ECON: Explicit Clothed humans Optimized via Normal integration

Yuliang Xiu\textsuperscript{1} Jinlong Yang\textsuperscript{1} Xu Cao\textsuperscript{2} Dimitrios Tzionas\textsuperscript{3} Michael J. Black\textsuperscript{1}

\textsuperscript{1}Max Planck Institute for Intelligent Systems, Tübingen, Germany
\textsuperscript{2}Osaka University, Japan \textsuperscript{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

Abstract

The combination of deep learning, artist-curated scans, and Implicit Functions (IF), is enabling the creation of detailed, clothed, 3D humans from images. However, existing methods are far from perfect. IF-based methods recover free-form geometry, but produce disembodied limbs or degenerate shapes for novel poses or clothes. To increase robustness for these cases, existing work uses an explicit parametric body model to constrain surface reconstruction, but this limits the recovery of free-form surfaces such as loose clothing that deviates from the body. What we want is a method that combines the best properties of implicit representation and explicit body regularization. To this end, we make two key observations: (1) current networks are better at inferring detailed 2D maps than full-3D surfaces, and (2) a parametric model can be seen as a “canvas” for stitching together detailed surface patches. Based on these, our method, ECON, has three main steps: (1) It infers detailed 2D normal maps for the front and back side (Sec. 3.1). (2) The normal maps are converted into detailed, yet incomplete, 2.5D front and back surfaces guided by a SMPL-X estimate (Sec. 3.2). (3) It then “inpaints” the missing geometry between two surfaces (Sec. 3.3). Face or hands can be optionally replaced with the cleaner ones from SMPL-X. See the video on our website for more results.

1. Introduction

Human avatars will be key for future games and movies, mixed-reality, tele-presence and the “metaverse”. To build realistic and personalized avatars at scale, we need to faithfully reconstruct detailed 3D humans from color photos taken in the wild. This is still an open problem, due to its challenges; people wear all kinds of different clothing and accessories, and they pose their bodies in many, often imaginative, ways. A good reconstruction method must accurately capture these, while also being robust to novel clothing and poses.
Initial, promising, results have been made possible by using artist-curated scans as training data, and implicit functions (IF) \([56, 59]\) as the 3D representation. Seminal work on PIFuHD \([70, 71]\) uses “pixel-aligned” IF and reconstructs clothed 3D humans with unconstrained topology. However, these methods tend to overfit to the poses seen in the training data, and have no explicit knowledge about the human body’s structure. Consequently, they produce disembodied limbs or degenerate shapes for images with novel poses; see the 2nd row of Fig. 2. Follow-up work \([26, 82, 96]\) accounts for such artifacts by regularizing the IF using a shape prior provided by an explicit body model \([52, 61]\), but regularization introduces a topological constraint, restricting generalization to novel clothing while attenuating shape details; see the 3rd and 4th rows of Fig. 2. In a nutshell, there are trade-offs between robustness, generalization and detail.

What we want is the best of both worlds; that is, the robustness of explicit anthropomorphic body models, and the flexibility of IF to capture arbitrary clothing topology. To that end, we make two key observations: (1) While inferring detailed 2D normal maps from color images is relatively easy \([31, 71, 82]\), inferring 3D geometry with equally fine details is still challenging \([9]\). Thus, we exploit networks to infer detailed “geometry-aware” 2D maps that we then lift to 3D. (2) A body model can be seen as a low-frequency “canvas” that “guides” the stitching of detailed surface parts.

With these in mind, we develop ECON, which stands for “Explicit Clothed humans Optimized via Normal integration”. It takes, as input, an RGB image and a SMPL-X body inferred from the image. Then, it outputs a 3D human in free-form clothing with a level of detail and robustness that goes beyond the state of the art (SOTA); see the bottom of Fig. 2. Specifically, ECON has three steps.

Step 1: Front & back normal reconstruction. We predict front- and back-side clothed-human normal maps from the input RGB image, conditioned on the body estimate, with a standard image-to-image translation network.

Step 2: Front & back surface reconstruction. We take the previously predicted normal maps, and the corresponding depth maps that are rendered from the SMPL-X mesh, to produce detailed and coherent front-/back-side 3D surfaces, \(\{M_F, M_B\}\). To this end, we extend the recent BiNI method \([7]\), and develop a novel optimization scheme that is aimed at satisfying three goals for the resulting surfaces: (1) their high-frequency components agree with clothed-human normals, (2) their low-frequency components and the discontinuities agree with the SMPL-X ones, and (3) the depth values on their silhouettes are coherent with each other and consistent with the SMPL-X-based depth maps. The two output surfaces, \(\{M_F, M_B\}\), are detailed yet incomplete, i.e., there is missing geometry in occluded and “profile” regions.

Step 3: Full 3D shape completion. This module takes two inputs: (1) the SMPL-X mesh, and (2) the two d-BiNI surfaces, \(\{M_F, M_B\}\). The goal is to “inpaint” the missing geometry. Existing methods struggle with this problem. On one hand, Poisson reconstruction \([38]\) produces “blobby” shapes and naively “infills” holes without exploiting a shape distribution prior. On the other hand, data-driven approaches, such as IF-Nets \([10]\), struggle with missing parts caused by (self-)occlusions, and fail to keep the fine details present on two d-BiNI surfaces, producing degenerate geometries.

We address above the limitations in two steps: (1) We extend and re-train IF-Nets to be conditioned on the SMPL-X body, so that SMPL-X regularizes shape “infilling”. We discard the triangles that lie close to \(\{M_F, M_B\}\), and keep the remaining ones as “infilling patches”. (2) We stitch together the front- and back-side surfaces and infilling patches via Poisson reconstruction; note that holes between these are small enough for a general purpose method. The result is a full 3D shape of a clothed human; see Fig. 2, bottom.

We evaluate ECON both on established benchmarks (CAPE \([55]\) and Renderpeople \([66]\)) and in-the-wild images. Quantitative analysis reveals ECON’s superiority. A perceptual study echoes this, showing that ECON is significantly preferred over competitors on challenging poses and loose clothing, and competitive with PIFuHD on fashion images. Qualitative results show that ECON generalizes better than the SOTA to a wide variety of poses and clothing, even with extreme looseness or complex topology; see Fig. 9.

With both pose-robustness and topological flexibility, ECON recovers 3D clothed humans with a good level of detail and realistic pose. Code and models are available for research purposes at econ.is.tue.mpg.de
2. Related Work

Image-based clothed human reconstruction. Regarding geometric representation, we group the mainstream clothed human reconstruction approaches into “implicit” and “explicit”. Note that with the terms implicit/explicit we mainly refer to the surface decoder rather than the feature encoder.

1) Explicit-shape-based approaches use either a mesh-based parametric body model [35, 52, 61, 69, 83], or a non-parametric depth map [18, 72] or point cloud [90], to reconstruct 3D humans. Many methods [15, 36, 39, 40, 42, 43, 75, 76, 87, 91, 92] estimate or regress minimally-clothed 3D body meshes from RGB pixels and ignore clothing. To account for clothed human shapes, another line of work [2–4, 41, 55, 62, 79, 100] adds 3D offsets on top of the body mesh. This is compatible with current animation pipelines, as they inherit the hierarchical skeleton and skinning weights from the underlying statistical body model. However, this “body-offset” approach is not flexible enough to model loose clothing, which deviates significantly from the body topology, such as dresses and skirts. To increase topological flexibility, some methods [5, 33] reconstruct 3D clothed humans by identifying the type of clothing and using the appropriate model to reconstruct it. Scaling up this “cloth-aware” approach to many clothing styles is nontrivial, limiting generalization to in-the-wild outfit variation.

2) Implicit-function-based approaches are topology-agnostic and, thus, can be used to represent arbitrary 3D clothed human shapes. SMPLicit [12], ClothWild [57] and DIG [46] learn a generative clothing model with neural distance fields [11, 56, 59] from a 3D clothing dataset. Given an image, the clothed human is reconstructed by estimating a parametric body and optimizing the latent space of the clothing model. However, the results usually do not align well with the image and lack geometric detail.

PIFu [70] introduces pixel-aligned implicit human shape reconstruction and PIFuHD [71] significantly improves the geometric details with a multi-level architecture and normal maps predicted from the RGB image. However, these two methods do not exploit knowledge of the human body structure. Therefore, these methods offset the body poses in the training data, e.g. fashion poses. They fail to generalize to novel poses, producing non-human shapes with broken or disembodied limbs. To address these issues, several methods introduce different geometric priors to regularize the deep implicit representation: GeoPIFu [26] introduces a coarse shape of volumetric humans, Self-Portraits [49], PINA [14], and S3 [85] use depth or LIDAR information to regularize shape and improve robustness to pose variation.

Another direction leverages parametric body models, which represent human body shape well, model the kinematic structure of the body, and can be reliably estimated from RGB images of clothed people. Such a representation can be viewed as a base shape upon which to model clothed humans. Therefore, several methods combine parametric body models with expressive implicit representations to get the best of both worlds. PaMIR [96] and DeepMultiCap [95] condition the pixel-aligned features on a posed and voxelized SMPL mesh. JIFF introduces a 3DMM face prior to improve the realism of the facial region. ARCH [30], ARCH++ [27] and CAR [50] use SMPL to unpose the pixel-aligned query points from a posed space to a canonical space. To further generalize to unseen poses on in-the-wild photos, ICON [82] regresses shapes from locally-queried features. However, the above approaches gain robustness to unseen poses at the cost of generalization ability to various, especially loose, clothing topologies. We argue that this is because loose clothing differs greatly from human body and that conditioning on the SMPL body in 3D makes it harder for networks to make full use of 2D image features.

Our work is also inspired by “sandwich-like” monocular reconstruction approaches, represented by Modulating Humans [18], FACSIMILE [72] and Any-Shot GIN [80]. Modulating Humans has two networks: a generator that estimates the visible (front) and invisible (back) depth maps from RGB images, and a discriminator that helps regularize the estimation via an adversarial loss. FACSIMILE further improves the geometric details by leveraging a normal loss, which is directly computed from depth estimates via differentiable layers. Recently, Any-Shot GIN generalizes the sandwich-like scheme to novel classes of objects. Given RGB images, it predicts front and back depth maps as well, and then exploits IF-Nets [10] for shape completion. We follow a similar path and extend it, to successfully reconstruct clothed human shapes with SOTA pose generalization, and better details from normal images.

3. Method

Given an RGB image, ECON first estimates front and back normal maps (Sec. 3.1), then converts them into front and back partial surfaces (Sec. 3.2), and finally “inpaints” the missing geometry with the help of IF-Nets+ (Sec. 3.3). See ECON’s overview in Fig. 3.

3.1. Detailed normal map prediction

Trained on abundant pairs of RGB images and normal images, a “front” normal map, \( \hat{N}^f \), can be accurately estimated from an RGB image using image-to-image translation networks, as demonstrated in PIFuHD [71] or ICON [82]. Both methods also infer a “back” normal map, \( \hat{N}^b \), from the image. But, the absence of image cues leads to over-smooth \( \hat{N}^b \). To address this, we fine-tune ICON’s backside normal predictor, \( \hat{N}^b \), with an additional MRF loss [77] to enhance the local details by minimizing the difference between the predicted \( \hat{N}^c \) and ground truth (GT) \( N^c \) in feature space.

To guide the normal map prediction and make it robust to various body poses, ICON conditions the normal map
We expect these 2.5D surfaces to satisfy three conditions: (1) a depth prior \( L \) applied to the front and back depth surfaces, and \( N^b \); (2) low-frequency surface variations, including high-frequency surface details agree with predicted clothed normal maps, \( N^c \); and (3) the depth-normal relationship of the front and back silhouettes are close to each other.

Apart from the \( L \) between 2D landmarks \( J \) and \( S \), we explicitly model the depth-normal relationship using variational normal integration methods \([7, 64]\). Specifically, we optimize SMPL-X’s shape, \( \beta \), pose, \( \theta \), and translation, \( \tau \), to minimize:

\[
L_{\text{SMPL-X}} = L_{N, \text{diff}} + L_{S, \text{diff}} + L_{J, \text{diff}} = \lambda_J |J^b - J^c|,
\]

where \( L_{N, \text{diff}} \) and \( L_{S, \text{diff}} \) are the normal-map loss and silhouette loss introduced in ICON \([82]\), and \( L_{J, \text{diff}} \) is the joint loss (L2) between 2D landmarks \( J^c \), which are estimated by a 2D keypoint estimator from the RGB image \( I \), and the corresponding re-projected 2D joint \( J^b \) from \( M^b \). For more implementation details, see Sec. A.1 in SupMat.

### 3.2. Front and back surface reconstruction

We now lift the clothed normal maps to 2.5D surfaces. We expect these 2.5D surfaces to satisfy three conditions: (1) high-frequency surface details agree with predicted clothed normal maps, (2) low-frequency surface variations, including discontinuities, agree with SMPL-X’s ones, and (3) the depth of the front and back silhouettes are close to each other.

Unlike PIFuHD \([71]\) or ICON \([82]\), which train a neural network to regress the implicit surface from normal maps, we explicitly model the depth-normal relationship using variational normal integration methods \([7, 64]\). Specifically, we tailor the recent bilateral normal integration (BiNI) method \([7]\) to full-body mesh reconstruction by harnessing the coarse prior, depth maps, and silhouette consistency.

To satisfy the three conditions, we propose a depth-aware silhouette-consistent bilateral normal integration (d-BiNI) method to jointly optimize for the front and back clothed depth maps, \( \tilde{Z}_F^c \) and \( \tilde{Z}_B^c \):

\[
d\text{BiNI}(\tilde{N}_F^c, \tilde{N}_B^c, \tilde{Z}_F^b, \tilde{Z}_B^b) \rightarrow \tilde{Z}_F^c, \tilde{Z}_B^c. \tag{2}
\]

Here, \( \tilde{N}_F^c \) is the front or back clothed normal map predicted by \( G^b \) from \( \{I, N^b\} \), and \( \tilde{Z}^b \) is the front or back coarse body depth image rendered from the SMPL-X mesh, \( M^b \).

Specifically, our objective function consists of five terms:

\[
\min_{\tilde{Z}_F^c, \tilde{Z}_B^c} L_n(\tilde{Z}_F^c; \tilde{N}_F^c) + L_n(\tilde{Z}_B^c; \tilde{N}_B^c) + \lambda_d L_d(\tilde{Z}_F^c; \tilde{Z}_F^b) + \lambda_d L_d(\tilde{Z}_B^c; \tilde{Z}_B^b) + \lambda_s L_s(\tilde{Z}_F^c; \tilde{Z}_B^c), \tag{3}
\]

where \( L_n \) is the BiNI loss term introduced by BiNI \([7]\), \( L_d \) is a depth prior applied to the front and back depth surfaces, and \( L_s \) is a front-back silhouette consistency term. For a more detailed discussion on these terms, see Sec. A.2 in SupMat.

With Eq. (3), we make two technical contributions beyond BiNI \([7]\). First, we use the coarse depth prior rendered from the SMPL-X body mesh, \( Z^b \), to regularize BiNI:

\[
L_d(\tilde{Z}^c; \tilde{Z}^b) = |\tilde{Z}^c - \tilde{Z}^b|_{\Omega_i \cap \Omega_j}, \quad i \in \{F, B\}. \tag{4}
\]

Figure 3. Overview. ECON takes as input an RGB image, \( I \), and a SMPL-X body, \( M^b \). Conditioned on the rendered front and back body normal images, \( N^b \), ECON first predicts front and back clothing normal maps, \( N^c \). These two maps, along with body depth maps, \( Z^b \), are fed into a d-BiNI optimizer to produce front and back surfaces, \( \{M_F, M_B\} \). Based on such partial surfaces, and body estimate \( M^b \), IF-Nets+ implicitly completes \( R \). With optional Face or hands from \( M^b \), screened Poisson combines everything as final watertight \( R \).
This addresses the key problem of putting the front and back surfaces together in a coherent way to form a full body. Optimizing BiNI terms \( \mathcal{L}_n \) leaves an arbitrary global offset between the front and back surfaces. The depth prior terms \( \mathcal{L}_d \) encourage the surfaces with undecided offsets to be consistent with the SMPL-X body, and is computed in the domains \( \Omega_n \cap \Omega_e \) (Fig. 4). For further intuitions on \( \mathcal{L}_n \) and \( \mathcal{L}_d \), see Fig. S.4 and Fig. S.5 in SupMat.

Second, we use a silhouette consistency term to encourage the front and back depth values to be the same at the silhouette boundary, which is computed in domain \( \partial \Omega_n \) (Fig. 4):

\[
\mathcal{L}_s(\tilde{z}_F, \tilde{z}_B) = |\tilde{z}_B - \tilde{z}_F|_{\partial \Omega_n}, \quad (5)
\]

The silhouette term improves the physical consistency of the reconstructed front and back clothed depth maps. Without this term, d-BiNI produces intersections of the front and back surfaces around the silhouette, causing “blobby” artifacts and hurting reconstruction quality; see Fig. S.6 in SupMat.

### 3.3. Human shape completion

For simple body poses without self-occlusions, merging front and back d-BiNI surfaces in a straightforward way, as done in FACSIMILE [72] and Modulating Humans [18], can result in a complete 3D clothed scan. However, often poses result in self-occlusions, which cause large portions of the surfaces to be missing. In such cases, Poisson Surface Reconstruction (PSR) [37] leads to blobby artifacts.

**PSR completion with SMPL-X (ECON_{EX}).** A naive way to “infill” the missing surface is to make use of the estimated SMPL-X body. We remove the triangles from \( \mathcal{M}^b \) that are visible to front or back cameras. The remaining triangle “soup” \( \mathcal{M}^{all} \) contains both side-view boundaries and occluded regions. We apply PSR [37] to the union of \( \mathcal{M}^{all} \) and d-BiNI surfaces \( \{ \mathcal{M}_F, \mathcal{M}_B \} \) to obtain a watertight reconstruction \( \mathcal{R} \). This approach is denoted as ECON_{EX}. Although ECON_{EX} avoids missing limbs or sides, it does not produce a coherent surface for the originally missing clothing and hair surfaces because of the discrepancy between SMPL-X and actual clothing or hair; see ECON_{EX} in Fig. 5.

**Inpainting with IF-Nets+ (R_{IF}).** To improve reconstruction coherence, we use a learned implicit-function (IF) model to “inpaint” the missing geometry given front and back d-BiNI surfaces. Specifically, we tailor a general-purpose shape completion method, IF-Nets [10], to a SMPL-X-guided one, denoted as IF-Nets+. IF-Nets [10] completes the 3D shape from a deficient 3D input, such as an incomplete 3D human shape or a low-resolution voxel grid. Inspired by Li et al. [44], we adapt IF-Nets by conditioning it on a voxelized SMPL-X body to deal with pose variation; for details see Sec. A.3 in SupMat. IF-Nets+ is trained on voxelized front and back ground-truth clothed depth maps, \( \{ \tilde{z}_F, \tilde{z}_B \} \), and a voxelized (estimated) body mesh, \( \mathcal{M}^b \), as input, and is supervised with ground-truth 3D shapes. During training, we randomly mask \( \{ \tilde{z}_F, \tilde{z}_B \} \) for robustness to occlusions. During inference, we feed the estimated \( \tilde{z}_F, \tilde{z}_B \) and \( \mathcal{M}^b \) into IF-Nets+ to obtain an occupancy field, from which we extract the inpainted mesh, \( \mathcal{R}_{IF} \), with Marching cubes [53].

**PSR completion with R_{IF} (ECON_{IF}).** To obtain our final mesh, \( \mathcal{R} \), we apply PSR to stitch (1) d-BiNI surfaces, (2) sided and occluded triangle soup \( \mathcal{M}^{all} \) from \( \mathcal{R}_{IF} \), and optionally, (3) face or hands cropped from the estimated SMPL-X body. Specifically, we tailor a general-purpose shape completion method, IF-Nets [10], to a SMPL-X-guided one, denoted as IF-Nets+. IF-Nets [10] completes the 3D shape from a deficient 3D input, such as an incomplete 3D human shape or a low-resolution voxel grid. Inspired by Li et al. [44], we adapt IF-Nets by conditioning it on a voxelized SMPL-X body to deal with pose variation; for details see Sec. A.3 in SupMat. IF-Nets+ is trained on voxelized front and back ground-truth clothed depth maps, \( \{ \tilde{z}_F, \tilde{z}_B \} \), and a voxelized (estimated) body mesh, \( \mathcal{M}^b \), as input, and is supervised with ground-truth 3D shapes. During training, we randomly mask \( \{ \tilde{z}_F, \tilde{z}_B \} \) for robustness to occlusions. During inference, we feed the estimated \( \tilde{z}_F, \tilde{z}_B \) and \( \mathcal{M}^b \) into IF-Nets+ to obtain an occupancy field, from which we extract the inpainted mesh, \( \mathcal{R}_{IF} \), with Marching cubes [53].

**Face and hand details.** The face and hands of the raw reconstruction can be replaced with the ones of the SMPL-X body.
body $M^b$. The necessity of (3) arises from the poorly reconstructed hands/face in $R_{IF}$, see difference in Fig. 6. The approach is denoted as ECON_{IF}.

Notably, although $R_{IF}$ is already a complete human mesh, due to the lossy voxelization of inputs and limited resolution of Marching cubes algorithm, it somehow smooths out the details of $\tilde{Z}_{E}^{c}_{EB}/M_{EB}$, which are optimized via d-BiNI (see $R_{IF}$ vs ECON_{IFEX} in Fig. 5). While ECON_{IFEX} preserves d-BiNI details better, only the side-views and occluded parts of $R_{IF}$ are fused in the Poisson step. In Tabs. 1 and 4, we use ECON_{IFEX} instead of $R_{IF}$ for evaluation.

4. Experiments

4.1. Datasets

Training on THuman2.0. THuman2.0 [88] contains 525 high-quality human textured scans in various poses, which are captured by a dense DSLR rig, along with their corresponding SMPL-X fits. We use THuman2.0 to train ICON, ECON_{IF} (IF-Nets+), IF-Nets, PIFu and PaMIR. Quantitative evaluation on CAPE & Renderpeople. We primarily evaluate on CAPE [55] and Renderpeople [66]. Specifically, we use the “CAPE-NFP” set (100 scans), which is used by ICON to analyze robustness to complex human poses. Moreover, we select another 100 scans from Renderpeople, containing loose clothing, such as dresses, skirts, robes, down jackets, costumes, etc. With such clothing variance, Renderpeople helps numerically evaluate the flexibility of reconstruction methods w.r.t. shape topology. Samples of the two datasets are shown in Fig. 7.

4.2. Metrics

Chamfer and P2S distance (cm). To capture large geometric errors, e.g. occluded parts or wrongly positioned limbs, we report the commonly used Chamfer (bi-directional point-to-surface) and P2S distance (1-directional point-to-surface) between ground-truth and reconstructed meshes. Normal difference (L2). To measure the fineness of reconstructed local details, as well as projection consistency from the input image, we also report the L2 error between normal images rendered from reconstructed and ground-truth surfaces, by rotating a virtual camera around these by $\{0^{\circ},90^{\circ},180^{\circ},270^{\circ}\}$ w.r.t. to a frontal view.

4.3. Evaluation

Quantitative evaluation. We compare ECON with body-agnostic methods, i.e., PIFu [70] and PIFuHD [71], and body-aware methods, i.e., PaMIR [96] and ICON [82], see in Tab. 1. For fair comparison, we use re-implementations of PIFu and PaMIR from ICON [82], because they have the same network settings and input data. ECON_{EX} performs on par with ICON, and outperforms other methods on images containing out-of-distribution (OOD) poses (CAPE), with a distance error below 1cm. In terms of out-of-distribution outfits (Renderpeople), ECON_{EX} performs on par with PaMIR, and much better than PIFuHD. When it comes to high-frequency details measured by normals, ECON_{EX} achieves SOTA performance on both datasets.

Perceptual study. Due to the lack of ground-truth geometry (clothed scan + underneath SMPL-X), we further conduct a perceptual study to evaluate ECON on in-the-wild images. Test images are divided into three categories: “challenging poses”, “loose clothing”, and “fashion images”. Examples of challenging poses and loose clothing can be seen in Fig. 9, and some of fashion images are in SupMat.’s Fig. S.2.

Participants are asked to choose the reconstruction they perceive as more realistic, between a baseline method and ECON. We compute the chances that each baseline is preferred over ECON in Tab. 2. The results of the perceptual study confirm the quantitative evaluation in Tab. 1. For “challenging poses” images, ECON is significantly preferred over PIFuHD and outperforms ICON. On images of people wearing loose clothing, ECON is preferred over ICON by a large

Table 1. Evaluation against the state of the art. All models use a resolution of 256 for marching cubes. * Methods are re-implemented in [82] for a fair comparison in terms of network settings and training data. † Official model is trained on the Renderpeople dataset. ECON_{EX} is optimization-based, thus requires no training (✓). “OOD” is short for “out-of-distribution”.

Table 2. Perceptual study. Numbers denote the chance that participants prefer the reconstruction of a competing method over ECON for in-the-wild images. A value of 0.5 indicates equal preference. A value of $< 0.5$ favors ECON, while $> 0.5$ favors competitors.
We compare d-BiNi with BiNi using 600 samples (200 scans x 3 views) from CAPE and Renderpeople where ground-truth normal maps and meshes are available. Table 3 reports the “root mean squared error” (RMSE) and “mean absolute error” (MAE) between the estimated and rendered depth maps. d-BiNi significantly improves the reconstruction accuracy by about 50% compared to BiNi. This demonstrates the efficacy of using the coarse body mesh as regularization and taking the consistency of both the front and back surface into consideration. Additionally, d-BiNi is 33% faster than BiNi.

**IF-Nets+ vs IF-Nets.** Following the metrics of Sec. 4.2, we compare IF-Nets [10] with our IF-Nets+ on \( R_{\text{IF}} \). We show the quantitative comparison in Tab. 4. The improvement for out-of-distribution (“OOD”) poses shows that IF-Nets+ is more robust to pose variations than IF-Nets, as it is conditioned on the SMPL-X body. Figure 5 compares the geometry “inpainting” of both methods in the case of occlusions.

**4.5. Multi-person reconstruction**

Thanks to the shape completion module, ECON can deal with occlusions. Unlike other crowd body estimators [73–75, 86], ECON makes it possible to reconstruct multiple detailed “clothed” 3D humans from an image with inter-person occlusions, even though ECON has not been trained for this. Figure 10 shows three examples. The occluded parts, colored in red, are successfully recovered.

**5. Discussion**

**Limitations.** ECON takes as input an RGB image and an estimated SMPL-X body. However, recovering SMPL-X bodies (or similar models) from a single image is still an open problem, and not fully solved. Any failure in this could lead to ECON failures, such as in Fig. 8-A and Fig. 8-B. As the synthetic data [6, 28, 78] is getting sufficiently realistic, their domain gap with real data is significantly narrowed, it is predictable that such limitations will be eliminated. The reconstruction quality of ECON primarily relies on the accuracy of the predicted normal maps. Poor normal maps can result in overly close-by or even intersecting front and back surfaces, as shown in Fig. 8-C and Fig. 8-D.

**Future work.** Apart from addressing the above limitations, several other directions are useful for practical applications. Currently, ECON reconstructs only 3D geometry. One could additionally recover an underlying skeleton and skinning weights [45, 84], to obtain fully-animatable avatars. Moreover, generating back-view texture [8, 67, 68, 94] would result in fully-textured avatars. Disentangling clothing [1, 63, 99], hairstyle [89], or accessories [19] from the recovered geometry, would enable the simulation [23], synthesis, editing and transfer of styles [16] for these. ECON’s reconstructions, together with its underneath SMPL-X body, could be useful as 3D shape prior to learn neural avatars [13, 24, 32].

In particular, ECON could be used to augment existing datasets of 2D images with 3D humans. Datasets of real clothed humans with 3D ground truth [55, 60, 66, 88, 97] are limited in size. In contrast, datasets of images without 3D ground truth are widely available in large sizes [17, 21, 51]. As the synthetic data [6, 28, 78] is getting sufficiently realistic, we can “augment” such datasets by reconstructing detailed 3D humans from their images. In SupMat., we apply ECON on SHHQ [17] and recover normal maps and 3D humans; see Fig. 8-C. As ECON-like methods mature, they could produce pixel-aligned 3D humans from photos at scale, enabling the training of generative models of 3D clothed avatars with details [20, 22, 29, 34, 58, 81, 93].

**Possible negative impact.** As the reconstruction matures, it opens the potential for low-cost realistic avatar creation. Although such a technique benefits entertainment, film production, tele-presence and future metaverse applications, it could also facilitate deep-fake avatars. Regulations must be established to clarify the appropriate use of such technology.
Figure 9. **Qualitative results on in-the-wild images.** We show 8 examples of reconstructing detailed clothed 3D humans from images with: (a) challenging poses and (b) loose clothing. For each example we show the input image along with two views (front and rotated) of the reconstructed 3D humans. Our approach is robust to pose variations, generalizes well to loose clothing, and contains detailed geometry.

Figure 10. **Multiple humans with occlusions.** We detect multiple people and apply ECON to each separately. Although ECON is not trained on multiple people, it is robust to inter-person occlusions. We show three examples, and for each: (top) input image and the predicted front and back normal maps, (bottom) ECON’s reconstruction. Red areas on the estimated mesh indicate occlusions.

### 6. Conclusion

We propose ECON to reconstruct detailed clothed 3D humans from a color image. It combines estimated 2.5D front and back surfaces with an underlying 3D parametric body in a highly effective way. On the one hand, it is robust to novel poses, while on the other hand, it is capable of recovering loose clothing and geometric details, since the reconstructed shape is not over-constrained to the topology of the body. ECON achieves this by using and extending recent advances in variational normal integration [7] and shape completion [10]. It effectively extends these to the task of image-based 3D human reconstruction. We believe ECON can lead to both real-world applications and useful tools for the 3D vision community. The code and models are available at [econ.is.tue.mpg.de](http://econ.is.tue.mpg.de) for research purposes.

**Acknowledgments.** We thank Lea Hering and Radek Daněček for proofreading, Yao Feng, Haven Feng and Weiyang Liu for valuable feedback, and Tsvetelina Alexiadis for perceptual study. We are especially grateful to Carlos Baretto (Blender Add-on), Teddy Huang (Docker Image), and Justin John (Windows Support). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.860768 (CLIPE project).

**Disclosure.** [is.tue.mpg.de/black/CoI_CVPR_2023.txt](http://is.tue.mpg.de/black/CoI_CVPR_2023.txt)
References


[9] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards generative detailed neural avatars. In Computer Vision and Pattern Recognition (CVPR), 2022. 2


[26] Tong He, John P. Colombo, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions

[27] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-ready clothed human reconstruction revisited. In International Conference on Computer Vision (ICCV), 2021. 3


[33] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In European Conference on Computer Vision (ECCV), 2020. 3


[38] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. Transactions on Graphics (TOG), 2013. 2, 10, 12


[46] Ren Li, Benoît Guillard, Edoardo Remelli, and Pascal Fua. Dig: Draping implicit garment over the human body. In Asian Conference on Computer Vision (ACCV), 2022. 3


[48] 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. 9


[50] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xi-angyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In Computer Vision and Pattern Recognition (CVPR), 2023. 3


[66] RenderPeople. renderpeople.com, 2018. 2, 6, 7


[73] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In International Conference on Computer Vision (ICCV), 2021. 7

[74] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In Computer Vision and Pattern Recognition (CVPR), 2023. 7

[75] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In Computer Vision and Pattern Recognition (CVPR), 2022. 3, 7

[76] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D Human Pose Estimation via Intuitive Physics. In Computer Vision and Pattern Recognition (CVPR), 2023. 3


[81] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D Human Pose Estimation via Intuitive Physics. In Computer Vision and Pattern Recognition (CVPR), 2023. 7

[82] Yuliang Xu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In Computer Vision and Pattern Recognition (CVPR), 2022. 2, 3, 4, 6, 9, 10, 13, 14, 15
[87] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In Computer Vision and Pattern Recognition (CVPR), 2023. 3
[99] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In Computer Vision and Pattern Recognition (CVPR), 2022. 7