Supplementary Material for ”3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation”

A. Volume Rendering Equations

We provide the mathematical derivation of the analytic form of the volume rendering equation in Equation 4 of the main article. Here we leave out the camera parameters Π by modeling in the camera coordinate. In the camera coordinate, a ray passing through pixel (i, j) can be written as,

$$\boldsymbol{r}(t) = t\boldsymbol{D} = t\left[\frac{i - O_y}{f}, \frac{j - O_x}{f}, 1\right]^T, \tag{1}$$

where (O_x, O_y) is the principal point of the camera and f is the focal length. Then the volume density along the ray at the k-th Gaussian kernel is,

$$\rho_k(\boldsymbol{r}(t)) = \exp\left(q_k - \frac{(t - l_k)^2}{2\sigma_k^2}\right) \tag{2}$$

where

$$\begin{aligned} l_k &= \frac{\boldsymbol{M_k}^T \boldsymbol{\Sigma_k}^{-1} \boldsymbol{D} + \boldsymbol{D}^T \boldsymbol{\Sigma_k}^{-1} \boldsymbol{M_k}}{2\boldsymbol{D}^T \boldsymbol{\Sigma_k}^{-1} \boldsymbol{D}} \\ q_k &= -\frac{1}{2}(\boldsymbol{M_k} - l_k \boldsymbol{D})^T \boldsymbol{\Sigma_k}^{-1} (\boldsymbol{M_k} - l_k \boldsymbol{D}) \\ \sigma_k^2 &= \frac{1}{\boldsymbol{D}^T \boldsymbol{\Sigma_k}^{-1} \boldsymbol{D}}. \end{aligned}$$

Intuitively, l_k is the global maximizer of $\rho_k(\boldsymbol{r}(t))$ that gives the peak density of the k-th Gaussian kernel which is $\exp(q_k)$. This enables us to calculate the integral for $T(t)$ in Equation 4,

$$\begin{aligned} T(t) &= \exp\left(-\int_{-\infty}^{\infty} \rho(\boldsymbol{r}(s))\mathrm{d}t\right) & (3) \\ &= \exp\left(-\sum_{k=1}^{K} \frac{e^{q_k}}{2}(\mathrm{erf}((t - l_k)/\sigma_k) + 1)\right), & (4) \end{aligned}$$

where $\mathrm{erf}(x)$ is the Error Function. Then we can analytically calculate the integral of volume rendering as,

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{r}) &= \int_{-\infty}^{\infty} T(t) \sum_{k=1}^{K} \rho_k(\boldsymbol{r}(t))\boldsymbol{\phi_k}\mathrm{d}t & (5) \\ &= \sum_{k=1}^{K} T(l_k)e^{q_k}\boldsymbol{\phi_k}. & (6) \end{aligned}$$

Here we assume the Gaussian kernels are far from the camera relative to its scale so that $t_n \approx -\infty$ and $t_f \approx \infty$.

B. Experiments

B.1. Evaluation on Human3.6M

We further evaluate our method on `Human3.6M` [1] dataset for reference. Note that this dataset is not our main focus because 1) it does not have any occlusion; 2) it is indoor setting and the train/test data are quite similar so the performance is highly saturated (e.g., for input image of size 224×224, assuming a 170cm tall human occupies the whole image, a 1px shift in the image space would correspond to 7.6mm already). Here we report the performance of other SOTA methods with the same ResNet50 backbone for a fair comparison. As shown in Table 1, we achieve SOTA performance.

B.2. Comparison to Multi-modal Methods

We further compare 3DNBF with a SOTA multi-modal method [7], which models the conditional distribution of 3D human pose given the test image, on `3DPW-AdvOcc@80`. Note that we only evaluate visible keypoints which excludes a certain amount of ambiguities. We run the official implementation and the mode prediction achieves MPJPE: 215.7 (74.9 ↑), PA-MPJPE: 97.1 (25.3 ↑), and PCKh: 71.7 (13.4 ↓). The 5-sample best scores are MPJPE: 146.5 (5.8 ↑), PA-MPJPE: 80.7 (9.0 ↑), and PCKh: 75.8 (9.2 ↓). This shows that our model outperforms also the multi-modal baseline. Although modeling multi-modal distributions is promising for handling severe occlusion, we consider it orthogonal to our approach.

B.3. Visualization of Image Features

We visualize the image features and pose predictions under varying occlusion levels in Fig. 1. The predicted correspondences between pixels and Gaussian kernels from the UNet features are color-coded. We can observe that the individual image features are quite robust to occlusion, only starting to get distorted when the occluder is about half the size of the human. However, as the features in the non-occluded regions are still predicted well, our method is able

Method	Human3.6M [1]	
	MPJPE↓	PA-MPJPE↓
HMR [3]	88.0	56.8
SPIN [5]	62.3	41.7
HMR-EFT [2]	-	46.0
PARE [4]	82.7	53.7
3DNBF	**58.7**	**38.9**

Table 1. **Evaluation on `Human3.6M` protocol 2.**

to correctly predict the non-occluded joints due to the robust likelihood (Eq.3)

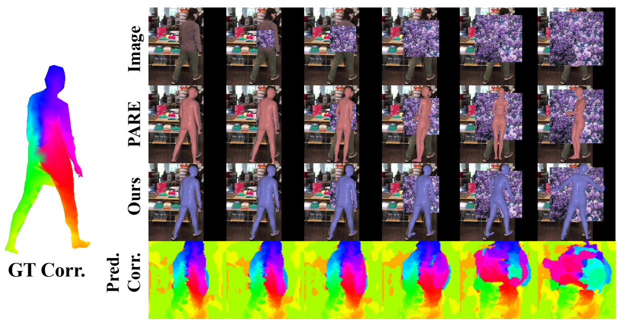


Figure 1. Visualizing image feature and pose estimation under varying occlusion levels. Left: GT image-3D correspondence. Right top to bottom: input occluded image, PARE output, our 3DNBF output, and predicted image-3D correspondence.

B.4. More Analysis of NBV compare to Mesh representation.

In our ablation study, the mesh baseline is implemented with the 3D-aware features, the contrastive training, and the robust likelihood. Compared to the mesh representation with SoftRas [6], the volume representation is analytically differentiable and hence can provide better gradients, particularly in the case of self-occlusion, due to the better volume density blending compared to the distance-based blending used in SoftRas. Fig. 2 illustrates the optimization process under self-occlusion with a mesh representation and our NBV. We set σ=10^{-3}, γ=10^{-2}, and K=40 for SoftRas which is consistent with the parameters used in the ablation. Note how initially the right arm is estimated to be behind the body, but from iteration 40 can be corrected to be in front of the body.

B.5. Qualitative Results

In Fig. 4, 5 and 6, we provide more qualitative results for 3DNBF comparing with state-of-the-art 3D human pose estimation methods on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. Qualitative results on 3DOH50K are provided in Fig. 7. The comparison between our 3DNBF and optimization-based methods: EFT [2] and 3D POF [8] are visualized in Fig. 8 and 9 respectively.

Figure 2. NBV better handles self-occlusion than mesh representation+SoftRas.

Figure 3. Failure cases where our model confuses front and back.

B.6. Limitations and Failure Cases

While being robust to occlusion, our method do have some limitations, which we leave for future work. The limitations of our method are: 1) the inference speed is not real-time; 2) more detailed body models may explain the observed features better and improve accuracy; 3) it should be extended to multi-person scenarios. One of the main failure cases we observe is the occasional front-back switch errors when the head is occluded as shown in Figure 3.

References

[1] 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, 2013. 1, 2

[2] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 2, 4, 6, 7, 8

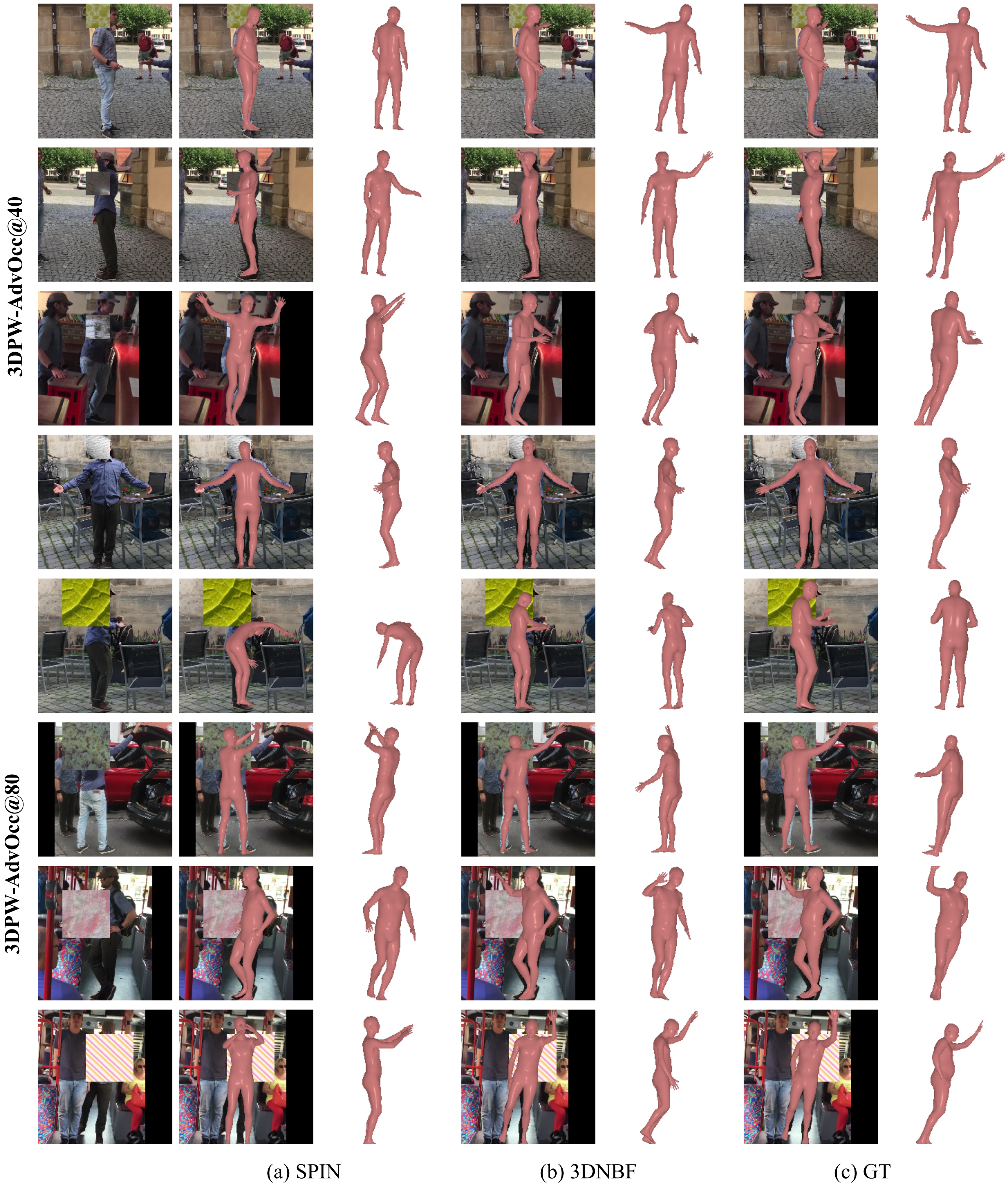


Figure 4. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) initial pose predicted by SPIN [5], (b) 3DNBF prediction and (c) ground truth pose.

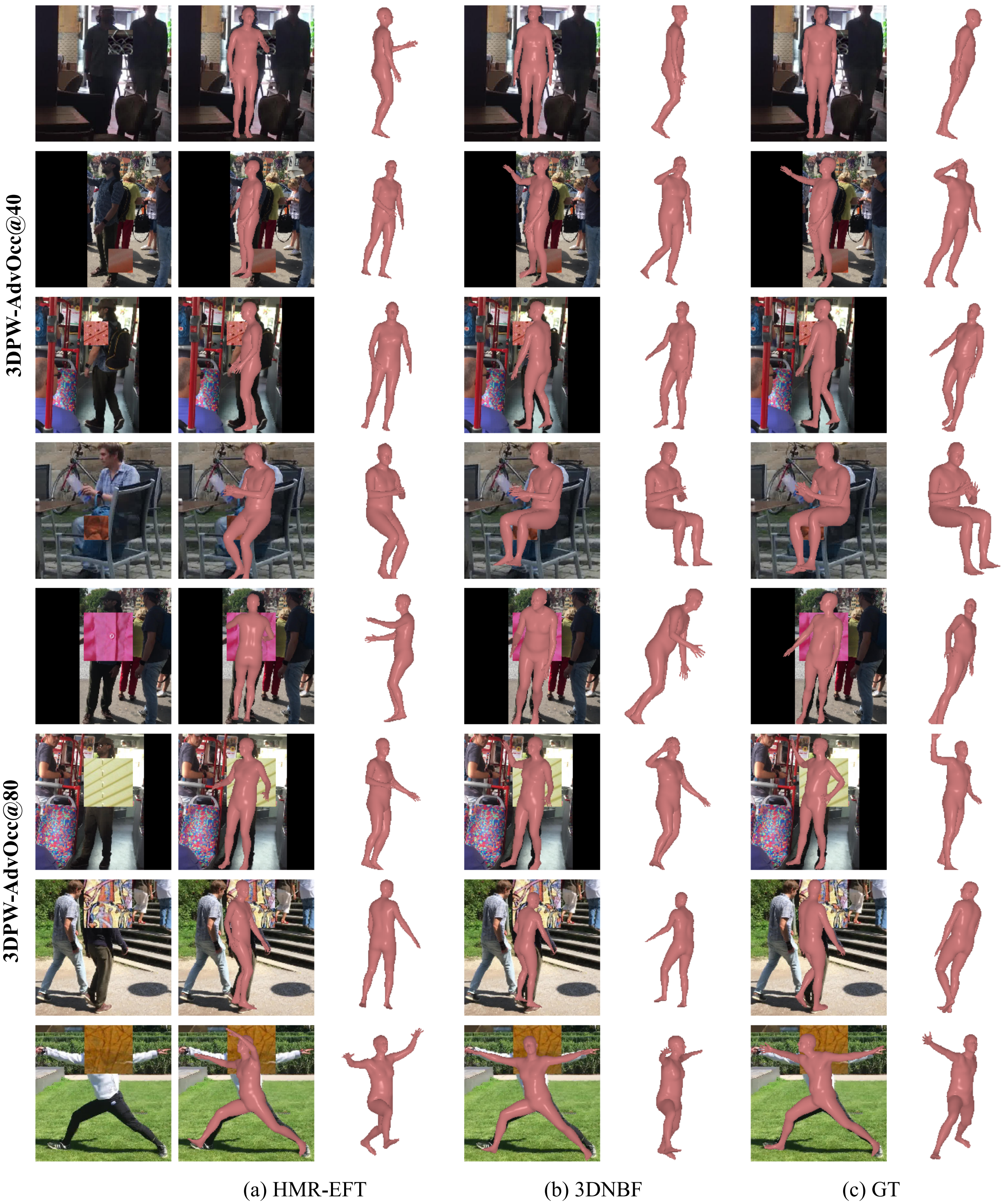


Figure 5. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) initial pose predicted by HMR-EFT [2], (b) 3DNBF prediction and (c) ground truth pose.

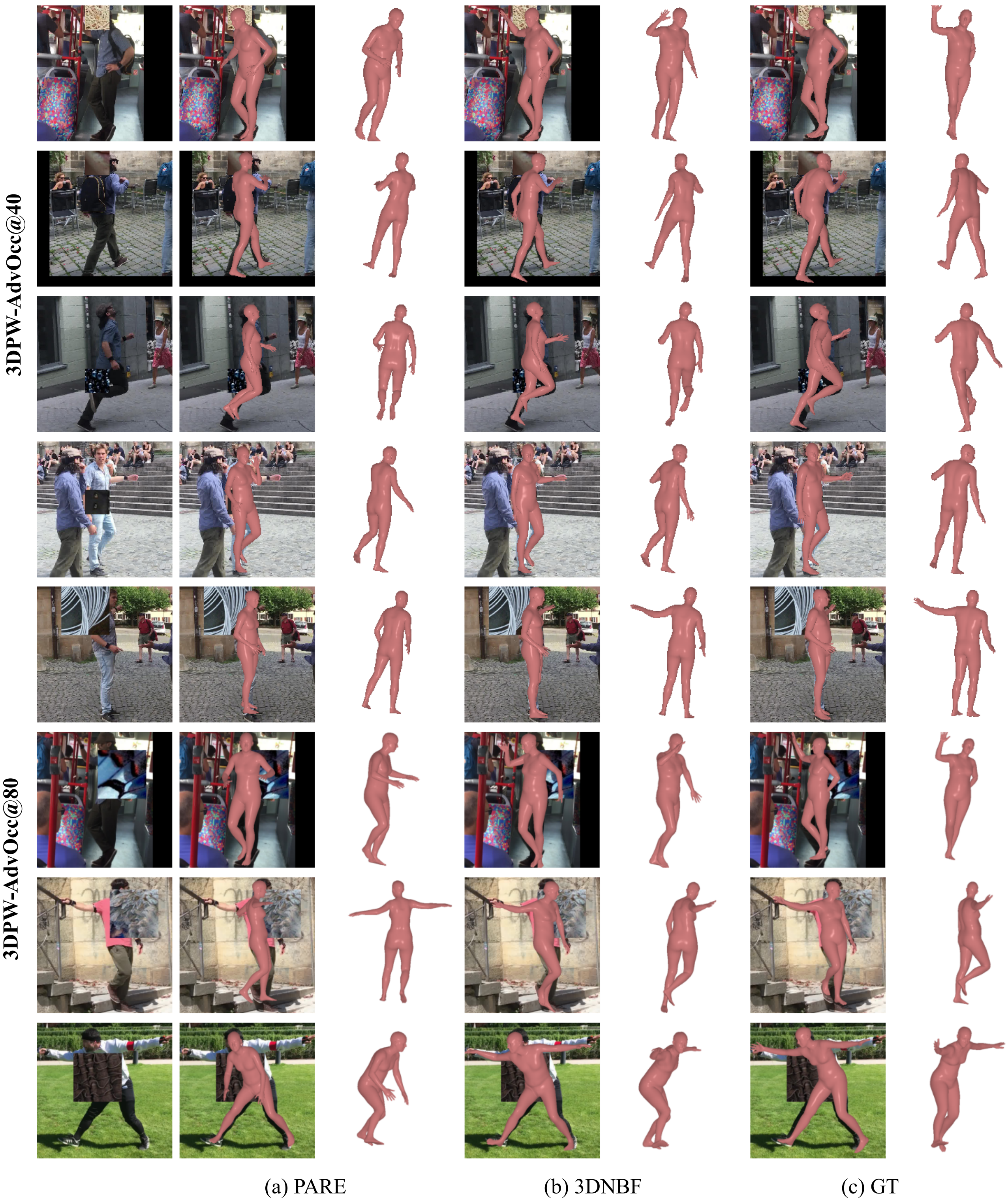


Figure 6. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) initial pose predicted by PARE [4], (b) 3DNBF prediction and (c) ground truth pose.

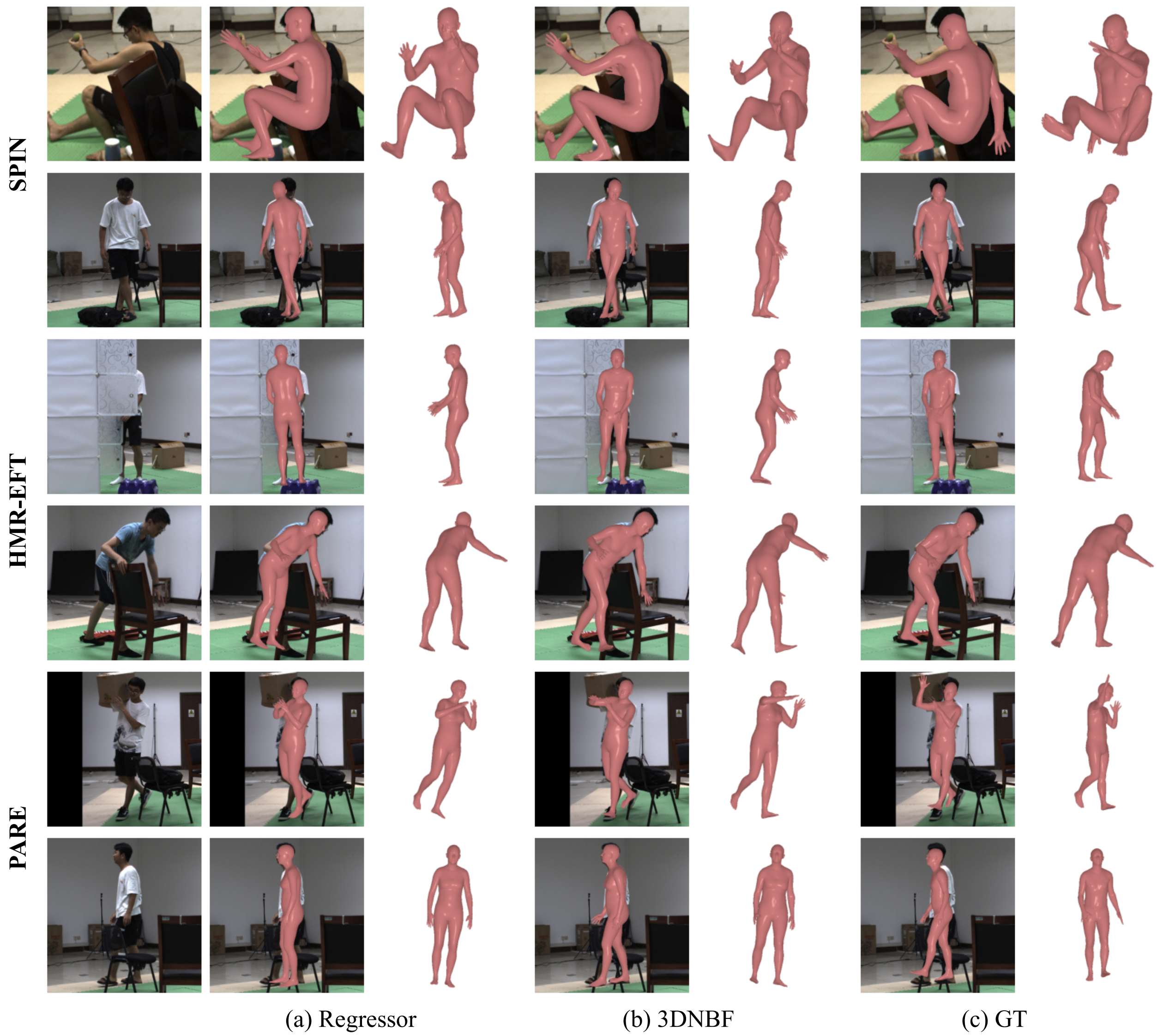


Figure 7. Qualitative results on 3DOH50K [9]. For left to right are the input image, (a) initial pose predicted by regression-based methods, (b) 3DNBF prediction and (c) ground truth pose. Row 1-2, 3-4, and 5-6 are results for SPIN [5], HMR-EFT [2], and PARE [4] respectively.

[3] 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. 2

[4] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11127–11137, 2021. 2, 5, 6

[5] 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. 2, 3, 6

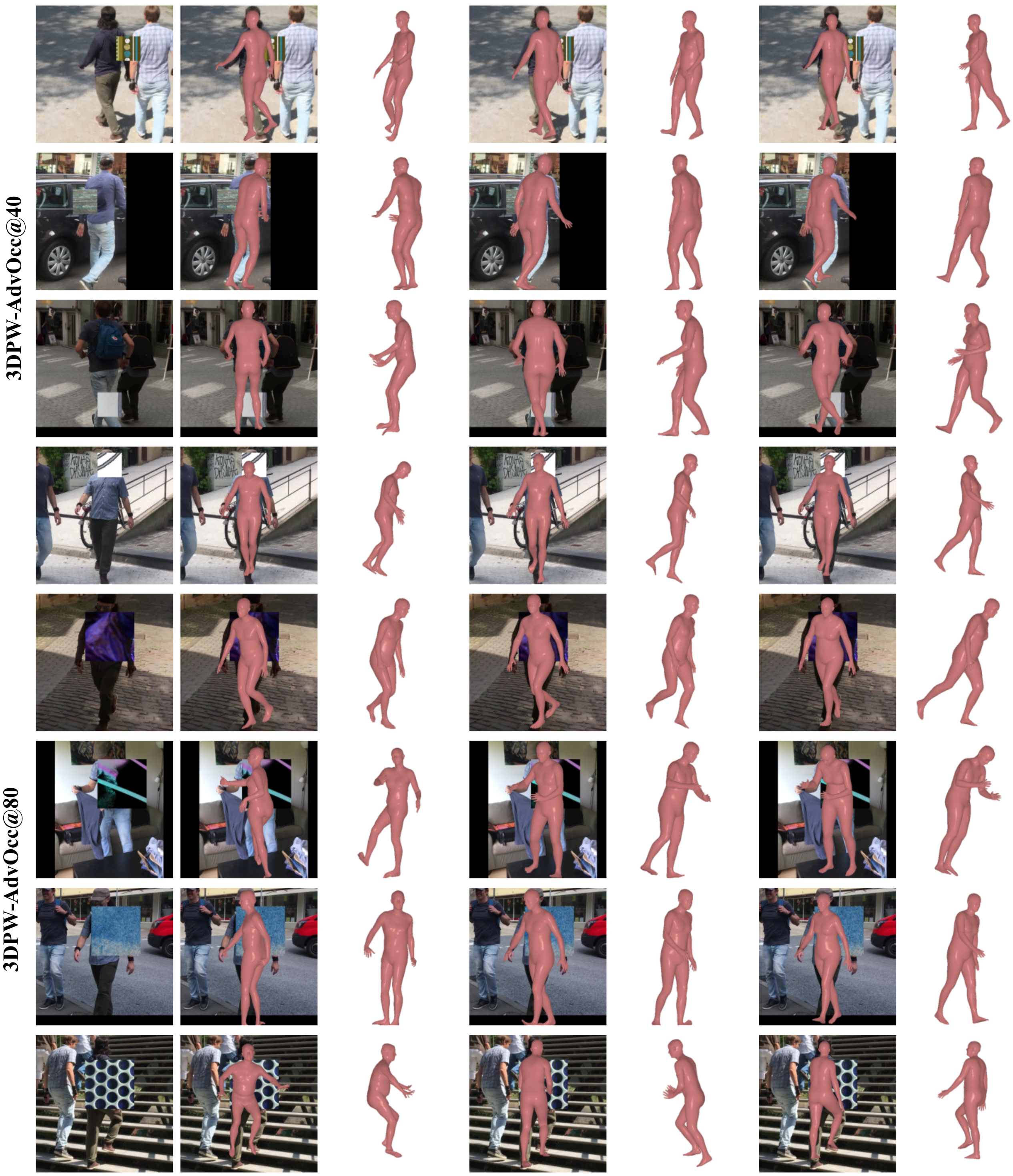
[6] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7708–7717, 2019. 2

[7] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11219–11229, 2021. 1

[8] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 2, 8

[9] 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. 6

Figure 8. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) Optimization results of EFT [2], (b) 3DNBF prediction and (c) ground truth pose. Initial poses are predicted by HMR-EFT [2].



(a) 3DPOF (b) 3DNBF (c) GT

Figure 9. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) Optimization results of 3DPOF [8], (b) 3DNBF prediction and (c) ground truth pose. Initial poses are predicted by HMR-EFT [2].