Delving Deep Into Hybrid Annotations for 3D Human Recovery in the Wild
Supplemental Material

Yu Rong[1] Ziwei Liu[1] Cheng Li[2] Kaidi Cao[4] Chen Change Loy[3]

[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong

[2]SenseTime Research [3]Nanyang Technological University [4]Stanford University

{ry017, zwliu}@ie.cuhk.edu.hk chengli@sensetime.com

kaidicao@cs.stanford.edu ccloy@ntu.edu.sg

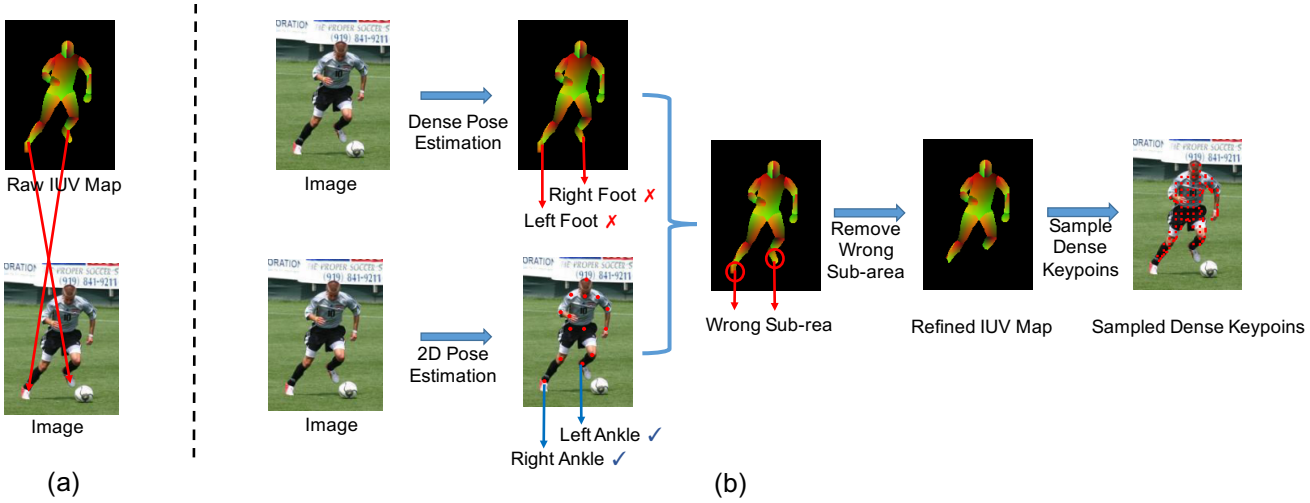

Figure 1: **Process of refining IUV Map.** Figure (a) demonstrates that the raw IUV map might contain errors. Figure (b) shows the process of refining the IUV maps. The generated IUV map is compared with the 2D keypoints. If they are not consistent, *e.g.*, the sub-area around "right ankle" is predicted as "left foot", then we discard this sub-area by assigning it as background. We compare each keypoint with the predicted IUV maps surrounding it and remove the inconsistent part.

1. Sampling Dense Keypoints

Since dense keypoint annotations are only available in COCO-DensePose dataset and training models purely using sparse 2D keypoints will lead to suboptimal results, we present an effective method for generating dense keypoints for other in-the-wild images that only annotated with sparse 2D keypoints. An effective way is to directly sample points from the IUV maps produced by the DensePose model.

The dense points drawn from IUV maps cannot be employed directly since the maps frequently contain wrong predictions. As Figure 1 (a) shows, the left foot is wrongly predicted as the right foot while the right foot is predicted as the opposite. To avoid erroneous points corrupting the learning of our model, we perform sampling of dense points by using accurate sparse keypoints as reference. Specifically, for each visible 2D keypoint, we check the values of IUV map in the 3×3 grid centering at it and select the value of 'I' (which indicates body part) that appears most frequently as the body part prediction of IUV map surrounding this keypoint. Then we chech whether the *surrounding IUV* is consistent with the 2D keypoint. For example, if a keypoint is labeled as "right ankle" but the *surrounding IUV* is "left foot", then this sub-area is assigned as erroneous region.

After finding the erroneous region, our sampling scheme will set the IUV map of this sub-area to be background in a recursive manner: We first set the IUV value of the keypoint to be background, then we check the 3×3 grid around it and determine the pixels whose value of 'I' equals to the *surrounding IUV* and set their IUV values to be background. Further, we check the 3×3 grids centering at these

Table 1: **Influence of 3D annotations.** This table lists detailed experiment results of Figure 4 of main paper.

Kept 3D Annotations (%) → Input ↓	100	80	60	40	20	10	5	1	0
IUV Map	125.2	125.9	128.3	132.3	133.6	136.8	144.0	144.3	191.5
Body Segment	124.8	126.7	128.9	131.3	132.3	135.9	143.0	148.5	196.7
Image	127.4	128.4	132.2	134.6	136.0	143.3	149.9	152.2	203.2
Image & IUV	125.5	126.2	130.1	131.6	135.3	135.9	140.6	148.0	197.0
Image & Body Segment	125.8	126.1	129.5	131.4	133.7	136.5	143.3	148.0	196.4

Table 2: **Influence of constrained 3D annotations.**. The inputs of the models are all single images.

Other Supervisions →	100% 3D & Sparse 2D	20% 3D & Sparse 2D	Dense & Sparse 2D	Sparse 2D Only
with Constained 3D	127.4	137.7	137.3	203.2
w/o Constrained 3D	128.9	138.1	173.4	230.9

pixels and determine more pixels using the same condition. The process is conducted recursively until there are no more pixels found. The above process is conducted on each keypoint to refine the whole IUV map before we use the map as the complementary input and for sampling dense keypoints. The sampling process is depicted in Figure 1 (b).

2. Implementation Details

In this section, we discuss more implementation details. In the training phase, the whole model is first pretraind using 3D data from Human3.6M dataset [2], then it is finetuned on the COCO-DensePose [1], UP-3D [4] and 3DPW [5]. For COCO-DensePose dataset, we train our model with ground truth dense keypoints and 2D keypoints. For UP-3D and 3DPW dataset, our model is trained with the combination of 3D annotations,2D keypoints and sampled dense keypoints. The sampled dense keypoints are obtained based on the method described in Section 1.

In the training phase, the batch size is set to 128. Adam optimizer [3] with $1e-4$ is adopted in the whole training phase. The model gets converged after $40 \sim 50$ epochs. Especially, if all the losses including 3D, dense and 2D are used in training, their balance weights are 10, 1, 10, respectively. If only two losses are used, their balance weights are set to be both 10.

3. Efficiency of 3D Annotations.

Detailed experiment results. Detailed experiment results in Figure 4 of the main paper is listed in Table 1. In experiments, the amount of paired 3D annotations used in the training phase is reduced gradually from 100% to 0% (0% means only using sparse 2D annotations in training). From the table, we find that 3D annotations are quite efficient. The reconstruction error only increases by 6% when 80% 3D annotations are excluded from training.

Influence of constrained 3D. We also investigate constrained annotations. The experiment results are listed in Table 2. When paired in-the-wild 3D annotations exist, using constrained 3D annotations barely brings improvement. However, when there are no paired in-the-wild 3D annotations exist, incorporating constrained 3D annotations into training improves the performance of models bys 30%.

References

[1] Rza Alp Gler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 2

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2

[4] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2

[5] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2