ECON: Explicit Clthed 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 \(\lambda_{\text{diff}}, \lambda_{\text{N-diff}}, \text{ and } \lambda_{\text{S-diff}}\) 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 \(\lambda_{\text{diff}} = 50.0\). 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 cloth and prior depth maps \(\{\tilde{z}_F, \tilde{z}_B, \tilde{z}_B, \tilde{z}_B\}\) within \(\Omega_n\) as \(\{\tilde{z}_F, \tilde{z}_B, \tilde{z}_F, \tilde{z}_B\}\); all vectors are of length \(|\Omega_n|\). d-BiNI then jointly solves for the front and back clothed depth \(\tilde{z}_F\) and \(\tilde{z}_B\) by minimizing the objective function consisting of the five terms:

\[
\mathcal{L}(\tilde{z}_F, \tilde{z}_B) = (A_F\tilde{z}_F - b_F)^\top W_F(A_F\tilde{z}_F - b_F) + (A_B\tilde{z}_B - b_B)^\top W_B(A_B\tilde{z}_B - b_B) + \lambda_d(\tilde{z}_F - z_F)^\top M(\tilde{z}_F - z_F) + \lambda_d(\tilde{z}_B - z_B)^\top M(\tilde{z}_B - z_B) + \lambda_s(\tilde{z}_F - z_B)^\top S(\tilde{z}_F - z_B)
\]

(S.1)

Here, \(A_F \in \mathbb{R}^{4|\Omega_n| \times |\Omega_n|}\) and \(b_F \in \mathbb{R}^{4|\Omega_n|}\) are constructed from the front normal map following Eq. (21) of BiNI \([1]\); \(A_B\) and \(b_B\) are from the back normal map. \(W_F\) and \(W_B \in \mathbb{R}^{4|\Omega_n| \times 4|\Omega_n|}\) 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. \(M\) and \(S\) are \(|\Omega_n| \times |\Omega_n|\) 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 \(M\) is

\[
m_i = \begin{cases} 1, & \text{if } \text{-th entry of } \tilde{z}_F \text{ in } \Omega_n \\ 0, & \text{otherwise} \end{cases}
\]

(S.2)

while the \(i\)-th diagonal entry \(s_i\) of \(S\) is

\[
s_i = \begin{cases} 1, & \text{if } \text{-th entry of } \tilde{z}_B \text{ in } \partial\Omega_n \\ 0, & \text{otherwise} \end{cases}
\]

(S.3)

Stacking \(\tilde{z}_F\) and \(\tilde{z}_B\) as \(\tilde{z} = [\tilde{z}_F \tilde{z}_B]\). Eq. (S.1) then reads

\[
\mathcal{L}(\tilde{z}) = (A\tilde{z} - b)^\top W(A\tilde{z} - b) + \lambda_d(\tilde{z} - z)^\top M(\tilde{z} - z) + \lambda_s\tilde{z}^\top S\tilde{z}
\]

(S.4)

where

\[
A = \begin{bmatrix} A_F & A_B \end{bmatrix}, \quad b = \begin{bmatrix} b_F \\ b_B \end{bmatrix}, \quad W = \begin{bmatrix} W_F \\ W_B \end{bmatrix}, \quad z = \begin{bmatrix} z_F \\ z_B \end{bmatrix}, \quad M = \begin{bmatrix} M \\ M \end{bmatrix}, \quad S = \begin{bmatrix} S \\ -S \end{bmatrix}.
\]

\(^1\)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 \( \mathbf{z} \), 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 0:

\[
(\mathbf{A}^\top \mathbf{W} + \lambda_\mathbf{d} \mathbf{M} + \lambda_\mathbf{s} \mathbf{S})\mathbf{z} = \mathbf{A}^\top \mathbf{Wb} + \lambda_\mathbf{d} \mathbf{Mz}. \tag{S.5}
\]

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.

**Hyper-parameters.** d-BiNI has three hyper-parameters: \( \lambda_s \), \( \lambda_d \), and \( k \). \( \lambda_s \) and \( \lambda_d \) 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 \( \lambda_d = 1e^{-4} \), \( \lambda_s = 1e^{-6} \), and \( k = 2 \).

**Discussion of hyper-parameters.** 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 \( \lambda_d \), which controls how much d-BiNI surfaces agree on the SMPL-X mesh. Small \( \lambda_d \) causes misalignment between the d-BiNI surface and the SMPL-X mesh, which will produce stitching artifacts. While an excessively large \( \lambda_d \) enforces d-BiNI to rely overly 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 “bloppy” artifacts after screened Poisson reconstruction [6].

### 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 \( \mathcal{F}_1^{d\text{-BiNI}} \) and \( \mathcal{F}_1^{\text{SMPL-X}} \), generating multi-scale deep feature grids to account for both local and global information, \( \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n, \mathcal{F}_k \in \mathbb{R}^{K \times K \times C \times \mathbf{K}}, n = 6 \). These deep features are with decreasing resolution \( K = \frac{N}{2^{n-1}}, N = 256 \) 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\(_\text{IF}\), 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\(_\text{IF}\) takes 112 secs, and ECON\(_\text{EX}\) 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\(^\circ\), 90\(^\circ\), 180\(^\circ\), 270\(^\circ\)\} 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 $\lambda_d$ on d-BiNI results.** $\lambda_d$ controls how much d-BiNI surfaces agree with the SMPL-X mesh. A small $\lambda_d$ causes a misalignment between the d-BiNI surface and the SMPL-X mesh, thus it produces stitching artifacts. An excessively large $\lambda_d$ 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


