Supplementary Material: Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network

Anonymous CVPR submission

Paper ID 5180

Results based on ground truth 2d joints Following [3], we use the ground truth 2d joints provided by the Human3.6 dataset instead of the output from the stacked hourglass as the input. The results are shown in Table 1, where our approach achieves the best performance. Moreover, the MPJEP improves by 14.9 mm compared to the results (52.7 mm) of using the 2d joint detections from the stacked hourglass.

More visualizations of our mixture density model We show more visualizing results of our mixture dentity model in Fig. 1. The level of ambiguity increases from top to bottom. There is no occlusion for the "standing" pose in the first row, hence, all the outputs from our mixture density model looks similar. As increasing number of joints are occluded (second to the fourth row), the output of each kernel becomes increasingly different. The last row shows a failure case of our model where none of the five outputs looks similar to the ground truth pose. The reason is that the output 2d pose from the stacked hourglass (the first column) is totally wrong, thus our mixture density model cannot recover the 3d pose from the wrong input. Moreover, we can see that our model try to generate 3d pose hypotheses that are consistent in 2d projections in all cases.

More qualitative results on MPII dataset We show more qualitative results on the MPII dataset in Fig. 2 to demonstrate the generalization capacity of our model. We choose images where the poses are not common in the Human3.6 dataset.

The *sum-exp-trick* used in the training process The loss function $\mathcal{L}_{3D}$ in Eqn. (10) in our paper for one pair of 2D joints and 3D poses $\{\mathbf{x}, \mathbf{y}\}$ can be expressed as:

$$\mathcal{L}_{3D} = -\ln \sum_{i=1}^{M} \frac{\alpha_i(\mathbf{x})}{(2\pi)^{d/2}\sigma_i(\mathbf{x})^d} \exp - \frac{\|\mathbf{y}-\mu_i(\mathbf{x})\|^2}{2\sigma_i(\mathbf{x})^2}. \tag{1}$$

Note that we omit the parameters of the deep network $\mathbf{w}$ for brevity. The right-most term is an exponential of values which tend to be very small, and this results in an underflow problem after applying the logarithm. We prevent the underflow problem by applying the *log-sum-exp* trick. Specifically, a logarithm of a sum of exponential terms can be expressed as:

$$\ln \sum_{i=1}^{n} \exp t_i = \max_i(t_i) + \ln \sum_{i=1}^{n} \exp\left(t_i - \max_i(t_i)\right). \tag{2}$$

We can extend the exponential function within the logarithm in Eqn. (1) to get:

$$\begin{aligned} \mathcal{L}_{3D} = -\ln \sum_{i=1}^{M} \exp\{\ln \alpha_i(\mathbf{x}) - \frac{d}{2} \ln 2\pi\sigma_i^2(\mathbf{x}) \\ - \frac{\|\mathbf{y}-\mu_i(\mathbf{x})\|^2}{2\sigma_i(\mathbf{x})^2}\}, \end{aligned} \tag{3}$$

which we can then apply the *log-sum-exp* trick expressed in Eqn. (2).

References

[1] M. R. I. Hossain and J. J. Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, pages 69–86. Springer, Cham, 2018. 2

[2] K. Lee, I. Lee, and S. Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–135, 2018. 2

[3] 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*, 2017. 1, 2

[4] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1561–1570. IEEE, 2017. 2

Table 1: Quantitative results of MPJPE in millimeter on the Human3.6M when the input is the ground truth 2d joints.

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Moreno et al.[4]	53.5	50.5	65.7	62.4	56.9	80.8	60.6	50.8	55.9	79.6	63.6	61.8	59.4	68.5	62.1	62.1
Martinez et al.[3]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Lee et al.[2]	34.6	39.7	37.2	40.9	45.6	50.5	42.0	39.4	47.3	48.1	39.5	38.0	31.9	41.5	37.2	40.9
Hossain et al.[1]	35.2	40.8	37.2	37.4	43.2	**44.0**	38.9	35.6	**42.3**	**44.6**	39.7	39.7	**40.2**	32.8	35.5	39.2
Ours	**31.1**	**38.2**	**33.5**	**35.5**	**39.1**	46.3	**35.6**	**34.6**	45.9	50.7	**39.4**	**36.1**	40.3	**29.6**	**31.1**	**37.8**

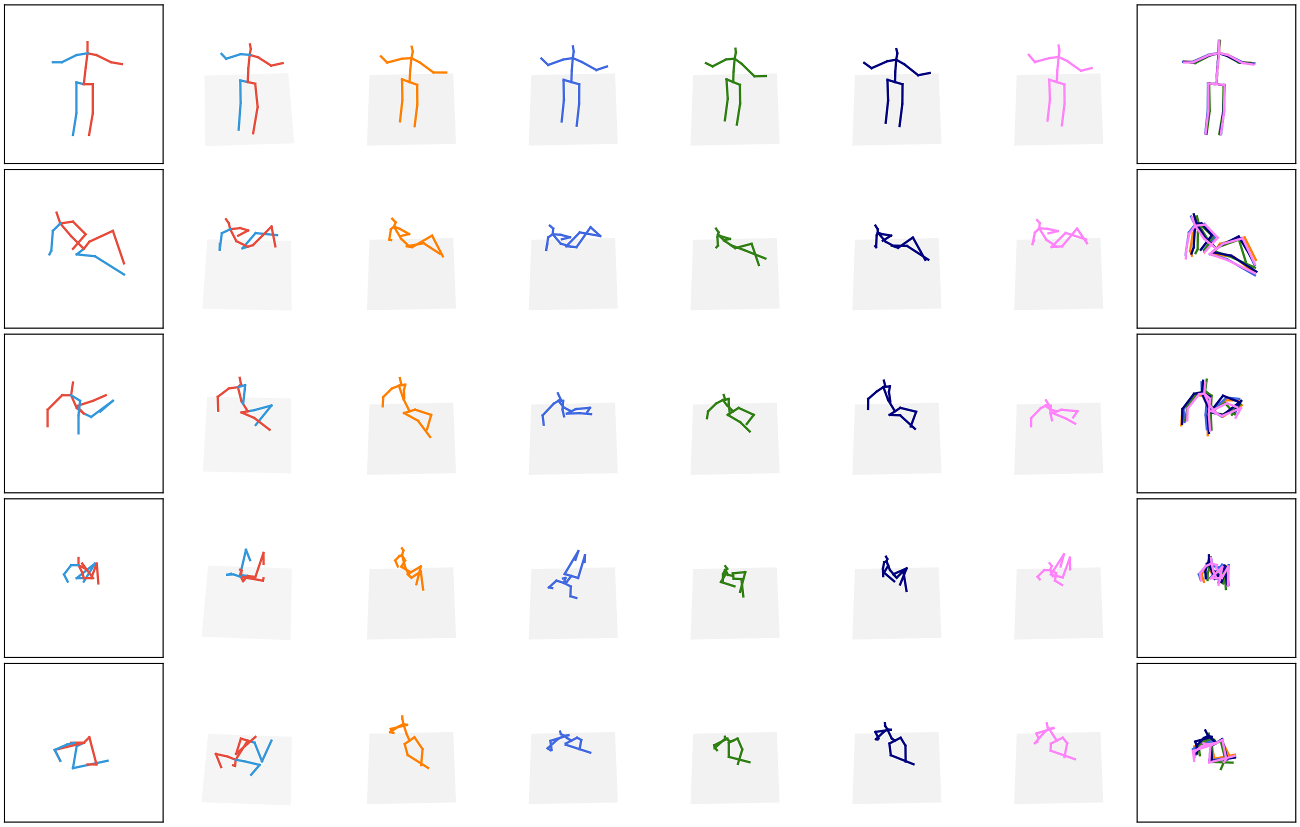

Figure 1: 3D Pose hypotheses generated by our network. The first column is the input of our network, i.e. the 2D joints estimated by the stacked hourglass network. The second column is the ground truth 3D pose, and the third to seventh columns are the hypotheses generated by our network. The last column is the 2D reprojections of all five hypotheses. The corresponding 2D projection and 3D pose are drawn in the same color. (Best view in color)

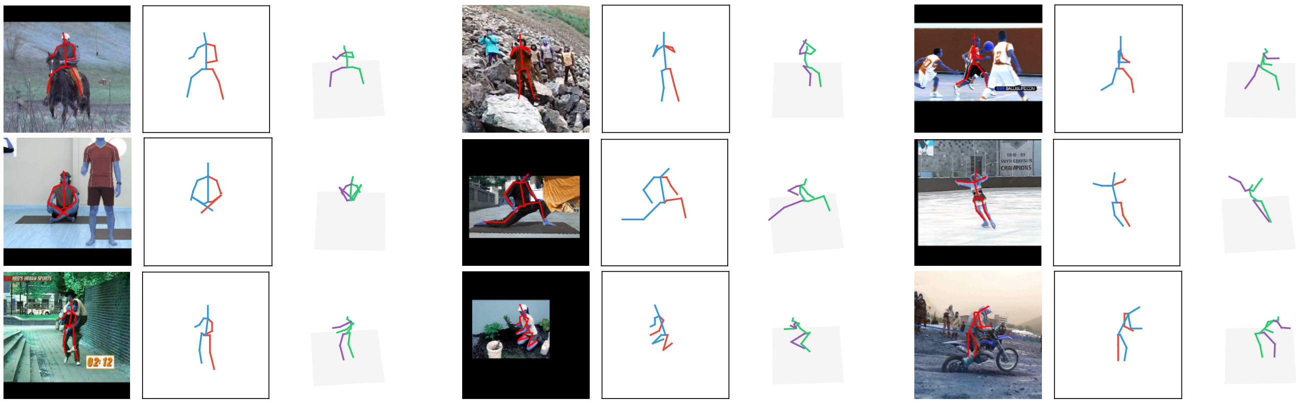

Figure 2: Qualitative results on the MPII test set. The first and second columns are the input images and output 2D joint detections of the stacked hourglass network, the last column is the 3D pose generated by our network