Towards More Stable Human Pose Estimation with Cross-View Attention and Foot Stabilization
Supplementary File

Li'an Zhuo*, Jian Cao*, Qi Wang, Bang Zhang, Liefeng Bo
Alibaba Group
{lianzhuo.zla, tanfeng.cj, wilson.wq, zhangbang.zb, liefeng.bo}@alibaba-inc.com

Section 1 illustrates more details about the RKTD. More comparisons and qualitative results are presented in Section 2.

1. Difference between KTD and RKTD

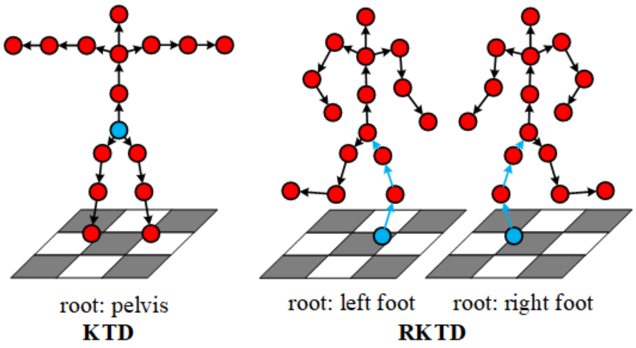

Figure 1. Difference between KTD and RKTD. The root node (blue nodes) in KTD is the pelvis, while the root node in our proposed RKTD is determined by the contact state, leading to a reversible kinematic tree (blue lines).

We further illustrate the difference between KTD and our proposed RKTD in Figure 1. KTD has a fixed kinectic tree, whose root node is the pelvis, while our proposed RTKD determine the root node by the contact states divided into three cases:

$$n_{\mathrm{root}} = \begin{cases} n_{\mathrm{l}} & \text{if } c_l > c_r > 0, \\ n_{\mathrm{r}} & c_r > c_l > 0, \\ n_{\mathrm{p}} & \text{otherwise}, \end{cases} \tag{1}$$

where n_{root}, n_l, n_r and n_p represents the root, the left foot, right foot and pelvis node separately. We then estimate each pose following the order in the corresponding kinectic tree, which takes the poses of the ancestor nodes into consideration.

2. More results

Table 1 is a comparison of the inference speed of our models and other works on a single NVIDIA A100 GPU with 1 batch size, which shows that our model not only achieves state-of-the-art accuracy, but also has good inference speed.

Method	3DPW		Batch 1 FPS
	MPJPE	PA-MPJPE	
MAED [2]	79.1	45.7	39
CLIFF [1]	**69.0**	43.0	22
Ours-Base	77.8	44.7	**77**
Ours-Large	70.8	**40.1**	46

Table 1. Comparison of the inference speed of our model with previous works.

Table 2 presents the accuracy of our model in contact prediction. Both our base model and the large model achieve very high precision and recall, with the F1 score above 0.90.

Method	AIST++		
	Precision	Recall	F1
Ours-Base	0.87	0.95	0.91
Ours-Large	0.89	0.96	0.92

Table 2. Foot-contact prediction accuracy for our model

More qualitative results about the foot pose reconstruction and our method are shown in Figure 2.

References

[1] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 1

* indicates the equal contributions.

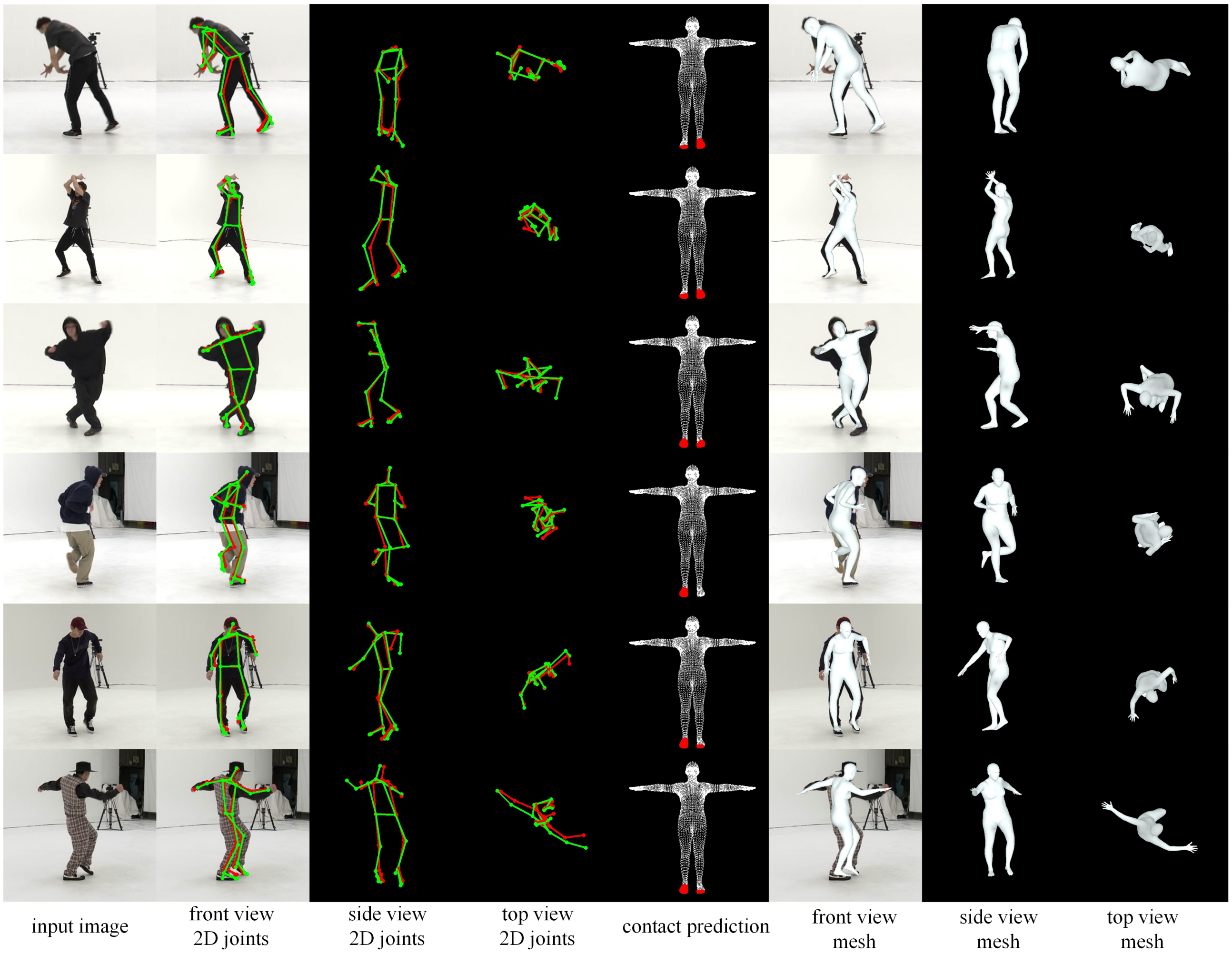

Figure 2. Qualitative results on AIST++ test set. From left to right: input image, tri-view 2D joints, tri-view 3D mesh (green for the ground truth and red for prediction). Additionally, the results about the foot pose reconstruction are reflected in the tri-view ground truth 2D joints.

[2] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *ICCV*, 2021. 1