Supplementary Material *for* Implicit 3D Human Mesh Recovery using Consistency with Pose and Shape from Unseen-view

Hanbyel Cho Yooshin Cho Jaesung Ahn Junmo Kim
Korea Advanced Institute of Science and Technology (KAIST), South Korea
{tlrl4658,choys95,jaesung02,junmo.kim}@kaist.ac.kr

1. Supplementary material

1.1. Implementation Details

Implementation of ImpHMR. Following the existing single image pose and shape estimation methods [6, 8, 10, 11], we use ResNet-50 [2] as the backbone model for our image encoder g. Note that, because ImpHMR is trained *end-to-end* manner, the weights of the image encoder g are *also updated* during training. For the feature fields module, foreground attention $\mathcal{A}$ adopts the spatial attention proposed in [20] to focus on human-related features. The neural feature field represented by the MLP network h adopts the object feature field suggested in [17]. For the MLP h, the size of the input latent and output feature vector is changed to 2048. For HMR tasks, since it is not necessary to infer the texture of humans, we condition h with a single latent vector $\mathbf{z}_{\text{fg}}$. For volume rendering, we sample 32 points on a ray direction $\mathbf{r}$, and the spatial resolution of the rendered 2D feature map is fixed to 4×4. In positional encoding, the frequency octave is set to 10 and 4 for $\mathbf{x}$ and $\mathbf{r}$, respectively. The regressor $\mathcal{R}$ has the same architecture as the regressor in SPIN [11] and is initialized with a pre-trained model in [11]. The maximum iteration number of $\mathcal{R}$ is set to 3 as in [8, 9, 11]. For the geometric guidance branch, deconvolution $\mathcal{D}$ is composed of 5 layers of transposed convolution along with BN [3] and ReLU activation, and outputs a 128×128 resolution silhouette. During training, we use Adam [7] optimizer with batch size 64. The λ_{2d}, λ_{3d}, λ_{pose}, λ_{shape}, and $\lambda_{silh.}$ are set to 300, 300, 60, 0.06, and 30, respectively. As in previous works, the learning rate is set to $5e{-}5$, and the network is trained with 50 epochs.

Implementation of the baseline using PTN. In ablation studies, we use the baseline model, Baseline-*PTN*, in which the feature fields module has been replaced by a voxel-based representation (*i.e.*, PTN [21]) in ImpHMR to verify the efficacy of using implicit representation. To implement Baseline-*PTN*, we use the *Decoder* network proposed in Yan *et al*. [21] consisting of *Volume Generator* and *Perspective Transformer*. In Baseline-*PTN*, Decoder takes a human-related feature vector $\mathbf{z}_{\text{fg}}$ as input. It generates a feature volume with $4 \times 4 \times 4$ resolution, and from the volume generates a projected feature viewed from a specific viewing direction ϕ by perspective transform. In order to generate an output of the same size as ImpHMR, we set the channel size of the feature volume to 2048. In addition, since there is no volume density in Baseline-*PTN*, the feature value to be projected is obtained through a max operation when performing perspective projection.

1.2. Datasets

The mixture of 3D (*i.e.*, MPI-INF-3DHP, Human3.6M) and 2D datasets (*i.e.*, MPII, COCO, LSPET) is used for training.

MPI-INF-3DHP [16] is a 3D human pose benchmark mostly taken in indoor environments. The dataset utilizes a markerless motion capture system and contains 3D body joint labels. We use the official training set of 8 subjects and 16 videos per subject for training.

Human3.6M [4] is a large-scale indoor 3D human pose dataset containing 3.6 million video frames and corresponding 2D and 3D body joint labels. It consists of 15 action categories and 7 subjects. We use 5 subjects (S1, S5, S6, S7, S8) for training and 2 subjects (S9, S11) for testing. Also, the dataset is subsampled from 50 to 25 frames per second.

3DPW [19] is a 3D human pose dataset obtained with an IMU sensor and RGB camera in an in-the-wild environment. It consists of 60 videos (24 train, 12 validation, 24 test) and 51K frames, annotated with SMPL parameters. We use both for training and evaluation, where noted.

In-the-wild 2D datasets. We use MPII [1], COCO [15], and LSPET [5] datasets for in-the-wild 2D human body keypoint datasets. The MPII, COCO, and LSPET datasets consist of human instances labeled with 2D human body joints in amounts of 14k, 75k, and 7k, respectively. As in PARE [10], we use the pseudo-ground-truth SMPL labels provided by [6] for each dataset for training.

Method	Res.1	Res.2	Res.4	Res.6
METRO [13]	37.6	-	-	-
MeshGraphormer [14]	18.2	-	-	-
HybrIK [12]	23.2	-	-	-
ImpHMR (Ours)	88.8	88.4	87.1	78.2

Table 1. **Comparison of inference speed.** The numbers are in *frames per second* (**fps**). The Res. denotes the spatial resolution of a 2D feature map in volume rendering for our method. Thanks to efficient spatial representation in feature fields, ImpHMR shows about 2 $\sim$ 4 *times faster* fps compared to METRO, MeshGraphormer, and HybrIK.

1.3. Entanglement Measurement

To quantitatively evaluate the 3D spatial construction capability of ImpHMR, we define **E**ntanglement between the **S**hape and the **V**iewing direction (in short, ESV). Let $\boldsymbol{\beta}_\phi = [\beta_{\phi,1}, \cdots, \beta_{\phi,10}] \in \mathbb{R}^{10}$, where $\beta_{\phi,i}$ denotes i-th coefficient in the PCA shape space, be SMPL shape parameters inferred by the model from an arbitrary viewing direction ϕ. Then, the standard deviation σ_i of the i-th shape parameters $\beta_{\phi,i}$ for the changing in viewing direction ϕ can be obtained as $\sigma_i = \sqrt{\frac{1}{360}\sum_{\phi=0^\circ}^{359^\circ}(\beta_{\phi,i} - \mu_i)^2}$, where μ_i denotes the mean value of the i-th shape parameters calculated by $\mu_i = \frac{1}{360}\sum_{\phi=0^\circ}^{359^\circ}\beta_{\phi,i}$. Finally, we can obtain ESV by averaging the standard deviation σ_i of each shape coefficient as $\text{ESV} = \frac{1}{10}\sum_{i=1}^{10}\sigma_i$. The smaller the deviation of the shape parameters inferred for the change in viewing direction, the lower the ESV value is measured. And this indicates that the shape and viewing direction are well disentangled.

1.4. More Comparison of Inference Speed

Table 1 compares the inference speed (*i.e.*, fps) between ImpHMR and the current best-performing methods [12–14]. For fair evaluation, we use the lightest backbone model proposed in each paper (R50 [2], HRNet-W64 [18], and R34 [2] for METRO [13], MeshGraphormer [14], and HybrIK [12], respectively). Also, frames per second (fps) is calculated by averaging the time it took for each model to infer 10000 times of an input image of 224 × 224 size on RTX 2080Ti GPU. As shown in Tab. 1, we can notice that ImpHMR has 2 $\sim$ 4 *times faster* fps than METRO, MeshGraphormer, and HybrIK, which are latest HMR methods.

1.5. Qualitative Results

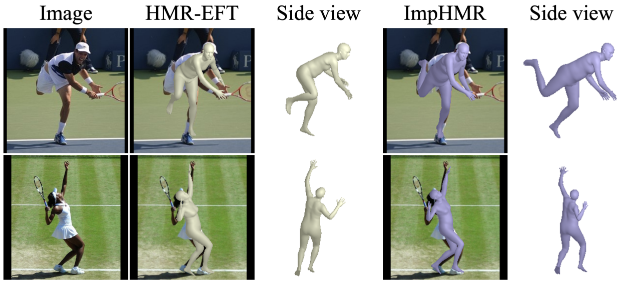

Figure 1. **Qualitative comparison by rotating the mesh inferred from the canonical viewing direction.** For each method, the front and side views of the mesh inferred from the canonical viewing direction are shown in order from left to right.

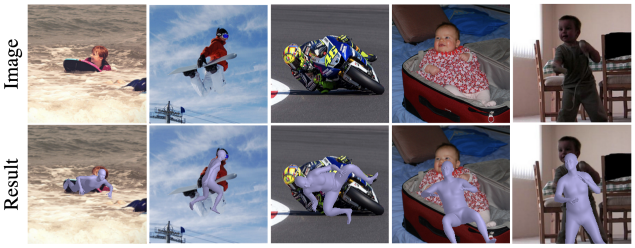

Figure 2. **Failure cases.** Inference results of challenging scenarios (*i.e.*, heavy occlusion, indistinguishable from background, and body shape of children) in which ImpHMR fails to reconstruct pleasing results.

Comparison by rotating the mesh inferred from the canonical viewing direction. The reason for showing the inferred SMPL mesh viewed from different viewing directions in the Fig. 7 (in Sec. 4.2) is to verify that our method has learned well the prior knowledge about human appearance on neural feature fields. In addition, since the MPJPE, PA-MPJPE, and PVE metrics mean reconstruction errors in 3D space, we can expect that our method will show good reconstruction results in other views. For example, as shown in Fig. 1, the mesh inferred from the canonical viewing direction by ImpHMR shows more plausible poses in the side view. Also, the first example in Fig. 1 shows the advantage of ImpHMR in the presence of self-occlusion.

1.6. Failure Cases

Figure 2 shows the inference result for challenging scenarios in which ImpHMR fails to reconstruct pleasing results. As can be seen in Fig. 2, ImpHMR fails to infer when most of the human body is occluded, or when the human body is indistinguishable from the background or object, and for the person who has the body shape of children.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1, 2

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co-

variate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 1

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1

[5] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472, 2011. 1

[6] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 1

[7] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 1

[8] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1

[9] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[10] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 1

[11] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[12] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 2

[13] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[14] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 2

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014. 1

[16] 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 *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 1

[17] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, June 2021. 1

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[19] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[20] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[21] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1