ISSN 2464-4617 (print)
ISSN 2464-4625 (DVD)
Computer Science Research Notes
CSRN 3201
WSCG 2022 Proceedings

Balanced Feature Fusion for Grouped 3D Pose Estimation

Jihua Peng[1,2], *Yanghong Zhou*[1], *P.Y. Mok*[1,2,*]
[1]*Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hong Kong.*
[2]*Laboratory for Artificial Intelligence in Design, Hong Kong Science Park, Hong Kong.*
* tracy.mok@@polyu.edu.hk

ABSTRACT

3D human pose estimation by grouping human body joints according to anatomical relationship is currently a popular and effective method. For grouped pose estimation, fusing features of different groups together effectively is the key step to ensure the integrity of whole body pose prediction. However, the existing methods for feature fusion between groups require a large number of network parameters, and thus are often computational expensive. In this paper, we propose a simple yet efficient feature fusion method that can improve the accuracy of pose estimation while require fewer parameters and less calculations. Experiments have shown that our proposed network outperforms previous state-of-the-art results on Human3.6M dataset.

Keywords

3D Human Pose Estimation, Grouping Feature Fusion, Anatomical Relationships

1 INTRODUCTION

3D human pose estimation is the computer vision task of estimating the articulated 3D joint locations of a human body from an input image or video, which has received much research attention because it supports many applications including cloth parsing [Don14a], surveillance [Liu18a], augmented reality [Lin10a], action prediction [Luv18a].

The main stream convolutional neural networks-based methods for 3D human pose estimation mainly follow two approaches: (1) directly regressing 3D coordinates of each joint from input images or sequences; (2) first predicting 2D joint coordinates from images and then matching these 2D key points to 3D coordinates. The second approach significantly outperforms the first one, since these methods benefit from the high performance of intermediate 2D pose detectors [Che18a, Sun19a].

Some researchers proposed to group human joints into parts, such as arms, legs and torso, based on kinematics and human anatomy to improve prediction accuracy [Cai19a, Fan18a, Lee18a, Wan19a]. They first encode each part to get the corresponding features, and then fuse these features together to get a complete human pose. Although the features of each part are independently predicted, different parts affect each other and their features are also related mutually. For example, there is interaction between the arm and the head for poses in "Eating", while the hand and the foot are closely related when performing poses "Running". In these methods, a feature fusion module is used to fuse the features of different parts together; while many existing feature fusion methods just use fully connected layers, resulting in large number of parameters and high computation costs.

In this paper, we propose a network framework with an optimized feature fusion (OFF) module for 3D pose estimation, as shown in Figure 1. Our method improves the accuracy of 3D pose estimation while requires fewer parameters and has lower computational complexity.

The remaining of this paper is organized as follows. We first review the related works on 3D human pose estimation. Then, we describe the proposed method in detail and experimentally demonstrate the effectiveness of our method by comparing with state-of-the-art methods and ablation studies. Finally, we conclude our work, and discuss limitations and future research directions.

2 RELATED WORK

3D human pose estimation has been studied since a very early time. Early traditional methods utilized pictorial structures to predict 3D coordinates of human joints. These methods usually require tedious manual operations and a large amount of calculations, and get bad results when encountering some complex poses. Thanks to the rapid development of deep learning methods, convolutional neural networks-based methods have become the main stream and achieved very promising results in recent years. These methods can be classified into two categories. Some early works

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

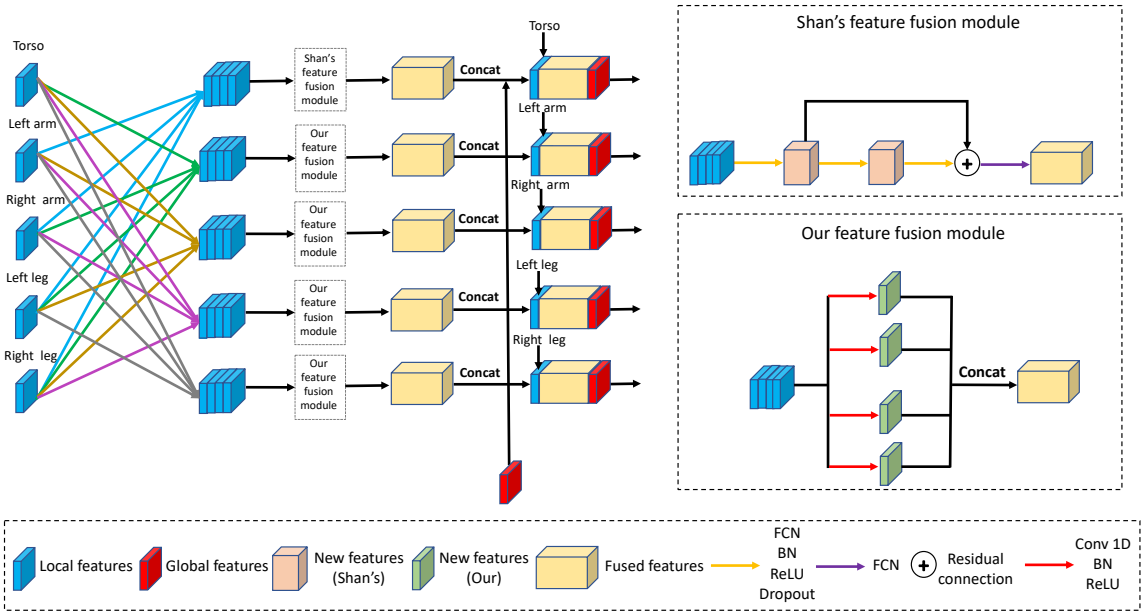

Figure 1: Our and Shan's [Sha21a] feature fusion modules. There are five groups of local features. We both concatenate the four groups of local features together. Then Shan [Sha21a] uses three modules based on FCN to raise the feature dimension and outputs fused features through a FCN, while we use four modules based on 1D convolution to conduct discriminative dimensionality reduction and obtain four new features respectively. We concatenate them together to get the fused features. For the torso part, we still use the feature fusion method proposed by [Sha21a]. Finally, the remaining local features, fused features, and global features are concatenated together to become a new feature of the body part. After performing such feature fusion five times, five new features of five body parts are obtained. (FCN) - Fully Connected Layer. (BN) - 1D Batch Normalization. (Conv 1D) - 1D convolution.

[Pav17a, Tek16a] directly predict the 3D human joints from the input image or video through neural networks, which is called one-step regression method. Recently, benefit from recent advancement in 2D pose detectors [Che18a, Liu18a, Liu19a, Sun19a, Hua20a], many methods first detect the coordinates of the 2D pose joints, and then use these 2D coordinates to regress the human pose joint in the 3D space. Among the methods of regressing 3D coordinates, the method using temporal information has the promising results.

The Long Short-Term Memory (LSTM) model is the first sequence-to-sequence model that extract 3D human pose information from videos. It mainly encodes the coordinates of 2d human joint, and then decodes them into coordinates in 3d space [Hos18a]. Lee et al. [Lee18a] design a propagating LSTM structure based on joint interdependency to learn the spatial position relationship of each joint of the human body. However, LSTM cannot process multiple frames in parallel, but store them sequentially in memory resulting in many parameters. To address this problem, temporal convolutional network (TCN) is proposed for 3D human pose estimation [Pav19a]. It performs 1D convolutions over 2d input pose sequences with fewer parameters. Liu et al. [Liu20a] apply attention mechanism to TCN, which determine key frames and output tensor in every layer. Chen et al. [Che21a] transform TCN to predict both direction and length of the bones.

Many current methods [Par18a, Wan19a, Zhe20a, Zen20a, Sha21a] group body parts to predict 3d human pose according to the anatomical relationship. Park et al. [Par18a] propose to divide human joints into 5 non-overlapping groups (torso, left arms, right arms, left legs, right legs). Then the features learned by these 5 groups are averaged to produce the feature of the whole pose. Wang et al. [Wan19a] define the degrees of freedom (DOF) and model limbs with higher DOFs and torso with lower DOFs. In this way, various body parts with different DOFs can supervise each other, leading to more reasonable prediction results. Zheng et al. [Zhe20a] treat each joint as a group and design a dual attention module to learn the feature relationship between each group. Zeng et al. [Zen20a] divide the human body into local groups of joints and develop a network to learn internal dependencies within each group and weak dependencies among groups. Shan et al. [Sha21a] split the human body into five groups (torso, left/right arms, left/right legs), encode the features of each group separately, and design a feature

fusion module to fuse the 5 groups of features to obtain a complete human body.

3 METHOD

We use the structure of [Sha21a] as the baseline from our network framework (Figure 1). The input of the network is the coordinate of the 2D pose that have been predicted from the images. It first processes the input 2D pose joints of target pose and other poses to obtain the positional and temporal information, and then divides these information into five groups (torso, left arm, right arm, left leg, right leg). After all the information encoded by the TCN network, we propose a new optimized feature fusion (OFF) module to fuse the five groups of features. We use the OFF module to fuse four groups of features (e.g., left and right hands, left and right legs) and then concatenate them with the remaining group of features (torso) to form a new torso feature. After performing five times of feature fusion, we get five new groups of features (torso, left/right hands, left/right legs). Finally, we decode the five new groups of features to get the coordinates of the joints of the five body parts and concatenate them together to get the complete human pose.

Our proposed feature fusion module is compared with that of [Sha21a] in Figure 1. Both of our inputs are five local features of five body parts, four of which are concatenated together. Shan et al. [Sha21a] use three modules composed of fully connected layer, 1D batch normalization [Iof15a], rectified linear units [Nai10a] and dropout [Sri14a] to raise the feature dimension to obtain the fused features. Shan et al. [Sha21a] also use residual connections to solve the problem of gradient explosion and disappearance when the number of network layers is deep. However, Cheng et al. [Che15a] illustrate that fully connected layers generally involve over 90% of the network parameters and generate redundancy of parameters in deep neural networks. Similarly, the method of [Sha21a] also generate some redundant information, since the features of each group are connected to each other when using fully connected layers to fuse the features of the four body parts. But in fact some body parts are not strongly related in some actions. For example, there is no strong correlation between head and hands in the action âWalkingâ. Compared with the method of [Sha21a], we only use four different 1D convolutions followed by 1D batch normalization [Iof15a] and rectified linear units [Nai10a], which can reduce the amount of parameters greatly and ensure that each part learns enough information.

Specifically, we denote each group of local features as F_i, $F_i \in \mathbb{R}^{B\times C}$, where B and C are the batch size and the number of channels. We concatenate the four groups of local features among them together according to the channel dimension as $[F_1,...,F_4]$. Then we aim to use the discriminative dimensionality reduction method [Sub17a, Gao19a] to make the features separate and obtain a new discriminative feature $F_i^{'}$ for group i by aggregating features from four groups. Therefore, we need to learn the transformation W for concatenated four groups of features $[F_1,...,F_4]$. We denote W as a 1D convolution with 1 stride and 1 kernel size. The input and output channels of this 1D convolution are $4C$ and C respectively. Therefore, this formula for obtaining the discriminative feature $F_i^{'}$ is as follows

$$F_i^{'} = BatchNorm1D(Conv1D([F_1,...,F_4])) \quad (1)$$

After performing four different groups of 1D convolution, batch normalization [Iof15a] and rectified linear units [Nai10a], we get four new groups of features $F_1^{'},F_2^{'},F_3^{'},F_4^{'}$. These new features are concatenated to form the fused features $\left[F_1^{'},F_2^{'},F_3^{'},F_4^{'}\right]$. Subsequently, the remaining local body features F_5 and global features (target pose) are concatenated with the fused feature and all the features are sent to the decoding layer to predict the coordinates of the 3D human pose. In addition, considering that the number of joints in the torso is the largest among the five groups, we still use the feature fusion method based on the fully connected layer proposed by [Sha21a] for the torso group. The remaining four groups of joints adopt our proposed optimized feature fusion (OFF) module.

4 EXPERIMENTS

4.1 Datasets and Evaluation

Dataset We evaluate our model on the public dataset Human3.6M [Ion13a]. Human3.6M is an indoor scenes dataset collected by motion capture system with 3.6 million video frames. It has 11 professional actors wearing clothes with markers which record the coordinates of each human body joint. These actors perform 15 actions in daily life under 4 synchronized camera views, such as walking dogs, photoing, sitting, greeting, eating and so on.

Following previous studies [Mar17a, Pav17a, Fan18a, Pav19a, Liu20a, Sha21a], we adopt five subjects (S1, S5, S6, S7, S8) for training and two subjects (S9 and S11) for testing. We use the commonly used protocols to evaluate our experimental results. Our model is trained in the PyTorch framework on one GeForce RTX 3070 GPU.

Evaluation protocol is denoted as Mean Per Joint Position Error (MPJPE) that is the average Euclidean distance between estimated human joint coordinates and ground-truth human joint coordinates. It is the most popular standard for evaluating the 3D human pose estimation.

Table 2 illustrates the computational complexity of different models. We compare our method with [Pav19a]

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Lee et al. [Lee18a]	32.1	36.6	34.4	37.8	44.5	49.9	40.9	36.2	44.1	45.6	35.3	35.9	37.6	30.3	35.5	38.4
Pavllo et al. [Pav19a]	35.2	40.2	32.7	35.7	38.2	45.5	40.6	36.1	48.8	47.3	37.8	39.7	38.7	27.8	29.5	37.8
Liu et al. [Liu20a]	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Zeng et al. [Zen20a]	34.8	32.1	**28.5**	30.7	31.4	36.9	35.6	30.5	38.9	40.5	32.5	31.0	29.9	**22.5**	24.5	32.0
Shan et al. [Sha21a]	29.5	30.8	28.8	29.1	30.7	**35.2**	31.7	27.8	**34.5**	**36.0**	30.3	29.4	28.9	24.1	24.7	30.1
Ours (T = 243 GT)	**27.8**	**28.2**	28.7	**27.7**	**30.4**	35.8	**31.1**	**26.6**	34.6	36.8	**30.2**	**28.3**	**28.0**	23.6	**24.1**	**29.5**

Table 1: Reconstruction error on Human3.6M under **evaluation protocol** with MPJPE (mm). The input 2d pose is ground truth. The lowest reconstruction error is bold, and the second lowest is underlined. (GT) - ground-truth.

Method	Parameters	≈ FLOPs	MPJPE
Pavllo et al. [Pav19a]	16.95M	33.87M	37.8
Shan et al. [Sha21a]	41.78M	36.03M	30.1
Ours	28.38M	22.39M	29.5

Table 2: Comparison with the computational complexity of different models. All models are trained on ground-truth 2D poses under **evaluation protocol** with MPJPE (mm). The input of all models is a 2D pose sequence of 243 frames. Both parameters and FLOPs of us and [Sha21a] are calculated at stage 2 or 3.

and [Sha21a] in terms of the number of model parameters and an estimate of the floating-point operations (FLOPs) because we all use the TCN as the baseline. FLOPs are the computational power required for forward propagation, which reflects the level of performance required for hardware such as GPU. In our experiments, we mainly calculate the computational power consumption of the convolutional layer, fully connected layer, BatchNorm and ReLU when the model is forward propagated. Both parameters and FLOPs of us and [Sha21a] are calculated at stage 2 or 3. Although our model has more parameters than that of [Pav19a], our MPJPE result is 8.3mm lower. In addition, the number of our model parameters is much less than that of [Sha21a] and our reconstruction error (MPJPE) is also lower. As for FLOPs, our model only has 22.39M which is the least among all models, almost half of the model proposed by [Sha21a].

4.2 Comparison with State-of-the-Art Methods

We compare our results with state-of-the-art works in recent years on the public dataset Human3.6M. Table 1 shows the comparison results between our method and recent methods on Human3.6M under **evaluation protocol**. We use ground-truth 2D poses as input and train under the recpetive field of 243 frames. Our method reaches 29.5mm in MPJPE, which is 0.6mm better than the best result. Besides, our model achieves the state-of-the-art in terms of multiple actions and also reaches the second best result in some complex actions, such as "Eat", "Photo", "Sitting", "Sitting Down", "Walk". The reason why we do not achieve the best results in these complex actions may be that these actions have severe deep ambiguity and occlusion, which requires other groups of joints to provide more information to the occluded or deeply ambiguous joints during feature fusion. Our method reduces nearly half of the parameters in the feature fusion stage, so it may lose some information for the learning of these complex actions.

4.3 Ablation Studies

In order to verify the validity of each part in our model, we perform ablation experiments on whether the torso part uses 1D convolution or fully connected layers on Human3.6M under **evaluation protocol** with MPJPE (mm). Our network takes the ground truth of 2D poses as input. Table 3 shows ablation experiments of different methods used in our network at stage 3. If all feature fusion modules use 1D convolution, although the parameter amount is the lowest, only 24.96M, the MPJPE is only reduced by 0.1mm. However, if the fully connected layer is used for feature fusion for the torso part and the 1D convolution is used for other parts for fusion, the MPJPE is reduced by 0.6mm and the parameter quantity is also much lower than the baseline. This is because the torso part has more joints than other parts and needs to learn more information during feature fusion. 1D convolution feature fusion method makes it not enough to learn features. The baseline uses FCN to perform feature fusion on five parts, so redundant information is learned for the part with few joints. Therefore, in order to balance the number of features that need to be learned in each part, we use the feature fusion module consisting of 1D convolution and FCN.

Method	MPJPE(mm)	△	Parameters
Baseline (FCN (all parts))	30.1	–	41.78M
+OFF(Conv 1D)	30.0	0.1	24.96M
+OFF(Conv 1D + FCN (torso))	29.5	0.6	28.38M

Table 3: Ablation study of different methods in our model at stage 3. OFF refer to optimized feature fusion proposed by us. The MPJPE results takes ground-truth 2D poses as input and is trained on Human3.6M under **evaluation protocol**. (FCN) - Fully Connected Layer. (Conv 1D) - 1D convolution.

5 CONCLUSION

We propose an optimized feature fusion module for grouped human pose in this paper. Compared with the feature fusion module [Sha21a] composed of fully connected layer, our optimized feature fusion module can

use fewer parameters to fuse different groups of human pose features and it can also reduce reconstruction errors. Experimental results prove that our method advances state-of-the-art performance on Human3.6M dataset. However, there are still limitations in our current method: we still require three stages of training and the improvement in prediction accuracy is observed but not in a significant percentage. In the future, we will research on one-stage training of the entire network to further improve prediction accuracy.

6 ACKNOWLEDGMENTS

This research is funded by the Laboratory for Artificial Intelligence in Design (Project Code: RP1-1) under the InnoHK Research Clusters, Hong Kong Special Administrative Region Government.

7 REFERENCES

[Don14a] Dong, J., Chen, Q., Shen, X., Yang, J., Yan, S. Towards unified human parsing and pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 843-850). 2014.

[Liu18a] Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J. Pose transferrable person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4099-4108). 2018.

[Lin10a] Lin, H. Y., Chen, T. W. Augmented reality with human body interaction based on monocular 3D pose estimation. In International Conference on Advanced Concepts for Intelligent Vision Systems (pp. 321-331). Springer, Berlin, Heidelberg. 2010.

[Luv18a] Luvizon, D. C., Picard, D., Tabia, H. 2d/3d pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5137-5146). 2018.

[Che18a] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7103-7112). 2018.

[Sun19a] 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 (pp. 5693-5703). 2019.

[Cai19a] Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T. J., Yuan, J., Thalmann, N. M. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2272-2281). 2019.

[Fan18a] Fang, H. S., Xu, Y., Wang, W., Liu, X., Zhu, S. C. Learning pose grammar to encode human body configuration for 3d pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1). 2018.

[Lee18a] Lee, K., Lee, I., Lee, S. Propagating lstm: 3d pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 119-135). 2018.

[Wan19a] 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 (pp. 7782-7791). 2019.

[Pav19a] Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7753-7762). 2019.

[Che21a] Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology, 32(1), 198-209. 2021.

[Liu20a] Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S. C., Asari, V. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5064-5073). 2020.

[Zen20a] Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In European Conference on Computer Vision (pp. 507-523). Springer, Cham. 2020.

[Iqb20a] Iqbal, U., Molchanov, P., Kautz, J. Weakly-supervised 3d human pose learning via multi-view images in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5243-5252). 2020.

[Mar17a] 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 (pp. 2640-2649). 2017.

[Sha21a] Shan, W., Lu, H., Wang, S., Zhang, X., Gao, W. Improving Robustness and Accuracy via Relative Information Encoding in 3D Human Pose Estimation. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 3446-3454). 2021.

[And09a] Andriluka, M., Roth, S., Schiele, B. Pictorial structures revisited: People detection and articulated pose estimation. In 2009 IEEE conference on computer vision and pattern recognition (pp. 1014-1021). IEEE. 2009.

[Ami13a] Amin, S., Andriluka, M., Rohrbach, M., Schiele, B. Multi-view pictorial structures for 3d human pose estimation. In Bmvc (Vol. 1, No. 2). 2013.

[Bel14a] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S. 3D pictorial structures for multiple human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1669-1676). 2014.

[Hos18a] Hossain, M. R. I., Little, J. J. Exploiting temporal information for 3d human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 68-84). 2018.

[Pav17a] 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 (pp. 7025-7034). 2017.

[Tek16a] Tekin, B., Rozantsev, A., Lepetit, V., Fua, P. Direct prediction of 3d body poses from motion compensated sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 991-1000). 2016.

[Par18a] Park, S., Kwak, N. 3d human pose estimation with relational networks. arXiv preprint arXiv:1805.08961. 2018.

[Wan19a] Wang, J., Huang, S., Wang, X., Tao, D. Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7771-7780). 2019.

[Zhe20a] Zheng, X., Chen, X., Lu, X. A joint relationship aware neural network for single-image 3D human pose estimation. IEEE Transactions on Image Processing, 29, 4747-4758. 2020.

[Iof15a] Ioffe, S., Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR. 2015.

[Nai10a] Nair, V., Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In Icml. 2010.

[Sri14a] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958. 2014.

[Che15a] Cheng, Y., Yu, F. X., Feris, R. S., Kumar, S., Choudhary, A., Chang, S. F. An exploration of parameter redundancy in deep networks with circulant projections. In Proceedings of the IEEE international conference on computer vision (pp. 2857-2865). 2015.

[Ion13a] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence, 36(7), 1325-1339. 2013.

[Liu18a] Liu, S., Li, Y., Hua, G. Human pose estimation in video via structured space learning and halfway temporal evaluation. IEEE Transactions on Circuits and Systems for Video Technology, 29(7), 2029-2038. 2018.

[Hua20a] Hua, G., Li, L., Liu, S. Multipath affinage stackedâhourglass networks for human pose estimation. Frontiers of Computer Science, 14(4), 1-12. 2020.

[Liu19a] Liu, S., Hua, G., Li, Y. 2.5 D human pose estimation for shadow puppet animation. KSII Transactions on Internet and Information Systems (TIIS), 13(4), 2042-2059. 2019.

[Sub17a] Su, B., Ding, X., Wang, H., Wu, Y. Discriminative dimensionality reduction for multi-dimensional sequences. IEEE transactions on pattern analysis and machine intelligence, 40(1), 77-91. 2017.

[Gao19a] Gao, Y., Ma, J., Zhao, M., Liu, W., Yuille, A. L. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3205-3214). 2019.