Supplementary Material
Learning Monocular 3D Human Pose Estimation from Multi-view Images

Helge Rhodin[1] Jörg Spörri[2,4] Isinsu Katircioglu[1] Victor Constantin[1]
Frédéric Meyer[3] Erich Müller[4] Mathieu Salzmann[1] Pascal Fua[1]

[1]CVLab, EPFL, Lausanne, Switzerland [2]Balgrist University Hospital, Zurich, Switzerland
[3]UNIL, Lausanne, Switzerland [4]University of Salzburg, Salzburg, Austria

Ski Dataset. The capture process for the ski dataset was as follows. The global locations of the cameras and of static reference markers (black spheres) were geodetically measured by a tachymeter theodolite (Leica Total station 1200). A trained annotator clicked on the visible reference markers in each view and frame to determine the camera pan, tilt and zoom. In the same way, 22 human joints were marked at their visual center and triangulated. The average Euclidean error of such procedure is 23 ± 10 mm, determined by comparing the photogrammetric marker estimates with the theodolite ones [1]. Possible click-inaccuracies were mitigated by normalizing bone length to be constant at their anthropometrically measured values [2]. Reprojection of the 3D annotation shows a very accurate overlap with the images, as visualized in the main paper. We make the dataset available to facilitate future work towards reliable monocular pose reconstruction (cvlab.epfl.ch/SkiPosePTZ-Dataset).

Action-Specific Evaluation on Human3.6M. In the main paper, we have shown that our weakly-supervised approach based on multi-view images for training consistently improves the accuracy of monocular human pose estimation across datasets. For the H36M dataset, we reported the errors averaged over all actions. Here, by contrast, we provide the NMPJPE for each individual actions of the H36M dataset. We refer the reader to the main paper for the definition of the notation and of the baseline. As can be seen from Table 1 and Fig. 1, our approach yields a consistent improvement over the baseline for all actions, with particularly high gains for the Sit (10 mm), Eat (7mm), and Walk (6 mm) classes. This confirms the qualitative and quantitative results reported in the main paper.

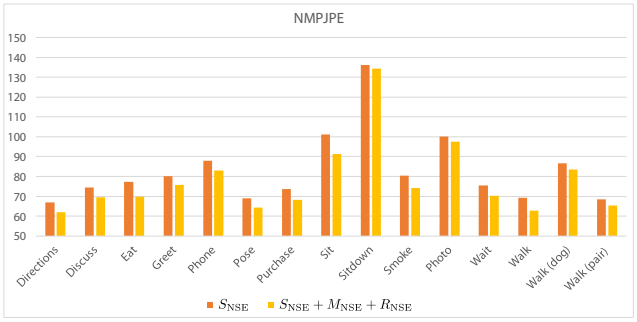

Figure 1. **Per-class accuracies on H36M.** We report the NMPJPE metric (in mm) for the baseline (S_{NSE}, orange) and our method ($S_{\text{NSE}} + M_{\text{NSE}} + R_{\text{NSE}}$, yellow) for every action of H36M. Note that our weakly-supervised scheme, leveraging multiple views during training only, consistent outperforms the baseline.

References

[1] M. Klous, E. Müller, and H. Schwameder. Collecting Kinematic Data on a Ski/snowboard Track with Panning, Tilting, and Zooming Cameras: Is There Sufficient Accuracy for a Biomechanical Analysis? *Journal of Sports Sciences*, 28(12):1345–1353, 2010. 1

[2] G. Smith. An iterative segment length normalization routine for use with linked segment models. In *Conference Proceedings of the18th Annual Meeting of the American Society of Biomechanics*, pages 35–6, 1994. 1

Method	Directions	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	Sitdown	Smoke	Photo	Wait	Walk	Walk Dog	Walk Pair	Total
S_{NSE}	67.0	74.5	77.3	80.1	87.8	69.1	73.6	101.0	136.0	80.5	100.0	75.4	69.1	86.7	68.5	83.4
$S_{\text{NSE}} + M_{\text{NSE}} + R_{\text{NSE}}$	**61.9**	**69.4**	**69.9**	**75.8**	**82.9**	**64.3**	**68.1**	**91.2**	**134.4**	**74.2**	**97.5**	**70.2**	**62.8**	**83.4**	**65.4**	**78.2**
Improvement	5.0	5.0	7.4	4.3	5.0	4.8	5.5	9.9	1.7	6.3	2.5	5.2	6.3	3.3	3.1	5.2

Table 1. Activity-wise NMPJPE scores in mm. As in Fig. 1, our method ($S_{\text{NSE}} + M_{\text{NSE}} + R_{\text{NSE}}$) is compared to the baseline (S_{NSE}).