Supplementary Material for "NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets"

Gyeongsik Moon[1] Hongsuk Choi[1] Kyoung Mu Lee[1,2]

[1]Dept. of ECE & ASRI, [2]IPAI, Seoul National University, Korea
{mks0601,redarknight,kyoungmu}@snu.ac.kr

In this supplementary material, we present more experimental results that could not be included in the main manuscript due to the lack of space.

A. Qualitative results

A.1. 3D body-only pseudo-GTs

Figure A (a) shows NeuralAnnot produces far better 3D body pseudo-GTs than SMPLify-X. In particular, it produces much better results when the poses in input images have truncations and complicated articulations. Figure A (b) shows our NeuralAnnot is highly robust to occlusions and truncation in crowd scenes of CrowdPose [5]. Figure A (c) shows additional results on Human3.6M [2] and MPI-INF-3DHP [7].

A.2. 3D hand-only pseudo-GTs

Figure B shows 3D hand pseudo-GTs of our NeuralAnnot on InterHand2.6M [8]. It successfully produces 3D pseudo-GTs from highly complicated interacting hand images.

A.3. 3D face-only pseudo-GTs

Figure C shows that NeuralAnnot and SMPLify-X produce similar 3D face pseudo-GTs. Unlike the body and hand parts, the face part does not involve complicated articulation, which makes SMPLify-X work well and produce similar results to those of NeuralAnnot.

A.4. 3D whole-body pseudo-GTs

Figure D (a) shows NeuralAnnot produces much better expressive whole-body 3D pseudo-GTs than SMPLify-X on MSCOCO. Figure D (b) shows more qualitative results of NeuralAnnot on MSCOCO.

B. Running SMPLify-X on 3D joint coordinates

As original SMPLify-X does not consider 3D joint coordinates during the optimization, we modified it to consider 3D joint coordinates for 3D pseudo-GTs of Human3.6M [2], MPI-INF-3DHP [7], and InterHand2.6M [8], which provide GT 3D joint coordinates. To this end, we made two modifications.

Camera initialization. We initialize extrinsic camera parameters R and t using hip and shoulder 3D joint coordinates by performing SVD. For the 3D pseudo-GTs of hands, we use five hand joints, which include the wrist, index root, middle root, ring root, and pinky root. R and t represent a 3D rotation matrix and 3D translation vector, respectively, from a human model coordinate system to a dataset coordinate system. We chose the hip and shoulder joints or the five hand joints as they can roughly decide the 3D global rotations of the human body or hands, respectively, while end-point joints (*e.g.*, wrists and ankles for the body and fingertips for the hands) cannot.

3D data term. We changed the 2D data term of SMPLify-X to the 3D data term, which calculates a distance between the GT 3D joint coordinates and 3D joint coordinates from a mesh. The 3D joint coordinates from a mesh are obtained by a joint regression matrix, defined in human models. We use a Geman-McClure error function [1] for the distance used in the 2D data term of the original SMPLify-X. We tried several other distances, such as $L1$ and $L2$, and found that Geman-McClure error function [1] and $L1$ work the best. The distance is calculated in a meter scale, and we set the weight of the data term to 10^6, which works the best.

C. Qualitative comparisons between two-stage NeuralAnnot and EFT

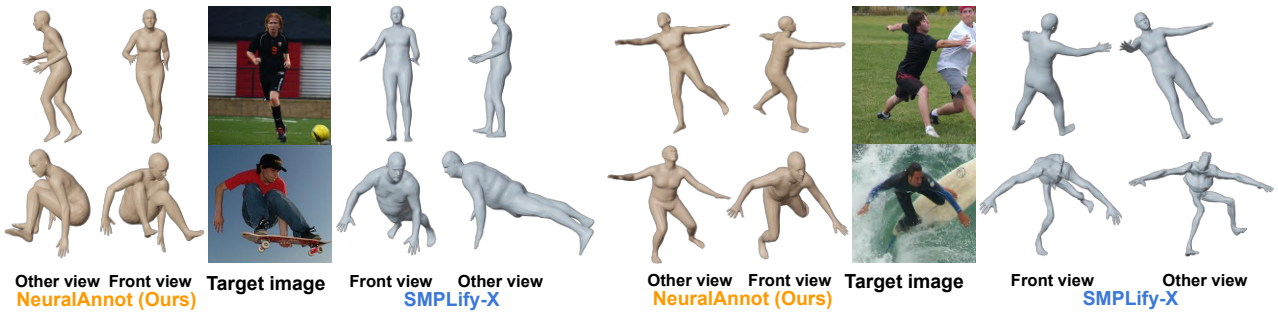
In Table 7 of the main manuscript, we showed our two-stage NeuralAnnot produces more beneficial 3D pseudo-GTs than previous two-stage annotators, such as SPIN [4]

and EFT [3]. To this end, we changed our NeuralAnnot to a two-stage annotator by training it on initial 3D pseudo-GTs of Human3.6M, MPI-INF-3DHP, and MSCOCO, where the initial 3D pseudo-GTs are obtained by our original one-stage NeuralAnnot. Figure E shows that changing our original one-stage NeuralAnnot to a two-stage annotator can correct possibly wrong 3D pseudo-GTs to better ones. Figure F shows that our two-stage NeuralAnnot produces better 3D pseudo-GTs than EFT.

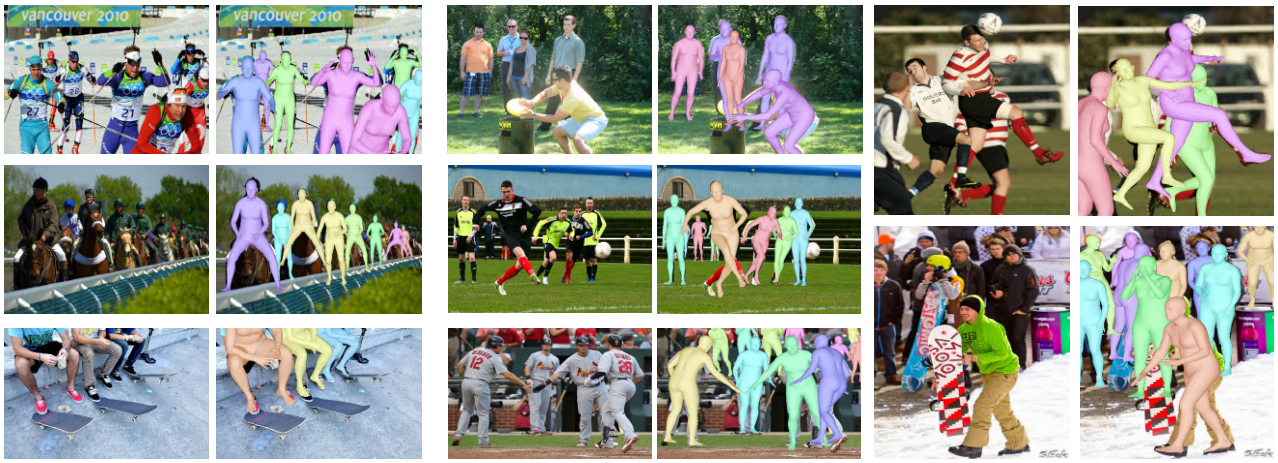
We believe there are two reasons why NeuralAnnot produces better ones. First, EFT is based on a pre-trained SPIN network, trained on 3D pseudo-GTs of SMPLify-X. SMPLify-X often suffers from inaccurate 3D pseudo-GTs, as shown in Figure 4 of the main manuscript and Figure A (a), which can affect 3D pseudo-GTs of SPIN and EFT. Second, as EFT fine-tunes a pre-trained SPIN network to 2D joint coordinates of each sample, it might produce inaccurate 3D pseudo-GTs when the input image has truncated or invisible joints. During the fine-tuning, their network can be overfitted to partial joints of a sample as truncated or invisible joints do not have 2D joint coordinates. Hence, the pre-trained SPIN network can be corrupted and produce wrong 3D poses for truncated or invisible joints as no supervisions are applied for those joints. On the other hand, our NeuralAnnot's network is not optimized for a specific sample; instead, it is optimized for entire samples of datasets. Therefore, it does not suffer from the overfitting to a specific sample and produces robust 3D pseudo-GTs when input images have truncated or invisible joints. The effects of the truncated or invisible joints are shown in the 1) first row and first column, 2) first row and second column, and 3) third row and first column of the Figure F. Figure A (b) additionally shows that NeuralAnnot produces robust 3D pseudo-GTs under severe truncations.

License of the Used Assets

- MSCOCO dataset [6] belongs to the COCO Consortium and are licensed under a Creative Commons Attribution 4.0 License.
- InterHand2.6M dataset [8] is CC-BY-NC 4.0 licensed.
- Human3.6M dataset [2]'s licenses are limited to academic use only.
- MPI-INF-3DHP dataset [7] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- 3DPW dataset [10] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- FreiHAND dataset [11] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- CrowdPose dataset [5] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- SMPLify-X [9] codes are released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- EFT [3] codes are CC-BY-NC 4.0 licensed.
- SPIN [4] codes are released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.



(a) 3D body-only pseudo-GT comparisons on MSCOCO



(b) 3D body-only pseudo-GTs on CrowdPose

(c) 3D body-only pseudo-GTs on Human3.6M and MPI-INF-3DHP

Figure A. (a) Qualitative comparisons between 3D body pseudo-GTs of NeuralAnnot and SMPLify-X on MSCOCO. (b) Visualized 3D body pseudo-GTs of NeuralAnnot on CrowdPose. (c) Visualized 3D body pseudo-GTs of NeuralAnnot on Human3.6M and MPI-INF-3DHP.

Figure B. Visualized 3D hand pseudo-GTs of NeuralAnnot on InterHand2.6M.

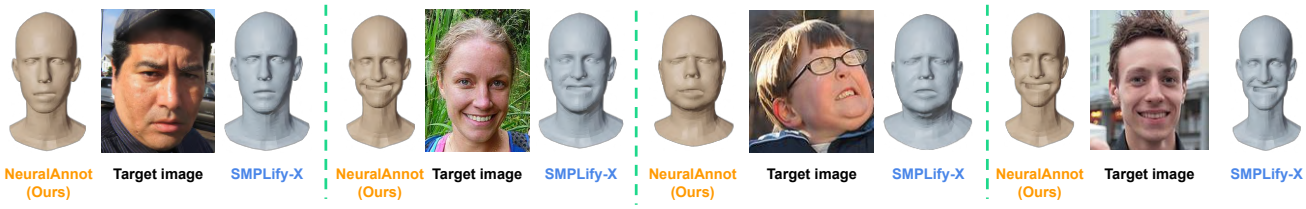
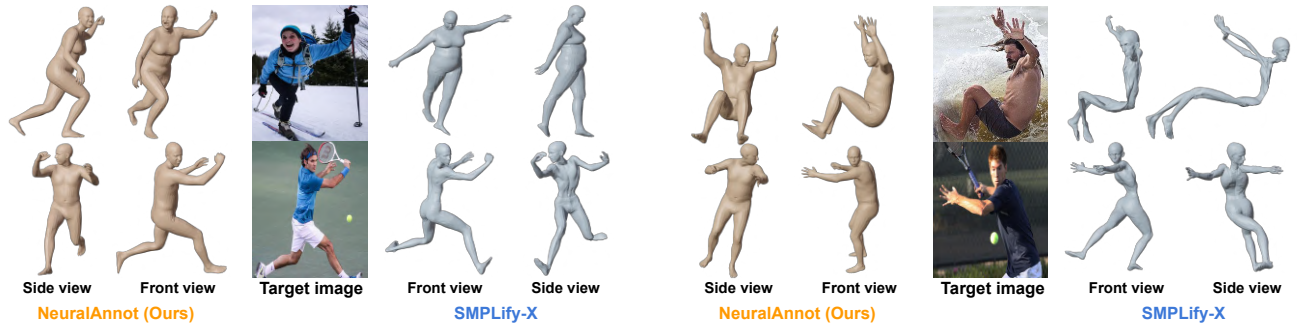


Figure C. Qualitative comparisons between 3D face pseudo-GTs of NeuralAnnot and SMPLify-X on MSCOCO. We normalized the global rotation of the face for visualization purpose.



(a) 3D expressive whole-body pseudo-GTs comparisons on MSCOCO

(b) 3D expressive whole-body pseudo-GTs on MSCOCO

Figure D. Visualized expressive whole-body 3D pseudo-GTs of NeuralAnnot on MSCOCO.

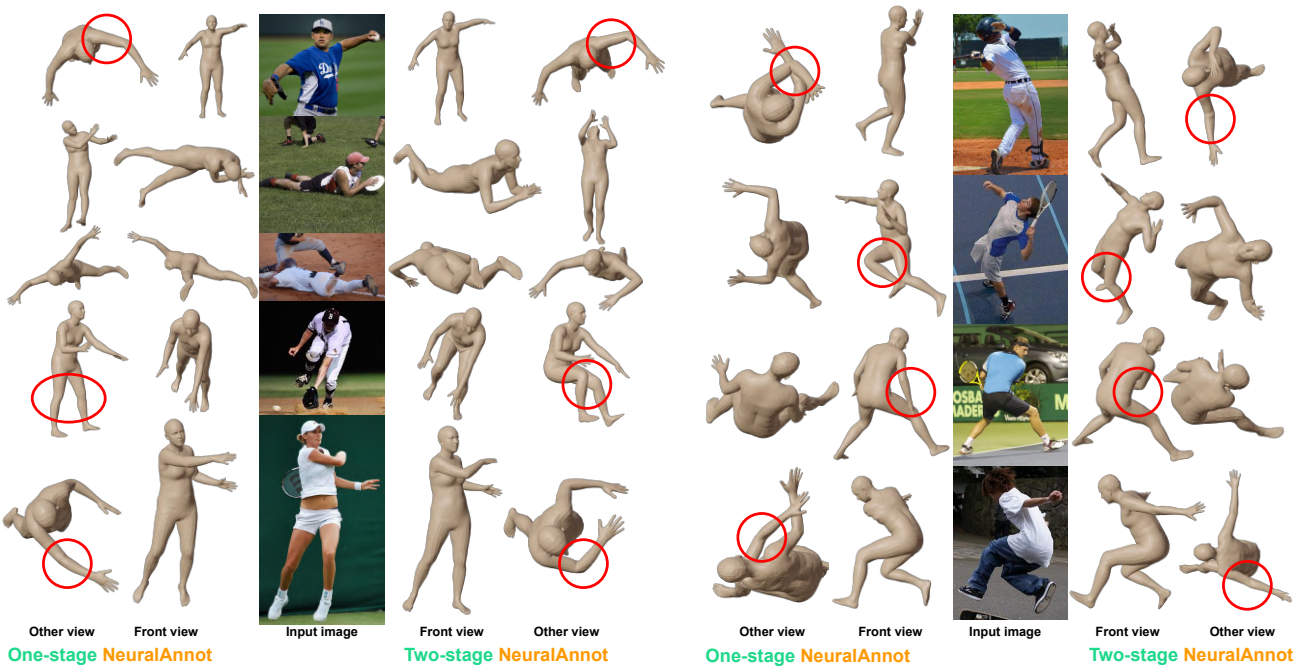


Figure E. Qualitative comparison between one-stage NeuralAnnot and two-stage NeuralAnnot on MSCOCO.

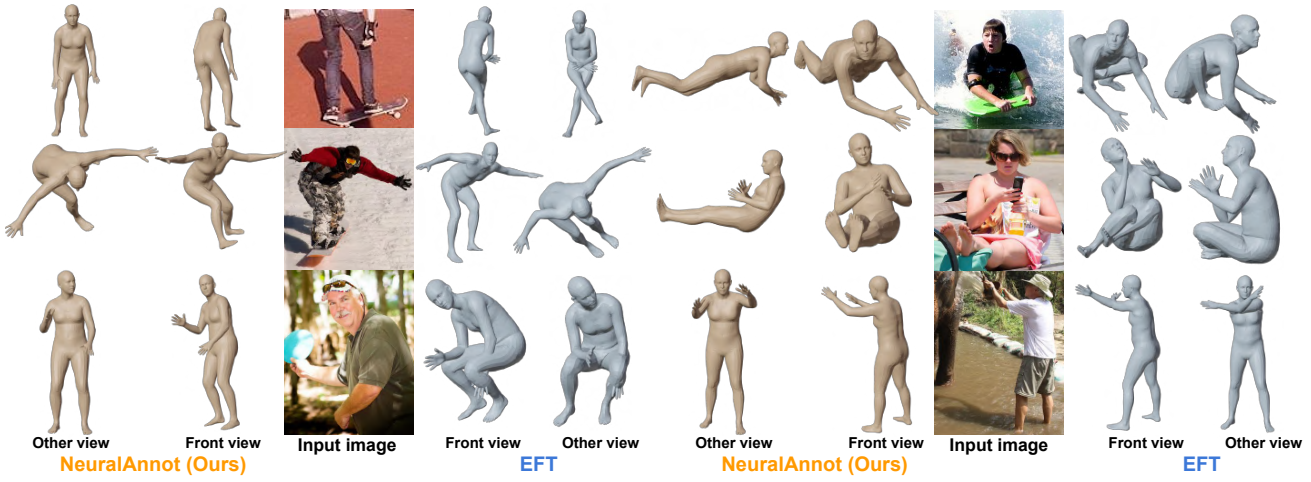


Figure F. Qualitative comparisons between NeuralAnnot and EFT on MSCOCO.

References

[1] Stuart Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 1987.

[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*, 2014.

[3] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020.

[4] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[5] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[7] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

[8] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 2020.

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019.

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

[11] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019.