ECON: Explicit Clothed humans Optimized via Normal integration
Supplementary Material

Yuliang Xiu[1] Jinlong Yang[1] Xu Cao[2] Dimitrios Tzionas[3] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Osaka University, Japan [3]University of Amsterdam, the Netherlands
{yuliang.xiu, jinlong.yang, black}@tue.mpg.de cao.xu@ist.osaka-u.ac.jp d.tzionas@uva.nl

In the following, we provide more details and discussion on normal prediction, d-BiNI and IF-Nets+, as well as more qualitative results in the perceptual study, as an extension of Sec. 3 and Sec. 4 of the main paper. We also explore future applications. Please check the video on our website for an overview of the method and more results.

1. Implementation details

1.1. Normal map prediction

We set the loss weights $\lambda_{\text{J\_diff}}$, $\lambda_{\text{N\_diff}}$, and $\lambda_{\text{S\_diff}}$ in Eq. (1) to 5.0, 1.0, and 1.0 respectively. However, if the overlap ratio between clothing and body mask is smaller than 0.5, it means humans are dressed with loose clothing. In this situation we trust the 2D joints more and increase the $\lambda_{\text{J\_diff}} = 50.0$. Similarly, when the overlap between body mask inside the clothing mask and full body mask is smaller than 0.98, occlusion happens. In such cases we set $\lambda_{\text{S\_diff}} = 0.0$ to avoid limb self-intersection after pose refinement.

During inference, following ICON [11], we iteratively refine SMPL-X and clothed-body normals for 50 iterations (1.10 iter/s on Quadro RTX 5000 GPU). We use `rembg`[1] plus `Mask R-CNN` [4] for multi-person segmentation, `Mediapipe` [9] to estimate full-body landmarks, `Open3D` for poisson surface reconstruction [5], and `MonoPort` [7,8] for fast implicit surface query.

1.2. d-BiNI

Optimization details. To better present the optimization details, we first write the d-BiNI objective function in a matrix form. Figure 4 shows the four inputs to d-BiNI. We vectorize the front and back clothed and prior depth maps $\{\widehat{\mathcal{Z}}_{\text{F}}^{\text{c}}, \widehat{\mathcal{Z}}_{\text{B}}^{\text{c}}, \mathcal{Z}_{\text{F}}^{\text{b}}, \mathcal{Z}_{\text{B}}^{\text{b}}\}$ within Ω_{n} as $\{\widehat{\mathbf{z}}_{\text{F}}, \widehat{\mathbf{z}}_{\text{B}}, \mathbf{z}_{\text{F}}, \mathbf{z}_{\text{B}}\}$; all vectors are of length $|\Omega_{\text{n}}|$. d-BiNI then jointly solves for the front and back clothed depth $\widehat{\mathbf{z}}_{\text{F}}$ and $\widehat{\mathbf{z}}_{\text{B}}$ by minimizing the objective function consisting of the five terms:

$$
\begin{aligned}
\mathcal{L}(\widehat{\mathbf{z}}_{\text{F}}, \widehat{\mathbf{z}}_{\text{B}}) = &(\mathbf{A}_{\text{F}}\widehat{\mathbf{z}}_{\text{F}} - \mathbf{b}_{\text{F}})^{\top}\mathbf{W}_{\text{F}}(\mathbf{A}_{\text{F}}\widehat{\mathbf{z}}_{\text{F}} - \mathbf{b}_{\text{F}}) \\
&+(\mathbf{A}_{\text{B}}\widehat{\mathbf{z}}_{\text{B}} - \mathbf{b}_{\text{B}})^{\top}\mathbf{W}_{\text{B}}(\mathbf{A}_{\text{B}}\widehat{\mathbf{z}}_{\text{B}} - \mathbf{b}_{\text{B}}) \\
&+\lambda_{\text{d}}(\widehat{\mathbf{z}}_{\text{F}} - \mathbf{z}_{\text{F}})^{\top}\mathbf{M}(\widehat{\mathbf{z}}_{\text{F}} - \mathbf{z}_{\text{F}}) \\
&+\lambda_{\text{d}}(\widehat{\mathbf{z}}_{\text{B}} - \mathbf{z}_{\text{B}})^{\top}\mathbf{M}(\widehat{\mathbf{z}}_{\text{B}} - \mathbf{z}_{\text{B}}) \\
&+\lambda_{\text{s}}(\widehat{\mathbf{z}}_{\text{F}} - \widehat{\mathbf{z}}_{\text{B}})^{\top}\mathbf{S}(\widehat{\mathbf{z}}_{\text{F}} - \widehat{\mathbf{z}}_{\text{B}}).
\end{aligned} \tag{S.1}
$$

Here, $\mathbf{A}_{\text{F}} \in \mathbb{R}^{4|\Omega_{\text{n}}|\times|\Omega_{\text{n}}|}$ and $\mathbf{b}_{\text{F}} \in \mathbb{R}^{4|\Omega_{\text{n}}|}$ are constructed from the front normal map following Eq. (21) of BiNI [1]; $\mathbf{A}_{\text{B}}$ and $\mathbf{b}_{\text{B}}$ are from the back normal map. $\mathbf{W}_{\text{F}}$ and $\mathbf{W}_{\text{B}} \in \mathbb{R}^{4|\Omega_{\text{n}}|\times 4|\Omega_{\text{n}}|}$ are bilateral weight matrices for front and back depth maps, respectively; both are constructed following Eq. (22) of BiNI [1] and depend on the unknown depth. $\mathbf{M}$ and $\mathbf{S}$ are $|\Omega_{\text{n}}| \times |\Omega_{\text{n}}|$ diagonal matrices whose diagonal entries indicate the pixels with depth priors and located at the silhouette, respectively. Specifically, the i-th diagonal entry m_i of $\mathbf{M}$ is

$$
m_i = \begin{cases} 1, & \text{if } i\text{-th entry of } \widehat{\mathbf{z}}_{\text{F}} \text{ in } \Omega_{\text{z}} \\ 0, & \text{otherwise} \end{cases}, \tag{S.2}
$$

while the i-th diagonal entry s_i of $\mathbf{S}$ is

$$
s_i = \begin{cases} 1, & \text{if } i\text{-th entry of } \widehat{\mathbf{z}}_{\text{F}} \text{ in } \partial\Omega_{\text{n}} \\ 0, & \text{otherwise} \end{cases}. \tag{S.3}
$$

Stacking $\widehat{\mathbf{z}}_{\text{F}}$ and $\widehat{\mathbf{z}}_{\text{B}}$ as $\widehat{\mathbf{z}} = \begin{bmatrix} \widehat{\mathbf{z}}_{\text{F}} \\ \widehat{\mathbf{z}}_{\text{B}} \end{bmatrix}$, Eq. (S.1) then reads

$$
\begin{aligned}
\mathcal{L}(\widehat{\mathbf{z}}) = &(\mathbf{A}\widehat{\mathbf{z}} - \mathbf{b})^{\top}\mathbf{W}(\mathbf{A}\widehat{\mathbf{z}} - \mathbf{b}) + \\
&\lambda_{\text{d}}(\widehat{\mathbf{z}} - \mathbf{z})^{\top}\widetilde{\mathbf{M}}(\widehat{\mathbf{z}} - \mathbf{z}) + \lambda_{\text{s}}\widehat{\mathbf{z}}^{\top}\widetilde{\mathbf{S}}\widehat{\mathbf{z}},
\end{aligned} \tag{S.4}
$$

where

$$
\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\text{F}} & \\ & \mathbf{A}_{\text{B}} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_{\text{F}} \\ \mathbf{b}_{\text{B}} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}_{\text{F}} & \\ & \mathbf{W}_{\text{B}} \end{bmatrix},
$$

$$
\mathbf{z} = \begin{bmatrix} \mathbf{z}_{\text{F}} \\ \mathbf{z}_{\text{B}} \end{bmatrix}, \quad \widetilde{\mathbf{M}} = \begin{bmatrix} \mathbf{M} & \\ & \mathbf{M} \end{bmatrix}, \quad \widetilde{\mathbf{S}} = \begin{bmatrix} \mathbf{S} & -\mathbf{S} \\ -\mathbf{S} & \mathbf{S} \end{bmatrix}.
$$

[1] https://github.com/danielgatis/rembg

To minimize Eq. (S.4), we perform an iterative optimization similar to BiNI [1]. At each iteration, we first fix the weights $\mathbf{W}$ and jointly solve for the front and back depth $\hat{\mathbf{z}}$, then compute the new weights from the updated depth. When $\mathbf{W}$ is fixed and treated as a constant matrix, solving for the depth becomes a convex least-squares problem. The necessary condition for the global optimum is obtained by equating the gradient of Eq. (S.4) to $\mathbf{0}$:

$$(\mathbf{A}^\top \mathbf{W}\mathbf{A} + \lambda_{\mathrm{d}}\widetilde{\mathbf{M}} + \lambda_{\mathrm{s}}\widetilde{\mathbf{S}})\hat{\mathbf{z}} = \mathbf{A}^\top \mathbf{W}\mathbf{b} + \lambda_{\mathrm{d}}\widetilde{\mathbf{M}}\mathbf{z}. \quad \text{(S.5)}$$

Equation (S.5) is a large-scale sparse linear system with a symmetric positive definite coefficient matrix. We solve Eq. (S.5) using a CUDA-accelerated sparse conjugate gradient solver with a Jacobi preconditioner [2].

Hyper-parameters. d-BiNI has three hyper-parameters: λ_{d}, λ_{s}, and k. λ_{d} and λ_{s} are used in the objective function Eq. (3), which control the influence of coarse depth prior term Eq. (4) and silhouette consistency term Eq. (5) separately. k is used in the original BiNI [1] to control the surface stiffness (See Sup.Mat-A in BiNI [1] for more explanation of k). Empirically, we set $\lambda_{\mathrm{d}} = 1e^{-4}$, $\lambda_{\mathrm{s}} = 1e^{-6}$, and $k = 2$.

Discussion of hyper-paramters. Figure S.4 shows the d-BiNI integration results under different values of k. It can be seen that a small k leads to tougher d-BiNI surfaces where discontinuities are not accurately recovered, while a large k softens the surface, and redundant discontinuities and noisy artifacts are introduced. Figure S.5 shows the effects of λ_{d}, which controls how much d-BiNI surfaces agree on the SMPL-X mesh. Small λ_{d} causes misalignment between the d-BiNI surface and the SMPL-X mesh, which will produce stitching artifacts. While an excessively large λ_{d} enforces d-BiNI to rely over heavily on SMPL-X, thus smoothing out the high-frequency details obtained from normals. Figure S.6 justifies the necessity of the silhouette consistency term. Without this term, the front and back d-BiNI surfaces intersect each other around the silhouettes, which will cause "blobby" artifacts after screened Poisson reconstruction [6].

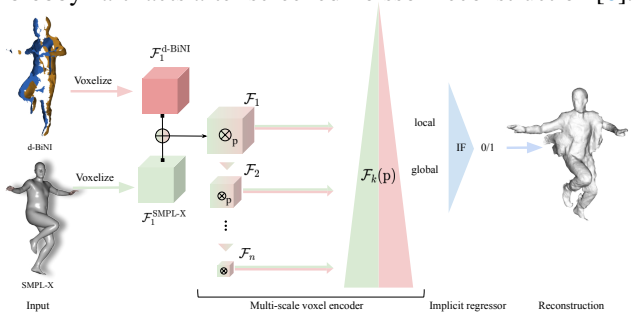

Figure S.1. **Overview of IF-Nets+.**

[2] https://docs.cupy.dev/en/stable/reference/generated/cupyx.scipy.sparse.linalg.cg.html

1.3. IF-Nets+

Network structure. As Fig. S.1 shows, similar to IF-Nets [2], IF-Nets+ applies multi-scale voxel 3D CNN encoding on voxelized d-BiNI and the SMPL-X surface, namely $\mathcal{F}_1^{\text{d-BiNI}}$ and $\mathcal{F}_1^{\text{SMPL-X}}$, generating multi-scale deep feature grids to account for both local and global information, $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n, \mathcal{F}_k \in \mathbb{R}^{K\times K\times K\times C_k}, n = 6$. These deep features are with decreasing resolution $K = \frac{N}{2^{k-1}}, N = 256$ and variable dimension channels $C = \{32, 32, 64, 128, 128, 128\}$. All these features are then fed into an implicit function regressor, parameterized by a Multi-Layer Perceptron (MLP), to predict the occupancy value of point P. This MLP regressor is trained with BCE loss.

Training setting. IF-Nets and IF-Nets+ share the same training setting. The voxelization resolution for both SMPL-X and d-BiNI surfaces is 256^3. We use RMSprop as an optimizer, with a learning rate $1e^{-4}$, and weight decay by a factor of 0.1 every 10 epochs. These networks are trained on an NVIDIA A100 for 20 epochs with a batch size of 48. Following ICON [11], we sampled 10000 points with the mixture of cube-uniform sampling and surface-around sampling, with standard deviation of 5cm.

Dataset details. We augment THuman2.0 [12] by (1) rotating the scans every 10 degrees around the yaw axis, to generate $525 \times 36 = 18900$ samples in total, and (2) randomly selecting a rectangle region from the d-BiNI depth maps, and erasing its pixels [14]. In particular, the erasing operation is being performed with $p = 0.8$ probability, the range of aspect ratio of erased area is between 0.3 and 3.3, and its range of proportion are $\{0.01, 0.05, 0.2\}$.

Speed analysis of ECON vs. ICON. d-BiNI takes 6.2 secs (150 iters). For ECON_{IF}, the IF-Nets+ plus Marching cubes takes 2.6 secs (for 256^3 resolution), and the Poisson step takes 10.7 secs (level=10). For a single image, ECON_{IF} takes 112 secs, and ECON_{EX} takes 97 secs. ICON, which shares the same SMPL-X fitting (w/ landmarks), takes 78 secs, and w/ cloth-refinement (50 iters) it takes 115 secs.

2. Qualitative results

Figure S.2 shows examples on SHHQ [3]. Figure S.3 shows PaMIR's results on the same photos in Fig. 9. Figures S.7 to S.9 show more comparisons used in our perceptual study, containing the results on in-the-wild images with challenging poses, loose clothing, and standard fashion poses, respectively. For each image, we display the results obtained by ECON, PaMIR [13], ICON [11], and PIFuHD [10]. In each row, we show normal maps rendered in $\{0°, 90°, 180°, 270°\}$ views. The video on our website shows more reconstructions with a rotating virtual camera.

Figure S.2. **SHHQ 3D reconstruction.** For each image we show a **front** and **side** view of ECON's reconstruction and a **SMPL-X** fit.

Figure S.3. **ECON** (Top) vs. **PaMIR** (Bottom) on loose clothes; **Zoom in** to see front/back 3D details.

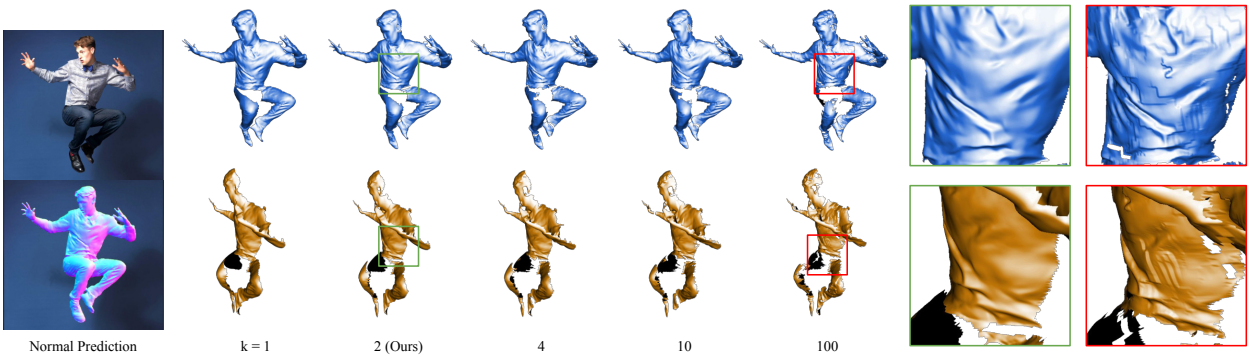

Figure S.4. **The effects of the hyper-parameter k on d-BiNI results.** k controls the stiffness of the target surface [1]. A smaller k leads to smooth d-BiNI surfaces, while a large k introduces unnecessary discontinuities and noise artifacts.

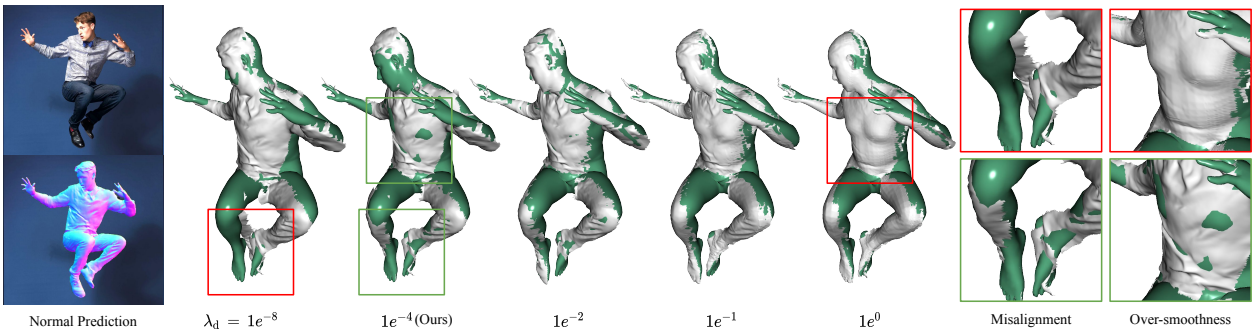

Figure S.5. **The effects of the hyperparameter λ_d on d-BiNI results.** λ_d controls how much d-BiNI surfaces agree with the SMPL-X mesh. A small λ_d causes a misalignment between the d-BiNI surface and the SMPL-X mesh, thus it produces stitching artifacts. An excessively large λ_d enforces d-BiNI to rely too heavily on SMPL-X, thus it smooths out the high-frequency details obtained from normals.

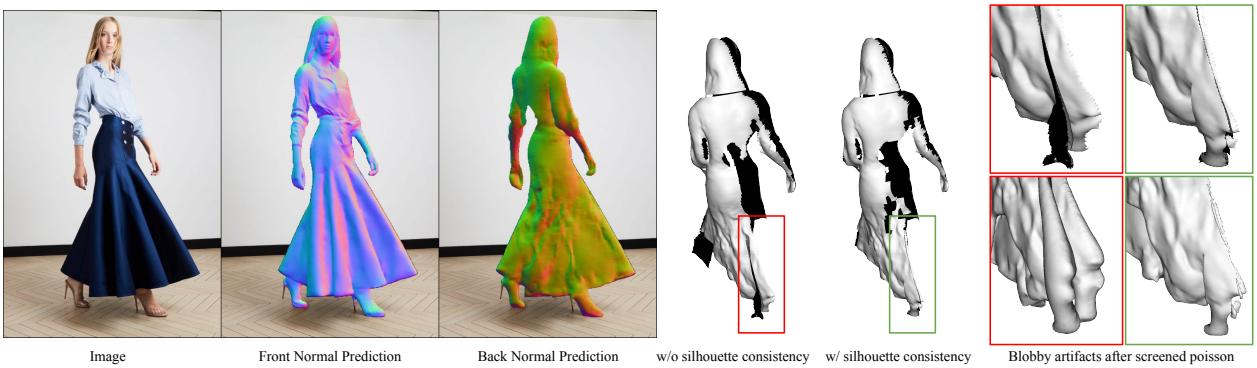

Figure S.6. **Necessity of silhouette consistency.** This term can be regarded as the mediator between front and back d-BiNI surfaces, preventing these surfaces from intersecting. Such intersection causes blobby artifacts after screened Poisson reconstruction [6].

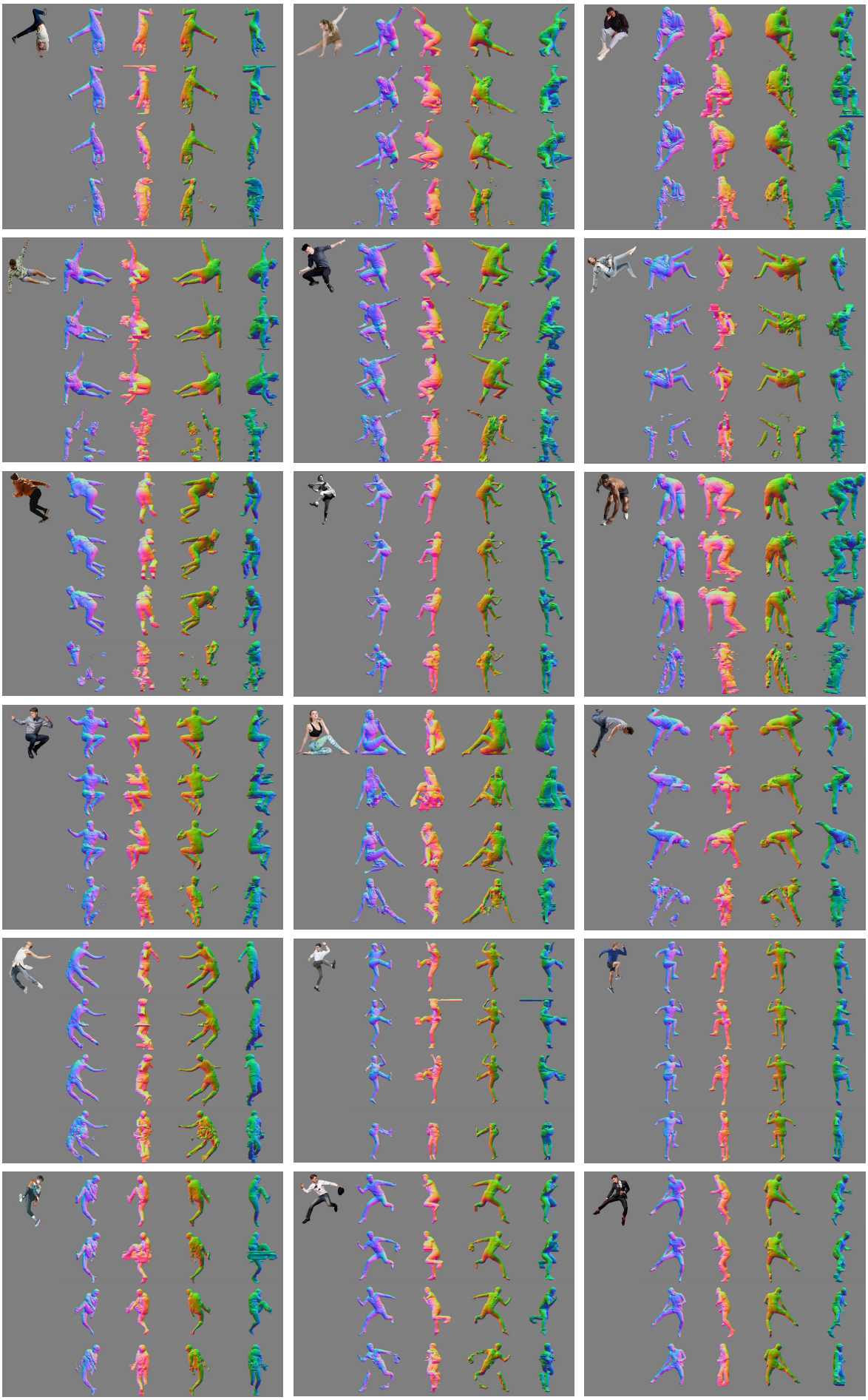

Figure S.7. **Results on in-the-wild images with challenging poses.** For each example the format is as follows: **Top** → **bottom:** ECON, PaMIR [13], ICON [11], and PIFuHD [10]. **Left** → **right:** Virtual camera rotated by $\{0°, 90°, 180°, 270°\}$. 🔍 **Zoom in** to see 3D details.

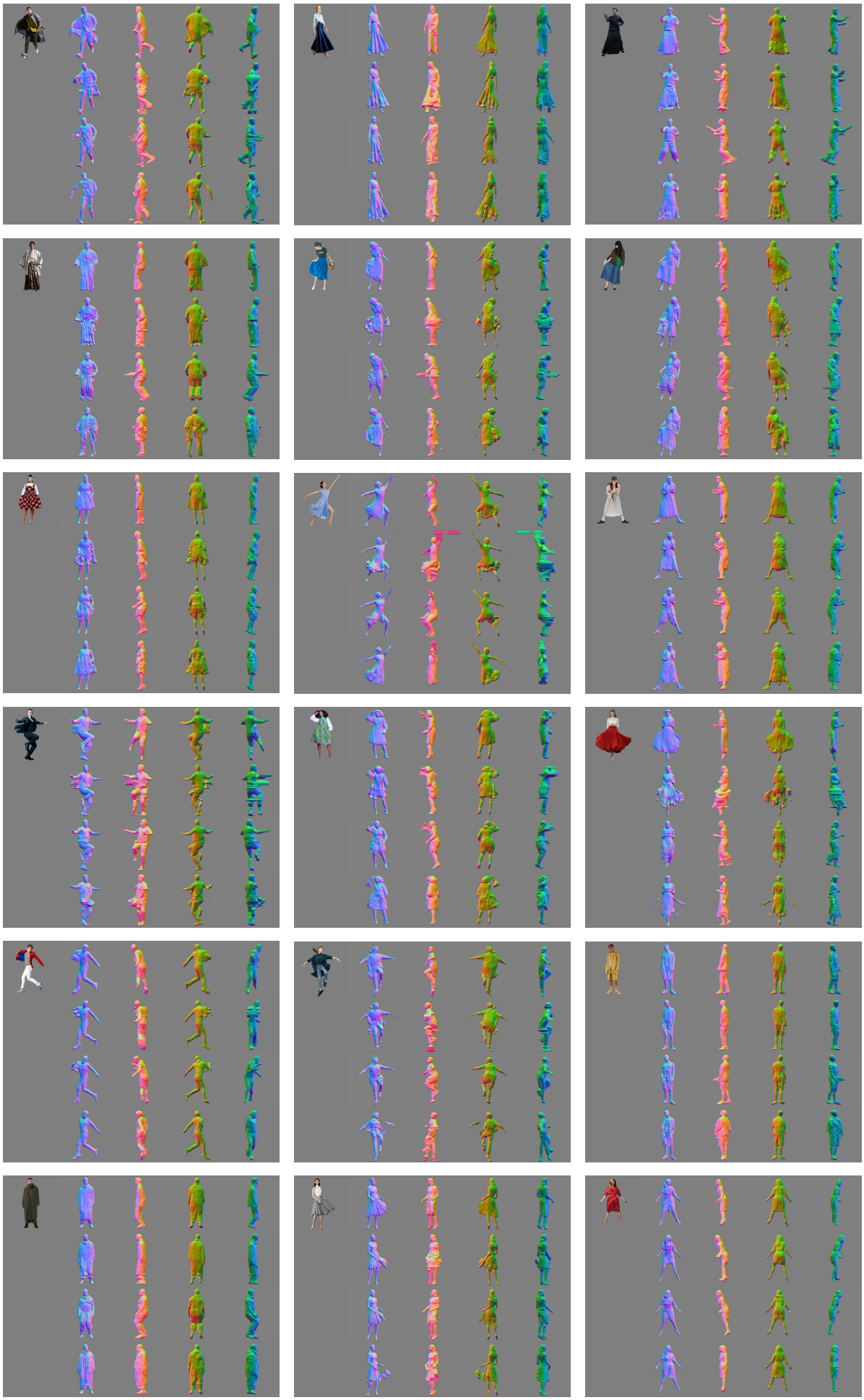

Figure S.8. **Results on in-the-wild images with loose clothing.** For each example the format is as follows: **Top → bottom:** ECON, PaMIR [13], ICON [11], and PIFuHD [10]. **Left → right:** Virtual camera rotated by $\{0°, 90°, 180°, 270°\}$. **Zoom in** to see 3D details.

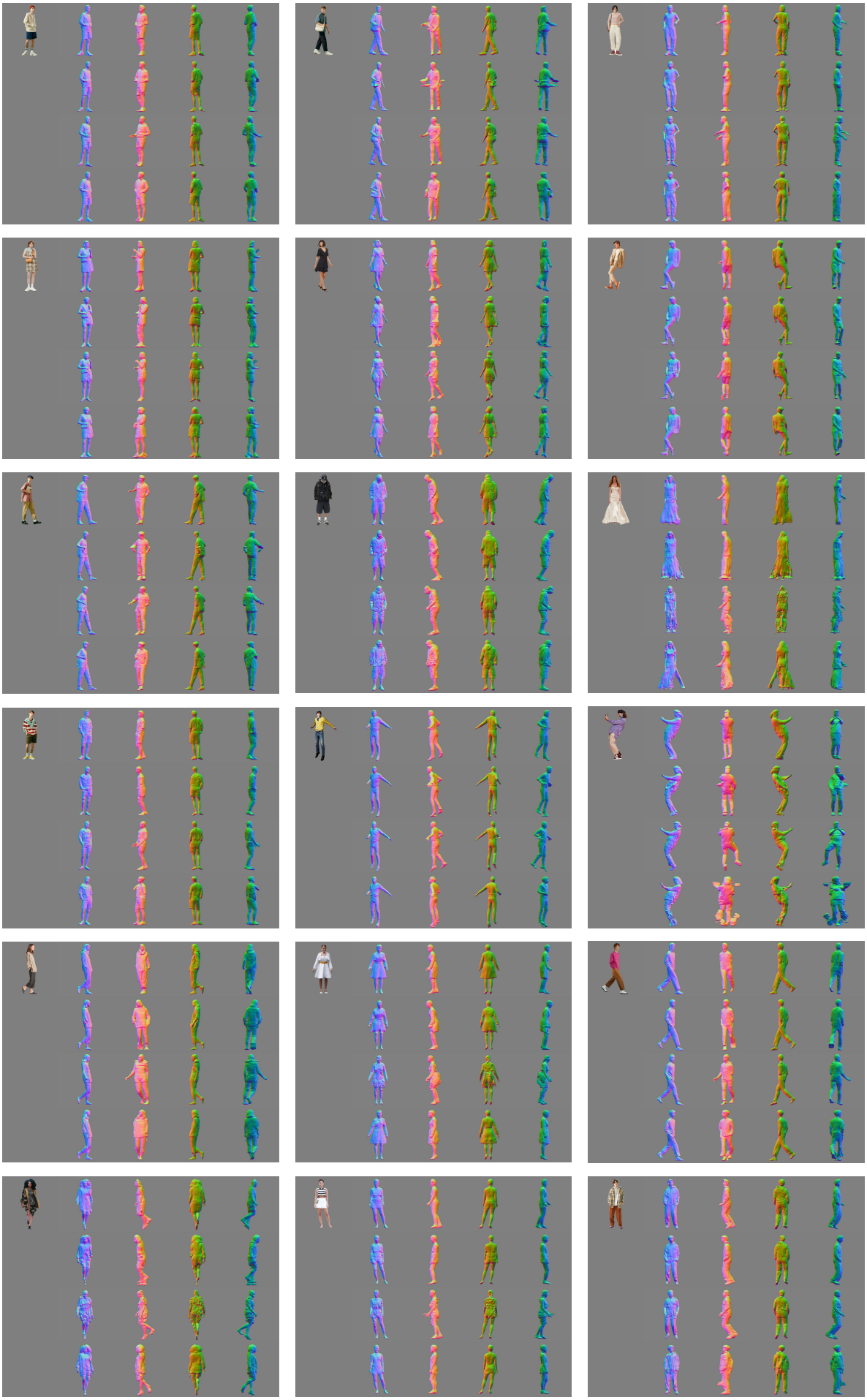

Figure S.9. **Results on in-the-wild fashion images.** For each example the format is as follows: **Top** → **bottom:** ECON, PaMIR [13], ICON [11], and PIFuHD [10]. **Left** → **right:** Virtual camera rotated by $\{0°, 90°, 180°, 270°\}$. 🔍 **Zoom in** to see 3D details.

References

[1] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4

[2] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3D shape reconstruction and completion. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[3] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[5] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing (SGP)*, 2006. 1

[6] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *Transactions on Graphics (TOG)*, 2013. 2, 4

[7] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *SIGGRAPH Real-Time Live*, 2020. 1

[8] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[9] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv:1906.08172*, 2019. 1

[10] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5, 6, 7

[11] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5, 6, 7

[12] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[13] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2, 5, 6, 7

[14] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI Conference on Artificial Intelligence*, 2020. 2