Instance-aware Contrastive Learning for Occluded Human Mesh Reconstruction

Supplementary Material

1. Introduction

In this supplementary material, we provide more explanations of the proposed method. First, details of the network architecture and the training procedure are described in Section 2. Further analysis of the proposed instance-aware contrastive learning scheme is conducted in Section 3. In the following, additional discussions about the results are given in Section 4.

2. Implementation Details

2.1. Architecture Details

For the reproducibility of the proposed method, the detailed architectural setting is specified in Table 1, except for the backbone network. Any backbone network can be adopted to encode the feature which is fed into network branches, while we basically use ResNet-50 [4] in this work. Note that the 2-dimensional coordinate map, which contains the coordinate index (x, y) at every location of the feature map [13], is concatenated with the backbone feature along the channel direction (see C_{in} for the first layer of each branch). The number of input and output channels for all fully connected layers included in nonlinear projectors (i.e., P_C and P_J in Fig. 3 of the main paper) is set to 64, without using the additive bias. The number of parameters in the proposed network is 35.6 million (3.1 million without the backbone network) and our model achieves the processing speed of 46.4 fps on a single NVIDIA GeForce RTX 3090 GPU, which enables the real-time operation.

Branch	Block	F	S	P	R_{in}	$\mathrm{R}_{\mathrm{out}}$	C_{in}	$\mathrm{C}_{\mathrm{out}}$
Center heatmap	ConvBlock	3	2	1	128	64	32+2	64
	ResBlock	3	1	1	64	64	64	64
	ResBlock	3	1	1	64	64	64	64
	1×1 Conv	1	1	0	64	64	64	1
Joint heatmap	ConvBlock	3	2	1	128	64	32+2	64
	ResBlock	3	1	1	64	64	64	64
	ResBlock	3	1	1	64	64	64	64
	1×1 Conv	1	1	0	64	64	64	24
Center-aligned instance map	ConvBlock	3	2	1	128	64	32+2	128
	ResBlock	3	1	1	64	64	128	128
	ConvBlock	3	1	1	64	64	128+1+24	128
	ResBlock	3	1	1	64	64	128	128
	1×1 Conv	1	1	0	64	64	128	64
Joint-aligned instance map	ConvBlock	3	2	1	128	64	32+2	128
	ResBlock	3	1	1	64	64	128	128
	ConvBlock	3	1	1	64	64	128+1+24	128
	ResBlock	3	1	1	64	64	128	128
	1×1 Conv	1	1	0	64	64	128	64
SMPL parameter map	ConvBlock	3	2	1	128	64	32+2	64
	ConvBlock(top)	3	1	1	64	64	64+1+64	64
	ConvBlock(bottom)	3	1	1	64	64	64+24+64	64
	ResBlock	3	1	1	64	64	64+64	128
	ResBlock	3	1	1	64	64	128	128
	1×1 Conv	1	1	0	64	64	128	142
Camera parameter map	ConvBlock	3	2	1	128	64	32+2	64
	ResBlock	3	1	1	64	64	64	64
	ResBlock	3	1	1	64	64	64	64
	1×1 Conv	1	1	0	64	64	64	3

Table 1. The detailed architecture of network branches in the proposed method. F, S, P, R, and C denote the size of filter, stride, padding, resolution, and the number of channels, respectively. Note that each ResBlock contains two convolution layers whose setting is same as shown in this Table.

2.2. Training

Loss calculation according to dataset. During training, 3D human pose datasets (i.e., Human3.6M [5], MPI-INF-3DHP [11], and MuCo-3DHP [12]) and 2D human pose datasets (i.e., MPII [1], LSP [6], COCO [10], and CrowdPose [9]) are utilized to optimize parameters of the proposed network. Since only the Human3.6M dataset provides the ground truth for pose and shape parameters of the SMPL model [2], we use the ground truth of 2D and 3D keypoints for MPI-INF-3DHP and MuCo-3DHP datasets instead. On the other hand, the pseudo ground truth for pose and shape parameters of the SMPL model, which is generated by using [8], is adopted for MPII, LSP, and COCO datasets. The annotation of 3D keypoints also can be acquired by linearly regressing the pseudo mesh vertices into joints with the pre-defined matrix. For the CrowdPose dataset, only 2D keypoints labels are used for training. The loss terms used when learning each dataset are shown in Table 2.

Datasets	$\mathcal{L}_{pose}$	$\mathcal{L}_{shape}$	$\mathcal{L}_{3d}$	$\mathcal{L}_{pa3d}$	$\mathcal{L}_{2d}$
Human3.6M	✓	✓	✓	✓	✓
MPI-INF-3DHP			✓	✓	✓
MuCo-3DHP			✓	✓	✓
MPII	✓	✓	✓	✓	✓
LSP	✓	✓	✓	✓	✓
COCO	✓	✓	✓	✓	✓
CrowdPose					✓

Table 2. Activated loss terms for each training dataset. Note that other loss terms (i.e., $\mathcal{L}_{cont}$, $\mathcal{L}_{center}$, $\mathcal{L}_{joint}$, and $\mathcal{L}_{prior}$) are always activated for all the datasets.

Training of heatmaps. In the proposed method, we leverage predicted heatmaps corresponding to the body center [13] and 24 joints [2] in the process of both contrastive learning and mesh regression. For training, we generate the ground truth for the heatmap by applying the Gaus-

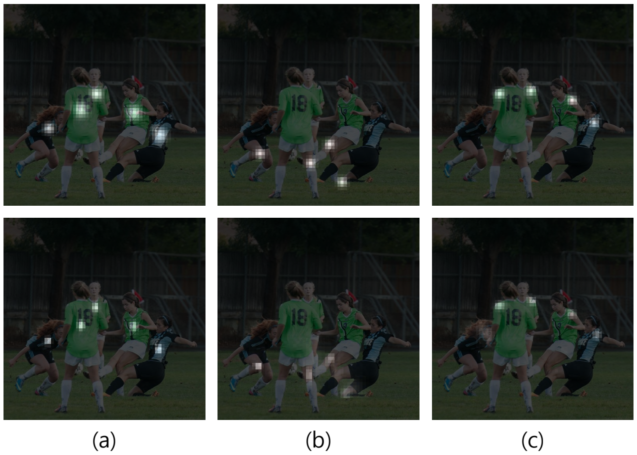

Figure 1. Visualization examples for the ground truth (top) and the prediction result (bottom) of the heatmap. (a) Center heatmap. (b) Joint heatmap corresponding to the right knee. (c) Joint heatmap corresponding to the left shoulder.

sian kernel, where the peak point is set to the value of 1, for each keypoint. By using the body center heatmap loss $\mathcal{L}_{center}$ and the joint heatmap loss $\mathcal{L}_{joint}$, the heatmap can be learned as shown in Fig 1. As can be seen, the heatmap is accurately activated at every keypoint location of multiple persons even under severe person-to-person occlusions.

3. Instance-aware Contrastive Learning

3.1. Convergence of Contrastive Loss

To embed the identity information in the latent space, parameters are optimized for center and joint features via the proposed contrastive loss $\mathcal{L}_{cont}$ (see Eq. (2) of the main paper). The convergence trend of the proposed contrastive loss is shown in Fig. 2. As can be seen, the loss value is stably converged without severe oscillations during training.

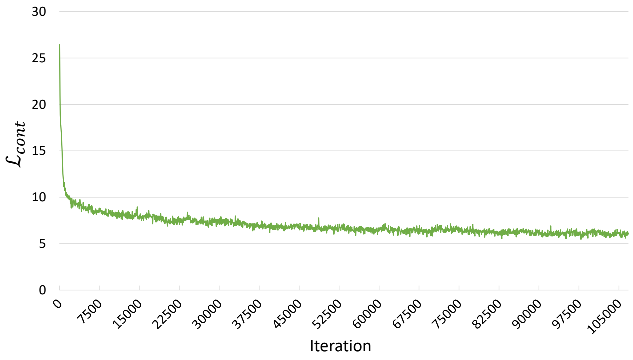

Figure 2. Convergence trend of the proposed contrastive loss during 10.5K iterations of training.

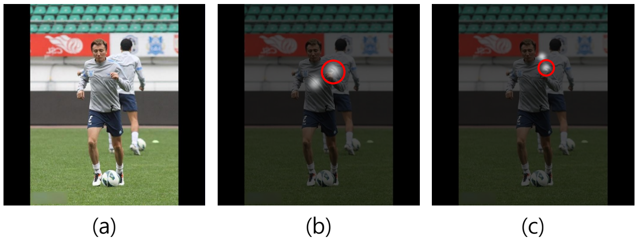

Figure 3. An example of overlapped case between body center and joint positions of different persons. (a) Input image. (b) Ground truth of the center heatmap. (c) Ground truth of the joint heatmap corresponding to the left shoulder.

3.2. Effect of Nonlinear Projector

In person-to-person occlusion scenes, locations of the body center for one person and the joint for another person are often overlapped. The corresponding example is shown in Fig. 3. As can be seen, the body center of the person in the back (see Fig. 3(b)) and the left shoulder of the person in the front (see Fig. 3(c)) are located at the same position. Since this problem may confuse the network in representing the identity information if a single instance map is used for instance-aware contrastive learning, we instead encode the center-aligned instance map and the joint-aligned instance map, respectively. To compare center and joint features that are sampled from two different instance maps, we transform them into the same latent space through each nonlinear projector, which is designed similarly to [3], while maintaining the essential information related to the personal identity. The importance of the nonlinear projector is demonstrated in Table 3. Based on the performance comparison, it is thought that the nonlinear projector plays a significant role in learning the identity representation.

Methods	MPJPE(↓)	PA-MPJPE(↓)
Ours (w/o projector)	105.1	80.0
Ours (w/ projector)	102.0	77.2

Table 3. Performance analysis according to the use of the nonlinear projector based on the 3DPW-PC dataset.

4. Discussion on Results

4.1. Qualitative Results

Additional examples of the occluded human mesh reconstruction by the proposed method are shown in Fig 4. Specifically, we provide results on several datasets, i.e., CMU-Panoptic [7], 3DPW [14], OCHuman [15], and CrowdPose [9]. As can be seen, human meshes are successfully reconstructed under diverse person-to-person occlusion situations. In particular, our model shows reliable

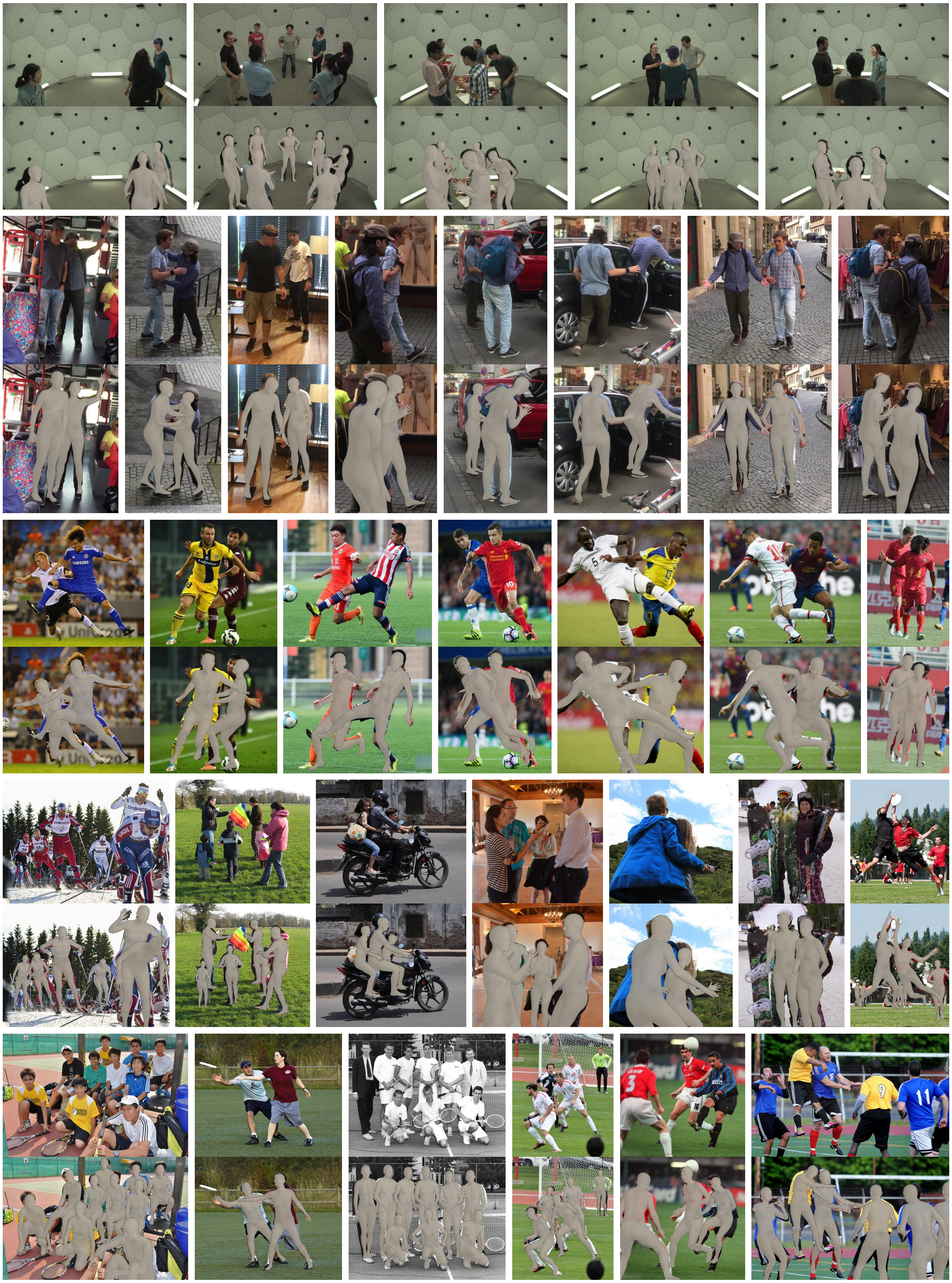

Figure 4. More results of occluded human mesh reconstruction by the proposed method on CMU-Panoptic (1st row), 3DPW (2nd row), OCHuman (3rd row), and CrowdPose (4-5th rows) datasets.

performance not only in constrained conditions such as laboratory environments (see the first row of Fig. 4), but also in outdoor circumstances (see the second row of Fig. 4). Moreover, as can be seen in the third to fifth rows in Fig. 4, results on sports scenes and real-life images, which contain complicated inter-person interactions with various dynamic poses, demonstrate the robustness of the proposed method against person-to-person occlusions.

4.2. Limitations

Some failure cases of the proposed method are shown in Fig. 5. As can be seen, the proposed method often suffers from ambiguities driven by self-occlusions in crowded contexts. Specifically, when certain body parts of target persons are occluded by themselves, the network tends to inappropriately exploit the reconstruction cue from other visible body parts, which have similar appearances to occluded parts, belonging to either the target person (see Fig. 5(a)) or non-target persons (see Fig. 5(b)). To overcome this limitation, efforts to distinguish invisible parts by self-occlusions could be made in our future works.

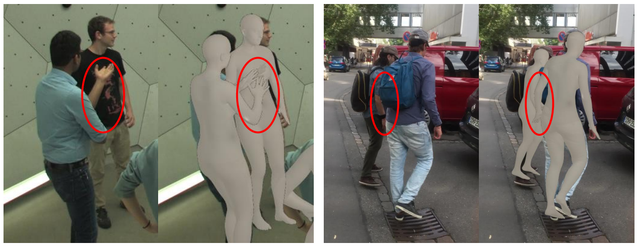

Figure 5. Examples of failure cases by the proposed method.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3686–3693, 2014. 1

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. Eur. Conf. Comput. Vis.*, pages 561–578, 2016. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn.*, pages 1597–1607, 2020. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 1

[5] 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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 1

[6] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. Brit. Mach. Vis. Conf.*, pages 1–11, 2010. 1

[7] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proc. Int. Conf. Comput. Vis.*, pages 3334–3342, 2015. 2

[8] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *Proc. Int. Conf. 3D Vis.*, pages 42–52, 2021. 1

[9] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10863–10872, 2019. 1, 2

[10] 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 *Proc. Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 1

[11] 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 *Proc. Int. Conf. 3D Vis.*, pages 506–516, 2017. 1

[12] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *Proc. Int. Conf. 3D Vis.*, pages 120–130, 2018. 1

[13] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Proc. Int. Conf. Comput. Vis.*, pages 11179–11188, 2021. 1

[14] 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 *Proc. Eur. Conf. Comput. Vis.*, pages 601–617, 2018. 2

[15] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2Seg: Detection free human instance segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 889–898, 2019. 2