

Supplementary Material for Graph-Based 3D Multi-Person Pose Estimation Using Multi-View Images

Size Wu^{1,3} Sheng Jin^{2,3} Wentao Liu^{3*} Lei Bai⁴ Chen Qian³ Dong Liu¹ Wanli Ouyang⁴

¹ University of Science and Technology of China ² The University of Hong Kong

³ SenseTime Research and Tetras.AI ⁴ The University of Sydney

wsz327471010@mail.ustc.edu.cn {jinsheng, liuwentao, qianchen}@sensetime.com

baisanshi@gmail.com dongeliu@ustc.edu.cn wanli.ouyang@sydney.edu.au

1. Generalization to Different Number of Camera Views

In this section, we evaluate the generalization to the different number of camera views. Specifically, we train our graph-based models on the five-camera setup (camera id: 3, 6, 12, 13, 23), and directly evaluate these models with different number of camera views, *i.e.* the five-camera setup (camera id: 3, 6, 12, 13, 23), four-camera setup (camera id: 6, 12, 13, 23) and the three-camera setup (camera id: 6, 12, 23).

Transferring the pre-trained models to a reduced number of camera views is challenging. First, reducing the number of cameras increases the ambiguity of occluded human poses. Second, the information in the fused features is less complete. Third, the feature distribution may vary in different camera setups. We find that Tu *et al.* [1] does not produce reliable prediction results when transferring to a reduced number of camera views. For example, when the number of camera views (# Views) is reduced to 3, the mAP drops dramatically from 96.73 to 68.14, and the MPJPE increases from 17.56mm to 37.14mm. Retraining the models with the test-time camera setups will mitigate this problem (marked with †). In comparison, our approach can better generalize to different camera setups *without* any fine-tuning. Although reducing the number of camera views will reduce the accuracy, we show that we still achieve reasonably good results, demonstrating that our proposed approach has strong generalization ability. For example, with only 3 camera views, we achieve 91.60mAP and 94.14mAR. We also show that our approach consistently outperforms the state-of-the-art approach [1] on generalization to different camera setups.

Table 1. Generalization to different number of camera views. All results are obtained using ResNet-50 as the backbone. † means the higher score the better, while ‡ means the lower the better. † means fine-tuning models under the test-time camera setups.

	#Views	mAP †	mAR †	MPJPE ‡
Tu <i>et al.</i> [1]	5	96.73	97.56	17.56mm
Ours	5	98.10	98.70	15.84mm
Tu <i>et al.</i> [1]	4	94.54	95.97	20.06mm
Tu <i>et al.</i> [1]†	4	95.60	96.80	18.63mm
Ours	4	97.65	97.89	17.87mm
Tu <i>et al.</i> [1]	3	68.14	72.14	37.14mm
Tu <i>et al.</i> [1]†	3	89.26	93.91	24.02mm
Ours	3	91.60	94.14	22.69mm

2. Network Architecture

In this section, we illustrate the detailed graph model architectures of MMG, CRG and PRG in Figure 1.

As shown in Figure 1 (a), the graph model of MMG consists of two layers of EdgeConv-E followed by two fully-connected (FC) layers. The input visual features \mathbb{R}^{512} extracted in the image plane are first updated by the EdgeConv-E layers. Then the fully-connected layers, whose input is the concatenation of target vertex feature and relative source vertex feature, predict whether an edge is connecting 2D centers of the same person.

As shown in Figure 1 (b), the input features come from three sources: (1) the visual features \mathbb{R}^{512} extracted in the image plane (2) the normalized 3D coordinates \mathbb{R}^3 of the query point (3) 2D center confidence score from the 2D backbone \mathbb{R}^1 . They are processed by fully-connected layers and then concatenated to produce a feature vector \mathbb{R}^{545} for each vertex. The features are then processed by three layers of EdgeConv for cross-view feature message passing. A max-pooling layer is used for feature fusion and fully-connected layers to predict the center confidence score. To facilitate training, we adopt residual connections in between

*Corresponding author.

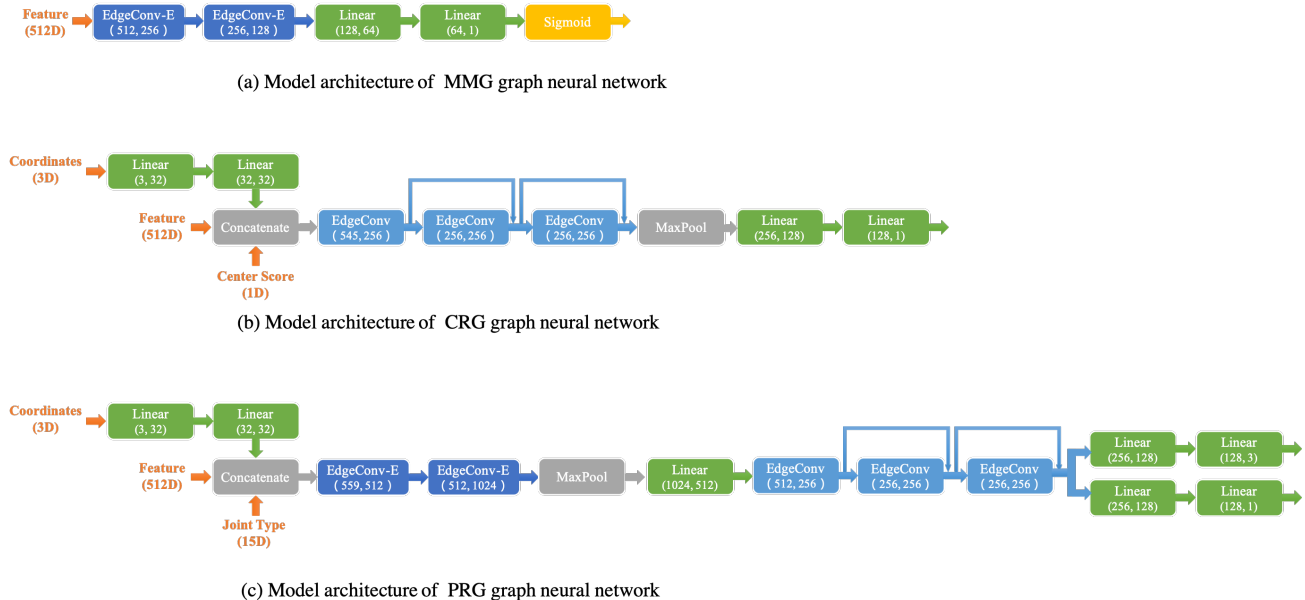


Figure 1. The model architectures of (a) MMG, (b) CRG and (c) PRG. ‘Linear’ denotes the fully-connected layer, and ‘MaxPool’ denotes the graph max-pooling layer. The input feature dimensions and the output feature dimensions are illustrated.

the EdgeConv layers.

As shown in Figure 1 (c), the input features also come from three sources: (1) the visual features \mathbb{R}^{512} extracted in the image plane (2) the normalized 3D coordinates \mathbb{R}^3 of each joint in the initial pose (3) one-hot feature of the joint type \mathbb{R}^{15} . They are processed by fully-connected layers and concatenated to produce a feature vector \mathbb{R}^{559} for each vertex. The features are then processed by two layers of EdgeConv-E for cross-view message passing. Then a max-pooling layer is applied to aggregate the cross-view features and coarsen the graph. The max pooled features are updated by the following three EdgeConv layers via effective information flow between the body joints. Similar to CRG, we add some residual connections to help model training. Finally, the extracted features are passed to two parallel MLPs (multi-layer perceptrons) to respectively regress a refinement vector and predict a confidence score for each joint. Both MLPs are composed of two fully-connected layers.

3. Qualitative Comparisons

In this section, we present more qualitative comparisons with Tu *et al.* [1] on the CMU Panoptic dataset (a, b, c) and the Shelf dataset (d).

As shown in Figure 2 (a), the arm of the man (purple) is only visible in two camera views, and is occluded by other people or by himself in most views. This results in large 3D pose errors for Tu *et al.* [1]. Our proposed PRG can fix such kinds of pose errors, by considering both the geometric constraints and the human body structural relations.

As shown in Figure 2 (b), many joints of the man (blue)

are self-occluded by his own body in many camera views. This makes the visual features unreliable, leading to false negatives (FN) for Tu *et al.* [1]. In comparison, our proposed MMG and CRG learn to detect human centers in a coarse-to-fine manner via GCN. We are able to obtain more robust human detection results. As shown in Figure 2 (c), accurately predicting the poses of the little child (green) is challenging, due to insufficient training data. This example indicates that our proposed approach has better generalization ability towards rare poses.

As shown in Figure 2 (d), there is a false positive pose in the red circle estimated by Tu *et al.* [1]. In comparison, our approach achieves better performance and gets fewer false positives. Our proposed CRG together with PRG can suppress these false positives, by considering the multi-view features as a whole via GCN.

References

- [1] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3

