PARE: Part Attention Regressor for 3D Human Body Estimation Supplementary Material

Muhammed Kocabas[1,2] Chun-Hao P. Huang[1] Otmar Hilliges[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zurich
{mkocabas,paul.huang,black}@tue.mpg.de otmar.hilliges@inf.ethz.ch

The Supplementary Material consists of this document and a video. They include acknowledgement, disclosure, additional information and visualizations of our method and results.

Acknowledgements: We thank Joachim Tesch for helping with Blender rendering. We thank Nikos Athanasiou, Vassilis Choutas, Emre Aksan, Stefan Stevsics, Xu Chen, Cornelia Kohler, Marilyn Keller, Shashank Tripathi, Yinghao Huang, Omid Taheri, Sai Kumar Dwivedi, Hongwei Yi, Dimitris Tzionas, Timo Bolkart, Yao Feng, and all Perceiving Systems department members for their feedback and fruitful discussions. This research was partially supported by the Max Planck ETH Center for Learning Systems.

Disclosure: https://files.is.tue.mpg.de/black/CoI/ICCV2021.txt

1. Methods

Implementation Details. In all our experiments, we use the weights pretrained on MPII [1] for a 2D pose estimation task to initialize both ResNet-50 and HRNet-W32, because we observe slower convergence with ImageNet pretrained weights. For Table 3-5 in the main paper, we train PARE and our baselines on COCO for 175K steps and evaluate on 3DPW and 3DPW-OCC datasets. We then include all the training data for the SOTA experiment in Table 1 of the main paper. For Table 2, we use the training data of [17] to align the experiment settings.

Loss. We use different weight coefficients λ for each term in the loss function. They are $\lambda_{3D} = 300$, $\lambda_{2D} = 300$, $\lambda_{SMPL} = 60$, $\lambda_{P} = 60$.

Body Part Segmentation labels. Since we have SMPL annotations for most of the samples in our datasets, we do not need additional body part segmentation annotations. We directly use the SMPL annotations to obtain supervision. In Fig 1, we visualize this body part labels. For each joint in the SMPL kinematic tree, we have a corresponding body part label.

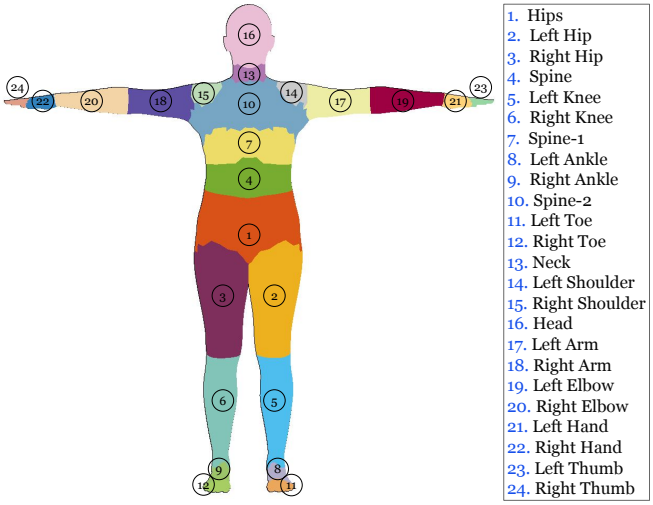

Figure 1: Body part segmentation labels used for the 2D part branch. For each joint in the SMPL kinematic tree, we have a body part label. Correspondences between joints *(right)* and body part labels *(left)* are shown in this figure.

Occlusion augmentation. In Fig. 2, we demonstrate the results of synthetic occlusion and random crop augmentations on two sample images.

Runtime PARE is only 1 ms/image slower than HMR, with runtime of 14.8ms on a GTX2080Ti.

2. Experiments

2.1. Training Datasets

Our training datasets closely follow previous work, namely EFT [5], SPIN [8], and HMR [6]. Here we provide the details for completeness.

MPI-INF-3DHP [12] is a multi-view indoor 3D human pose estimation dataset. 3D annotations are captured via a commercial markerless mocap software, therefore it is less accurate than some of the 3D datasets e.g. Human3.6M [3].

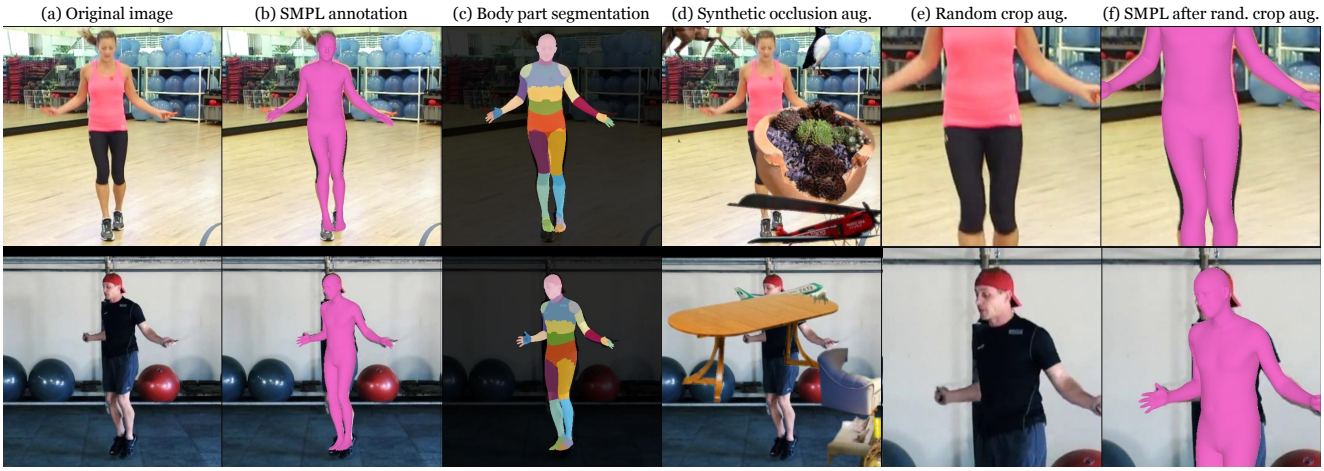

Figure 2: Training samples after synthetic occlusion and random crop augmentations are applied.

We use all of the training subjects S1 to S8 which makes 90K images in total.

Human3.6M [3] is an indoor, multi-view 3D human pose estimation dataset. Following previous methods, for training, we use 5 subjects (S1, S5, S6, S7, S8) which means 292K images.

In-the-wild 2D datasets COCO [9], MPII [1] and LSPET [4] are in-the-wild 2D keypoint datasets. MPII has 14K, COCO has 75K, LSPET has 7K instances labeled with 2D keypoints. In addition to 2D keypoint annotations, we utilize the pseudo SMPL annotations provided by the EFT [5] method.

Training Dataset Ratios. To obtain the final best performing model, we follow EFT [5] and SPIN [8] which use fixed data sampling ratios for each batch. After training 100% with COCO-EFT for 175K steps, we incorporate 50% Human3.6M, 30% In-the-wild (i.e. [COCO, MPII, LSPET]-EFT), and 20% MPI-INF-3DHP datasets into training. We also observe that using [50% Human3.6M, 30% COCO-EFT, 20% MPI-INF-3DHP] or [20% Human3.6M, 30% COCO-EFT, 50% MPI-INF-3DHP] gives equivalent performance on the 3DPW dataset.

Failure Cases. In Fig. 3, we show a few examples where PARE fails to reconstruct reasonable human body poses. The scenarios range from (a-b) too many people in the crop, (b-d) rarely-seen extreme poses, (e-f) children whose body shapes cannot be fully explained by the SMPL model, and (g-h) extreme occlusion.

Comparing to [17]. Zhang et al. [17] parameterize human meshes as UV maps where each pixel stores the 3D location of a vertex. They leverage saliency masks as visibility information and cast occlusions as an image-inpainting problem. However, we find that the raw predicted vertex locations from [17] yield arbitrary global scale, rotation, translation, and there are no camera parameters associated with the output. What they show in the main paper are the post-processed results after fitting a SMPL model, which is not described in their released implementation; how to visualize the unprocessed, raw predicted meshes is unclear. Thus, we bring them to the same camera coordinate frame as PARE through Procrustes Alignment and overlay them on the input as shown in Fig. 5(c). One clearly sees mesh artifacts (red ovals), which is common for non-parametric models. The requirement of accurate saliency maps certainly limits the performance of [17] on in-the-wild images.

		3DPW	
	Method	MPJPE ↓	PA-MPJPE ↓
Temporal	HMMR [6]	116.5	72.6
	Doersch et al. [2]	-	74.7
	Sun et al. [14]	-	69.5
	VIBE [7]	93.5	56.5
	MEVA [10]	86.9	54.7
	PARE (ResNet-50)	82.9	52.3
	PARE (HRNet-W32)	**82.0**	**50.9**

Table 1: **Evaluation on the 3DPW dataset.** The numbers are average joint errors in mm. PARE models outperform video-based methods which leverage temporal information.

Comparing to state-of-the-art Temporal Models. In Table 1, we compare PARE to recent state-of-the-art video based models. To do so, we run a SOTA multi-object tracker and then run PARE independently on each frame of the tracklets, with no temporal smoothing. Even the SOTA video methods have access to extra temporal information, PARE outperforms them. We show some qualitative results

of VIBE and PARE on some challenging images in Fig 6. Please see the supplemental video for a better visualization of the video results (starts at 05:21).

	SPIN [8]	HMR-EFT	PARE
PCK ↑	81.5	83.4	**85.1**

Table 2: **Evaluation of 2D keypoint project accuracy on 3DPW dataset.**

2D keypoint projection accuracy We evaluate the 2D keypoint accuracy of our method by projecting the 3D keypoints to the image space using the estimated camera parameters on 3DPW test set. Percentage of correct keypoints (PCK) is used as the evaluation metric. The results are reported in Table 2.

3. More on visualizing attention of networks

Two new visualizations are proposed in this work: (1) an occlusion sensitivity map/mesh and (2) a part attention map. We provide more examples and discussions for both visualizations. Please see the video for an animation of the sensitivity analysis, which more clearly illustrates the approach.

Occlusion Sensitivity. There are many visualization techniques [11, 13, 16, 18] available to inspect what CNNs learn. We are, however, more interested in studying how perturbations in the input image affect the output rather than visualizing the internal filters learned by CNNs. We therefore follow the framework of [16] and replace the classification score with an error measure for body poses, as described in the main paper. We choose MPJPE as the error measure *without* Procrustes Alignment, because PA-MPJPE leads to artificially low error by aligning global orientations, which are a major source of error.

This analysis is not limited to a particular network architecture so we also apply it on PARE and visualize the error maps together with those from SPIN [8] in Fig. 7. Warmer colors correspond to higher MPJPEs w.r.t. ground truth when those pixels are occluded, suggesting that methods rely on the regions to estimate body poses. One clearly sees that PARE is more robust to localized part occlusion. Please see the video for animation (starts at 00:53).

Additionally, we also map the per-pixel error to the overlaying 3D vertex, and aggregate over the whole 3DPW dataset [15]. In this way, we visualize the per-joint error on the SMPL template mesh, which we term the *occlusion sensitivity mesh*. Fig. 8 shows the occlusion sensitivity mesh for four different joints and averaged over all joints from both SPIN and PARE. We again observe that SPIN is very sensitive to localized part occlusion. For example, occlusions of right arm or face regions result in high error for right wrist. On the other hand, occlusion sensitivity meshes of PARE have more consistent cold colors over the body, again confirming that it is more robust to partial occlusion.

Part Attention. We also visualize the estimated part attention P before softmax in Fig. 9 for four sample images from 3DPW [15]. When body parts are visible, the shapes of warm regions resemble part segmentation labels, which means the network focuses on body part regions (e.g. Left/Right Knee and Ankle in the third row). For naturally occluded body parts, the attended regions get wider, covering other parts and the scene. This suggests that PARE implicitly learns to reason about the visibility of body parts and leverages available information to solve the task. In particular, Fig. 4 illustrates the progression of attention maps during training for two occluded parts Left/Right Ankles. We see that deactivating the part supervision helps attention maps to focus on more meaningful and explainable regions.

In addition to part attention maps, we also visualize the results as segmentation maps in Fig. 10. We visualize the results of two different models; (a) a model trained with full part segmentation supervision, (b) a model trained with part segmentation initially and unsupervised for the final stages. Note that part segmentation IoU decreases significantly when we do not use part segmentation, however we see an increase in body reconstruction accuracy especially in the case of occlusion.

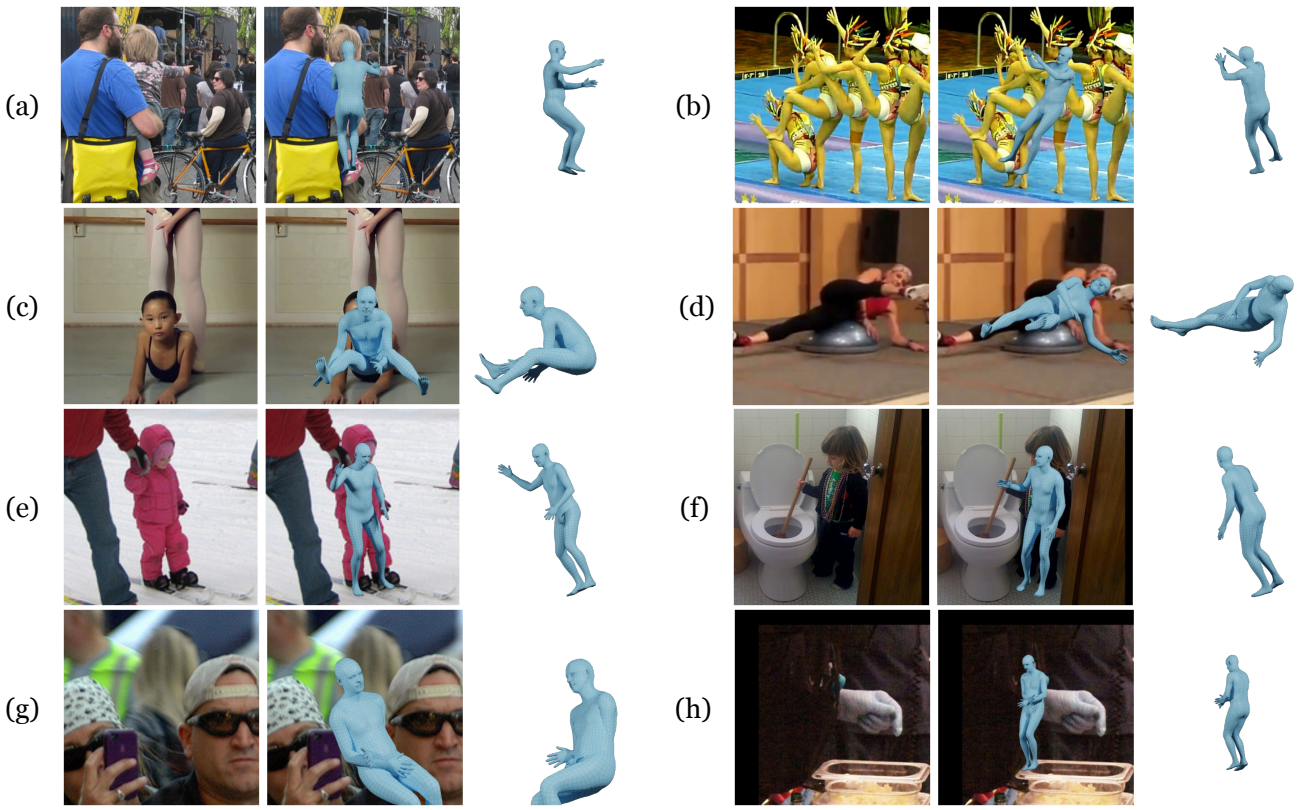

Figure 3: Challenging scenarios where PARE fails to produce fairly good reconstructions.

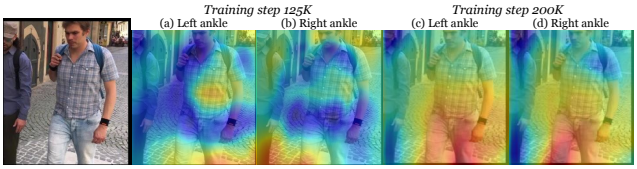

Figure 4: Attention map progression during training. Training uses body-part supervision only until step 125K (a-b). Note that the final attention maps for occluded parts (at 200K (c-d)) focus on visible parents.

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*, 2014. 1, 2

[2] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D pose estimation: motion to the rescue. In *Advances in Neural Information Processing*, 2019. 2

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2014. 1, 2

[4] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[5] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv:2004.03686*, 2020. 1, 2, 5

[6] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 2

[7] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5252–5262, 2020. 2, 6

[8] 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 *International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 2, 3, 5, 7

[9] 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 *European Conference on Computer Vision*, 2014. 2

[10] Zhengyi Luo, S. Golestaneh, and Kris M. Kitani. 3D human motion estimation via motion compression and refinement.

Figure 5: **Qualitative comparison.** Here, we compare PARE with recent state-of-the-art methods i.e. SPIN [8], HMR-EFT [5], and Zhang et al. [17].

Figure 6: Comparison of VIBE [7] with our method, PARE. Note that VIBE is a video-based method, while PARE is run on each video frame independently.

In *Asian Conference on Computer Vision*, pages 324–340, 2020. 2

[11] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015. 3

[12] 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 *International Conference on 3DVision*, 2017. 1

[13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017. 3

[14] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao

Figure 7: Occlusion Sensitivity Maps of SPIN [8] and PARE

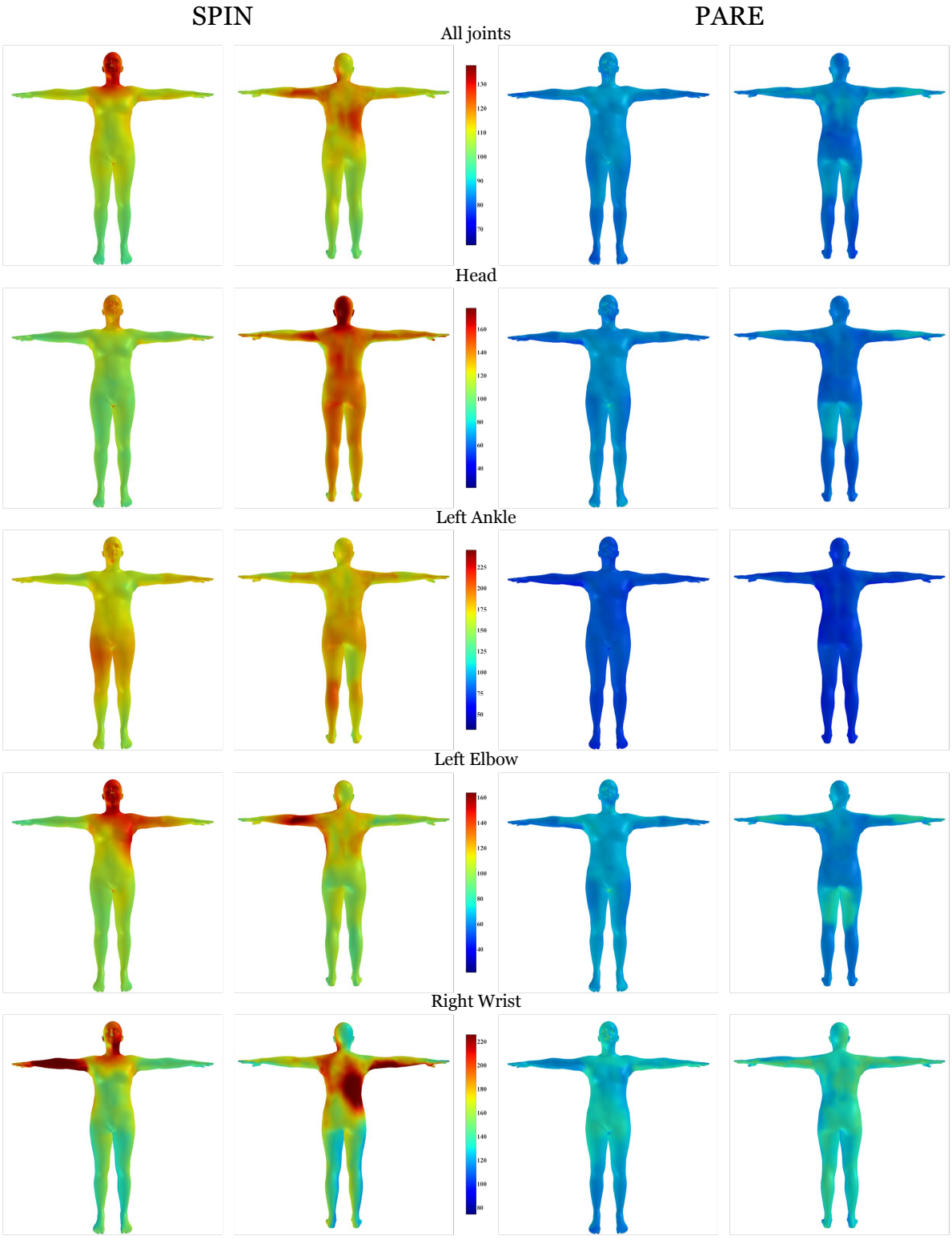

Figure 8: Occlusion sensitivity meshes per joint.

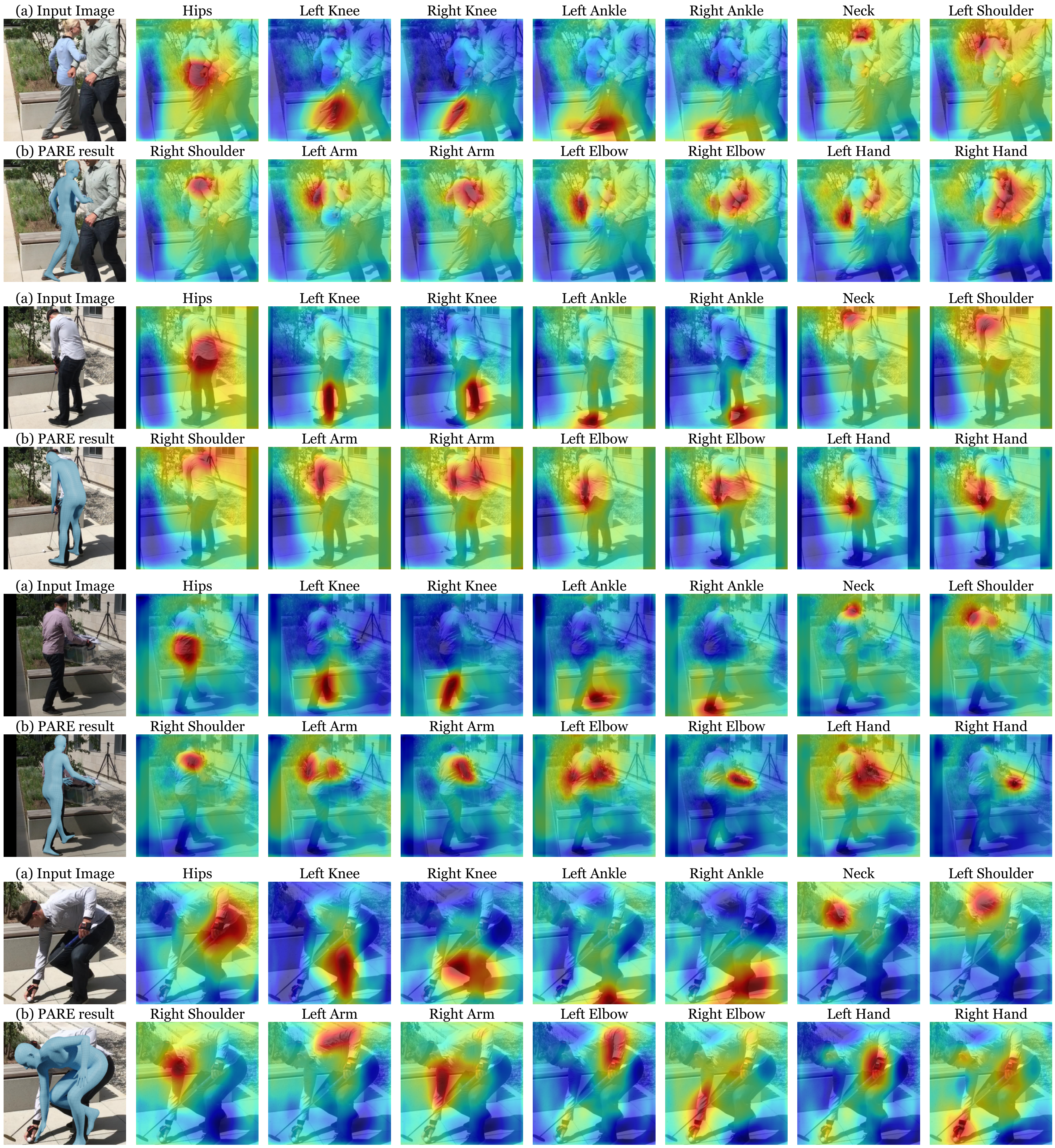

Figure 9: Part attention maps.

(a) Part segmentation results with full part supervision

(b) Part segmentation results with part supervision + unsupervised

Figure 10: Part segmentation results in two different scenarios: (a) full part segmentation supervision is applied during training, (b) part segmentation supervision is applied at the initial stages and training is continued without part supervision. At the top of each result, we denote the part segmentation IoU, MPJPE and PA-MPJPE. Notice how part segmentation IoU decreases, but per-joint accuracy improves.

Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *International Conference on Computer Vision*, 2019. 2

[15] 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 *European Conference on Computer Vision*, pages 614–631, 2018. 3

[16] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference*

on Computer Vision, 2014. 3

[17] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7374–7383, 2020. 1, 2, 5

[18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 3