$

For the causal situation, y_t should be obtained only from $x_0, \ldots, x_t$ instead of the subsequent input $x_{t+1}, \ldots, x_T$.

3.2 The Proposed Pose Sequence Model

The encoder-decoder model is a common scheme in time series modeling. The encoding can convert the input time series into vectors, and the decoding can convert the vectors into output sequences. The combination of CNN and RNN can form an encoder and decoder structure. The encoder part of the PSM model we use is the LSTM structure, and the decoder part is a fully convolutional network, forming the RNN-CNN structure. It can effectively use the ability of LSTM to extract time information in time series modeling and the advantages of CNN in processing spatial information. The PSM model is shown in Fig. 3.

Because LSTM can be used to deal with the long-term dependence in time series modeling, its structure is relatively simple and its parameters are few,

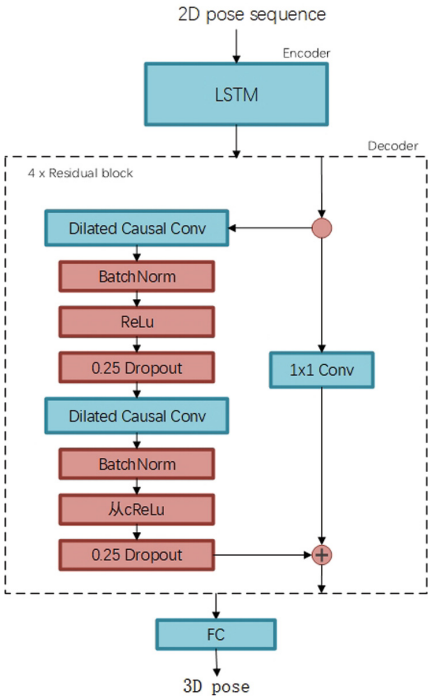

Fig. 3. The proposed PSM for 3D pose estimation.

so it is used as the encoder of the model in this paper. In addition, a dilated convolution is used to decode the LSTM encoding. When causal PSM is used for training, the encoder uses LSTM, while noncausal PSM uses bidirectional LSTM. There are some differences between bidirectional LSTM and LSTM. The current state $t$is not only related to the previous state $t-1$, but also related to the next state $t+1$.

The PSM decoder consists of the full convolution model of the dilated convolution, and the dilated convolution refers to the standard convolution with holes. The reception field of each convolution kernel can be changed by adjusting the kernel spacing, and the dilated convolution obtains multi-scale information by setting different dilation rates. We set different expansion rates for different blocks. This strategy can play the advantages of parallel processing and reduce the loss of information at the hole. For a 2D sequence $\mathrm{x} \in \Re^2$ and a function $f : \{0, \ldots, k-1\} \Rightarrow \Re$, operation of dilated convolution F acting on any element e of the sequence x is expressed as follow.

$$F(e) = (x *_D f)(e) = \sum_{i=0}^{k-1} f(i) \cdot X_{s-D \cdot i} \tag{2}$$

k is the size of the convolution kernel and D is the expansion factor. The fully convolutional neural network includes the Batch Norm (BN), Relu, and dropout. The BN layer is to normalize the batch of data, a BN layer is after the fully connected layer to ensure each layer remains uniformly distributed. Moreover, the Relu function is chosen as the activation function. For the dropout, each neuron stops with a probability of p. Moreover, a residual connection is used to superimpose the input and the output, which solves the problem of gradient disappearance caused by deep networks. After obtaining the 2D joints J in each image, LSTM is used to perform encoding, and a fully convolutional neural network is used to complete the decoding of temporal convolution. For the LSTM, the number of hidden layers is set to $J * 2$. For the decoder, kernel with size K is set to 3, the output is dilated convolution with a size C=1024 and an expansion factor $D = K^N$, where N are the n-th residual modules. The next part is BatchNorm, Relu, and dropout layers.

3.3 Training Details

The training process is shown in Fig. 4.

Here, the 2D pose represents the 2D sequences. The branch of the 3D pose prediction learns the mapping relationship from 2D to 3D pose using PSM. The $y^{(j)}_{f_{3d,t}}(i)$ is the joint j in the t-th frame predicted by the model, and $y^{(j)}_{f_{gt,t}}(i)$ is the ground truth for the t-th frame. The loss function of Mean Per Joint Position Error (MPJPE) can be defined as:

$$L_{3d=}\frac{1}{N_T}\frac{1}{N_S}\sum_{j=1}^{N_T}\sum_{i=1}^{N_S}\left\|y^{(j)}_{f_{3d,t}}(i)-y^{(j)}_{gt,t}(i)\right\|_2 \tag{3}$$

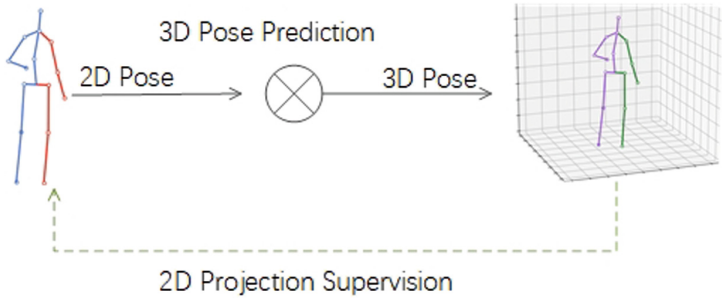

Fig. 4. The training process.

N_T is the number of video frames, and N_S is the number of joints.

Simultaneously, weak supervision of 2D projection is used. We project the estimated 3D pose to the 2D space and get the MPJPE loss L_{proj}, then the loss of the overall task is :

$$L = L_{3d} + L_{proj} \tag{4}$$

Where L_{3d} is 3D loss, L_{proj} is 2D loss.

During training, each step represents as N. The framework consists of the below steps. First, the 2D joints J^{2D} are normalized, and then the 3D joint $\widehat{J^{3D}}$ are predicted through the PSM. $\widehat{J^{3D}}$ can be calculated with the ground truth to obtain L_{3d}. The projected 2D joint can be obtained through projection, and then we calculated with the ground truth to obtain L_{proj}, where the camera parameter is C.

4 Experimental Verifications

First, the data sets used for training and testing and the overall evaluation indicators are introduced, and the proposed framework is compared with the baselines method in different data sets. After verification, our framework has achieved good results.

4.1 Datasets and Evaluation

In the experiment, we mainly used two data sets, HumanEVA and Human3.6m. Hman3.6M is a general data set in the field of 3D human pose estimation. It includes 15 groups of actions completed by motion capture, and a total of 3.6 million videos are provided in 50 HZ format. 17 joint point models are used, 5 object groups (S1, S5, S6, S7, S8) are used as training sets, and (S9, S11) are used as test sets. HumanEVA is another data set used in the experiment, with a total of 4 test objects. According to actions, it can be divided into single-action SA protocol and multi-action MA protocol.

In the evaluation process, we used two protocols: P1 is used to calculate the Euclidean distance between the predicted 3D coordinates and the ground truth, which is averaged according to the number of joints and frames, namely MPJPE. P2 uses Procrustes analysis to evaluate the error between the rigid body transformation result and the ground truth, which is P-MPJPE.

4.2 Implementation Details

Our 2D detector can use different networks, including Mask R-CNN [9] and HRNet [24]. For Mask R-CNN, the ResNet-101 backbone network can be used. The learning rate starts from 1e–3, the attenuation rate is 0.995, and 80k iterations of training have been conducted. For HRNet, starting from 1e–4, it was reduced to 1e–6 in the 15th iteration, and a total of 20k iteration trainings were conducted.

In addition, the Human3.6m dataset has been translated and rotated. The receptive field of the PSM model is set to 243 and the attenuation factor is 0.95. For the HumanEVA dataset, the attenuation factor is 0.99, and 800 cycles of training were conducted.

4.3 Experiment on 3D Datasets

Comparison Results on Human3.6m Dataset. Comparative experiments are carried out on the Human3.6m dataset and the results are as follows.

Table 1. The value of P1 on Human3.6m dataset

	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	Smoke	Wait	Walk	Avg
Fang et al. AAAI (2018) [8]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	60.3	57.7	47.5	60.4
Yang et al. CVPR (2018) [31]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	57.4	58.4	60.1	58.6
Pavllo et al. CVPR (2019) [21]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	47.1	**44.0**	32.8	46.8
Wu et al. AAAI (2020) [29]	**36.9**	43.9	39.5	60.4	45.3	**51.6**	**38.1**	41.9	54.1	44.4	57.6	32.2	47.3
Ours, causal	41.7	44.1	41.4	43.1	46.0	52.4	44.9	43.2	54.4	44.2	45.1	32.8	44.7
Ours, non-causal	41.3	**43.8**	**39.1**	**42.5**	**45.1**	51.8	44.7	**41.5**	**52.8**	**43.9**	44.8	**32.0**	**43.9**

Table 1 and Table 2 show the results on the Human3.6m dataset under the evaluation indicators P1 and P2. The model uses HRNet as a two-dimensional attitude detector, and the data in the table contains the results of multiple actions. The smaller the value of the evaluation index P1 and P2, the better. The last column of the table is the average value of multiple groups of actions. Cause and effect represent cause and effect PSM, which takes the previous frame as input, rather than cause and effect represents PSM, and the input data includes future frames. The best result in the table is shown in bold, and the second-best result is shown in the underline. It can be seen from the table that non-causal PSM achieves better results than causal PSM. Our method ranks first in most actions and second in some actions.

Table 2. The value of P2 on Human3.6m dataset

	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	Smoke	Wait	Walk	Avg
Fang et al. AAAI (2018) [8]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	47.2	44.3	36.7	45.7
Yang et al. CVPR (2018) [31]	**26.9**	**30.9**	36.3	39.9	43.9	47.4	38.8	**29.4**	**36.9**	41.5	30.5	42.5	37.7
Pavllo et al. CVPR (2019) [21]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	47.1	44.0	32.8	46.8
Wu et al. AAAI (2020) [29]	32.1	36.2	33.9	41.2	**37.4**	**40.6**	**30.7**	33.4	45.0	37.4	38.8	**25.7**	37.3
Ours, causal	33.1	39.0	33.2	36.8	39.9	40.9	31.2	32.3	43.7	37.0	35.9	26.1	35.7
Ours, non-causal	32.7	38.6	**32.9**	**35.3**	39.8	39.5	30.9	31.9	42.2	**36.8**	**35.2**	**25.7**	**35.2**

Comparison Results on HumanEVA Dataset. We conducted experiments in the HumanEVA dataset to prove the effectiveness of our framework on small-scale datasets. Three participants were selected as test subjects, S1, S2, and S3. Then, using the two-dimensional attitude detector HRNet, multiple actions (MA) and single action (SA) strategies are selected for experiments. As can be seen from Table 3, our framework has generally achieved good results on P2. Especially in the case of MA, the best results are obtained.

Table 3. Comparative experiments on the HumanEVA dataset

Subjects	Walk(S1)	Walk(S2)	Jog(S1)	Jog(S2)	Box(S1)	Box(S2)
Martinez et al. (SA) [17]	19.7	17.4	26.9	18.2	-	-
Lee et al. [11]	18.6	19.9	25.7	16.8	42.8	48.1
Pavllo et al. (SA)	14.5	10.5	21.9	13.4	24.3	34.9
Pavllo et al. (MA)	13.9	10.2	20.9	<u>13.1</u>	<u>23.8</u>	33.7
ours(SA)	<u>12.6</u>	<u>10.0</u>	<u>18.6</u>	13.4	24.1	30.4
ours(MA)	**12.4**	**9.8**	**18.2**	**11.4**	**21.8**	**29.4**

5 Conclusion

The framework proposed in this paper is multi-stage, which is used to realize 3D human pose estimation in monocular video. First, obtain the 2d pose of the human body from the video, and then use the 2d pose to predict the 3d pose. Our model adopts PSM, which can realize the sequence modeling of 2d to 3d pose. PSM is an encoded second structure, which makes full use of the multi-level features extracted by a fully convolutional neural network and LSTM. In addition, since our framework is multi-stage, we can use different 2D detectors to improve performance. Compared with the corresponding baseline methods, our method has achieved good results on HumanEVA and Human3.6m datasets.

Acknowledgements. This work is supported by Key Research and Development Projects of Hebei Province under Grant 21310102D.

References

1. Bishop, C.M.: Mixture Density Networks. IEEE Computer Society, Washington, DC (1994)
2. Bridgeman, L., Volino, M., Guillemaut, J.Y., Hilton, A.: Multi-person 3D pose estimation and tracking in sports. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2487–2496. IEEE, Long Beach, CA, USA, June 2019. https://doi.org/10.1109/CVPRW.2019.00304, https://ieeexplore.ieee.org/document/9025555/

3. Cai, Y., et al.: Exploiting Spatial-Temporal Relationships for 3D pose estimation via graph convolutional networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2272–2281. IEEE, Seoul, Korea (South), October 2019. https://doi.org/10.1109/ICCV.2019.00236, https://ieeexplore.ieee.org/document/9009459/
4. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7035–7043, July 2017
5. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3D human pose estimation with bone-based pose decomposition. IEEE Trans. Circ. Syst. Video Technol. **32**(1), 198–209 (2022). https://doi.org/10.1109/TCSVT.2021.3057267, https://ieeexplore.ieee.org/document/9347537/
6. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation from multiple views. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7784–7793. IEEE, Long Beach, CA, USA, June 2019. https://doi.org/10.1109/CVPR.2019.00798, https://ieeexplore.ieee.org/document/8953350/
7. Fabbri, M., Lanzi, F., Calderara, S., Alletto, S., Cucchiara, R.: Compressed volumetric heatmaps for multi-person 3D pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7202–7211. IEEE, Seattle, WA, USA, June 2020. https://doi.org/10.1109/CVPR42600.2020.00723, https://ieeexplore.ieee.org/document/9156316/
8. Fang, H., Xu, Y., Wang, W., Liu, X., Zhu, S.: Learning pose grammar to encode human body configuration for 3D pose estimation. In: Proceedings of AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, pp. 6821–6828 (Feb2018)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN (2017). http://arxiv.org/abs/1703.06870
10. Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3d human pose estimation via part guided novel image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6152–6162, June 2020
11. Lee, K., Lee, I., Lee, S.: Propagating LSTM: 3D pose estimation based on joint interdependency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 123–141. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_8
12. Li, C., Lee, G.H.: Generating multiple hypotheses for 3D human pose estimation with mixture density network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9887–9895, June 2019
13. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: a hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3383–3393, June 2021
14. Li, S., Ke, L., Pratama, K., Tai, Y.W., Tang, C.K., Cheng, K.T.: Cascaded deep monocular 3D human pose estimation with evolutionary training data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6173–6183, June 2020
15. Liu, J., et al.: Feature Boosting Network For 3D Pose Estimation. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 494–501 (2020). https://doi.org/10.1109/TPAMI.2019.2894422, https://ieeexplore.ieee.org/document/8621059/

16. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2640–2649, October 2017
17. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy. pp. 2659–2668 (Oct2017)
18. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/8edd72158ccd2a879f79cb2538568fdc-Paper.pdf
19. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4903–4911, July 2017
20. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7025–7034, July 2017
21. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 7753–7762, June 2019
22. Sengupta, A., Budvytis, I., Cipolla, R.: Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11199–11209. IEEE, Montreal, QC, Canada, October 2021. https://doi.org/10.1109/ICCV48922.2021.01103, https://ieeexplore.ieee.org/document/9709969/
23. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703, June 2019
24. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 5693–5703, June 2019
25. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2602–2611, October 2017
26. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1653–1660, June 2014
27. Wandt, B., Rosenhahn, B.: RepNet: weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7782–7791, June 2019
28. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686
29. Wu, H., Xiao, B.: 3D human pose estimation via explicit compositional depth maps. In: Proceedings of AAAI Conference on Artificial Intelligence New York, NY, USA, 7–12 February 2020, pp. 12378–12385, Feb 2020

30. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 472–487. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_29
31. Yang, W., Ouyang, W., Wang, X., Ren, J.S.J., Li, H., Wang, X.: 3D human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA. pp. 5255–5264, June 2018
32. Ye, Q., Kim, T.K.: Occlusion-aware hand pose estimation using hierarchical mixture density network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–817, September 2018
33. Zhang, J., Wang, Y., Zhou, Z., Luan, T., Wang, Z., Qiao, Y.: Learning dynamical human-joint affinity for 3d pose estimation in videos. IEEE Trans. Image Process. **30**, 7914–7925 (2021). https://doi.org/10.1109/TIP.2021.3109517, https://ieeexplore.ieee.org/document/9531423/