Sparse-View 3D Reconstruction of Clothed Humans via Normal Maps

Jane Wu[1] Diego Thomas[2] Ronald Fedkiw[1,3]
[1]Stanford University [2]Kyushu University [3]Epic Games
janehwu@stanford.edu, thomas@ait.kyushu-u.ac.jp, fedkiw@cs.stanford.edu

Abstract

We present a novel deep learning-based approach to the 3D reconstruction of clothed humans using weak supervision via 2D normal maps. Given a single RGB image or multiview images, our network is optimized to infer a person-specific signed distance function (SDF) discretized on a tetrahedral mesh surrounding the body in a rest pose. Subsequently, estimated human pose and camera parameters are used to generate a normal map from the SDF. A key aspect of our approach is the direct use of the Marching Tetrahedra algorithm in end-to-end optimization, and in order to do so we derive analytical gradients to facilitate straightforward differentiation (and thus backpropagation). Additionally, predicted normal maps allow us to leverage pretrained image-to-normal networks in order to minimize a surface error instead of a photometric error. We demonstrate the efficacy of our approach on both labeled and in-the-wild data in the context of existing clothed human reconstruction methods.

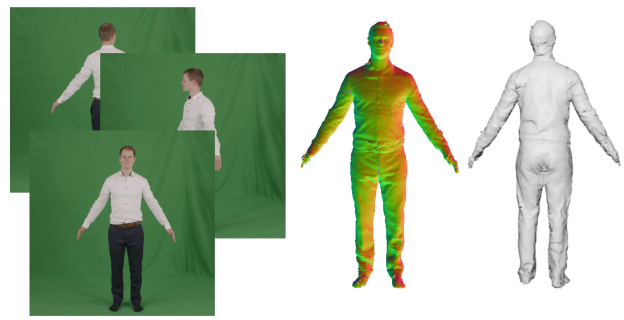

Figure 1. Given in-the-wild image(s) or video, our method is able to reconstruct clothed humans using inferred normal maps as the supervisory signal. Sample frames from the input video are shown to the left, and the predicted triangle mesh is shown to the right (the front-facing mesh is shaded with its normal map).

1. Introduction

Learning-based approaches for 3D human digitization have largely focused on the fully-supervised setting, where deep neural networks (DNNs) are trained to explicitly fit ground truth 3D geometry [66, 83, 84, 96]. In such approaches, high-end capture setups (with 4D scanners or a large number of cameras) are typically used to obtain high-quality, multiview training data [60, 82]. However, DNNs trained on scan data (typically captured in controlled environments) have not generalized well to in-the-wild scenarios with diverse people and lower-quality images. As a consequence, significant efforts have been made to reconstruct 3D geometry without relying on any 3D training data. Recently, Neural Radiance Fields (NeRFs) [47] and 3D Gaussian Splatting [32] approaches using dense multiview images have demonstrated impressive results in terms of reconstruction accuracy and novel-view synthesis [79,80,93]. However, these methods require on the order of hundreds of input images, which is largely inapplicable to in-the-wild scenarios where cost and compute budget are limited. Therefore, 3D human reconstruction from a few input RGB images is crucial to the democratization of digital humans required for many applications in AR/VR, robotics, etc.

In this paper, we investigate the problem of reconstructing 3D geometry of humans from a sparse set of input images (down to a single image). Instead of relying on (1) limited 3D human datasets or (2) dense multiview images, we propose using inferred 2D normal maps to guide network parameter optimization using one or more RGB images. By formulating our primary loss function with respect to image-based normals, we aim to better represent fundamental correlations between 2D images and 3D geometry in order to facilitate subsequent democratization to consumer-grade devices.

Given a sparse set of RGB images, we estimate the clothed body with a graph-based network that infers signed distance function (SDF) values for each vertex of a skinned tetrahedral mesh surrounding the body (see e.g. [36, 37, 54, 81]). Our network architecture builds on TetraTSDF [54], and we add a second explicit representation of the surface via a triangle mesh. We thus have access to two explicit versions of the neural SDF, and energies can be conveniently formulated for either the volume or the surface or both. In addition, the explicit surface can be efficiently rendered via image rasterization to obtain normal maps. See Figure 1.

Importantly, our proposed approach directly uses the Marching Tetrahedra algorithm [74] in end-to-end optimization. We demonstrate that the Marching Tetrahedra algorithm is straightforward to differentiate, and we derive analytical gradients to backpropagate through both Marching Tetrahedra and image rasterization. Our approach alleviates issues in the recently presented HAVE-FUN [87] and TeCH [24] papers, which use DMTet [69] as their tetrahedral representation. Section 2.2 and [46] explain the approximate/incorrect gradients that arise from such neural network meshing approaches [62, 69]. We also avoid concerns associated with the ray tracing of an implicit surface (see e.g. [9, 52, 76, 88]) by computing an explicit triangle mesh.

To summarize, our contributions:

1. A novel and efficient framework that can be used to reconstruct 3D geometry from a sparse set of multiview RGB data obtained with consumer-grade cameras.
2. Exact gradients to differentiate through Marching Tetrahedra via a Lagrangian formulation, which enables differentiable mesh generation and thus end-to-end optimization with both volumetric and surface-based energies.
3. Differentiable image rasterizer that: (a) allows us to use normal maps for optimization and (b) can efficiently compute normal maps from triangle meshes with over 300k triangles during network optimization.
4. Regularization energies that coerce inferred implicit surfaces to: (a) resemble true SDFs and (b) be locally smooth.
5. Silhouette energies enforcing 3D boundary matching.

2. Related Work

2.1. Human Shape Estimation

Various works use parametric body models such as SMPL [43] to estimate human body pose and shape without clothing [18, 34, 42, 50]. While existing methods are able to generalize to in-the-wild images, the inferred body mesh is often quite different from the underlying body shape and does not capture clothing.

In order to reconstruct humans wearing clothing from a single image [73], template-based approaches either rely on parametric models [10, 30, 35, 92, 98] or use person-specific meshes [21, 89]. For instance, GTA [92] projects SMPL onto a learned 3D triplane representation, and [98] constructs local implicit fields centered around locations on the SMPL-X model [58]. Limitations of template-based approaches to clothed human reconstruction include the output being constrained by the topology of the template as well as a reliance on accurate pose estimation. Template-free methods typically leverage 2D signals or 3D geometric representations to recover geometry. In [17], reconstruction is achieved by generating front and back depth images that are later combined into a 3D surface. [22] builds on this idea and proposes a coarse-to-fine reconstruction method leveraging both predicted depth and normal images. Inspired by shape-from-silhouette techniques, SiCloPe [51] recovers geometry by predicting silhouette images and 3D joint positions. [75, 97] predict volumetric occupancy on a uniform voxel grid directly, while [15] proposed learning a Fourier subspace of 3D occupancy; in both cases, Marching Cubes can be used to generate a triangle mesh. PIFu [65] and PIFuHD [66] infer 3D shape with neural implicit functions sampled onto a grid. Follow-up work [7, 84] leverages predicted normal maps to improve depth inference. PAMIR [96] extends PIFuHD to increase generalizability by regularizing the implicit function using semantic features from a parametric model. ICON [84] and ECON [83] leverage inferred front and back normal maps as an intermediate encoding of 3D geometry, but these methods still rely on ground truth 3D scan data during training.

Instead of a single input image, other works aim to construct animatable avatars from a sparse set of cameras [23, 67, 93, 99], video [12, 16, 20, 27–29, 56, 71, 79, 90], depth [13, 77, 85, 95], point clouds [33, 45], 4D capture [25], or scans [39, 41, 68]. Most similar to our work, SelfRecon [27] uses normal maps inferred from PIFuHD [66] to supervise network training, and SeSDF [5] can either take a single image or uncalibrated multiview images.

Moving towards improved generalization, weakly supervised methods have also been explored for human pose estimation [31, 53, 59, 86], human body shape [49, 91], and garment template reconstruction (on the SMPL body) [11, 49]. A number of methods also optimize network parameters to obtain person-specific 3D avatars [24, 26, 79, 80, 87, 93].

2.2. Differentiable Marching Cubes / Tetrahedra

A number of recent works have proposed methods for backpropagating through Marching Cubes [44] / Marching Tetrahedra [74] by either leveraging properties of a point-to-SDF network [62, 69] (e.g. as in DeepSDF [57]) or by training a DNN for mesh generation [40]. MeshSDF [62] builds on DeepSDF [57], where a network f_η is trained to infer an SDF value at a location x conditioned on a latent shape code η. In the forward step, $f_\eta(x)$ is computed for every vertex of a fixed voxel grid so that Marching Cubes can be used to generate a triangle mesh. The authors postulate that a small increase in the SDF values would move a triangle vertex in the normal direction, which is only true idealistically when

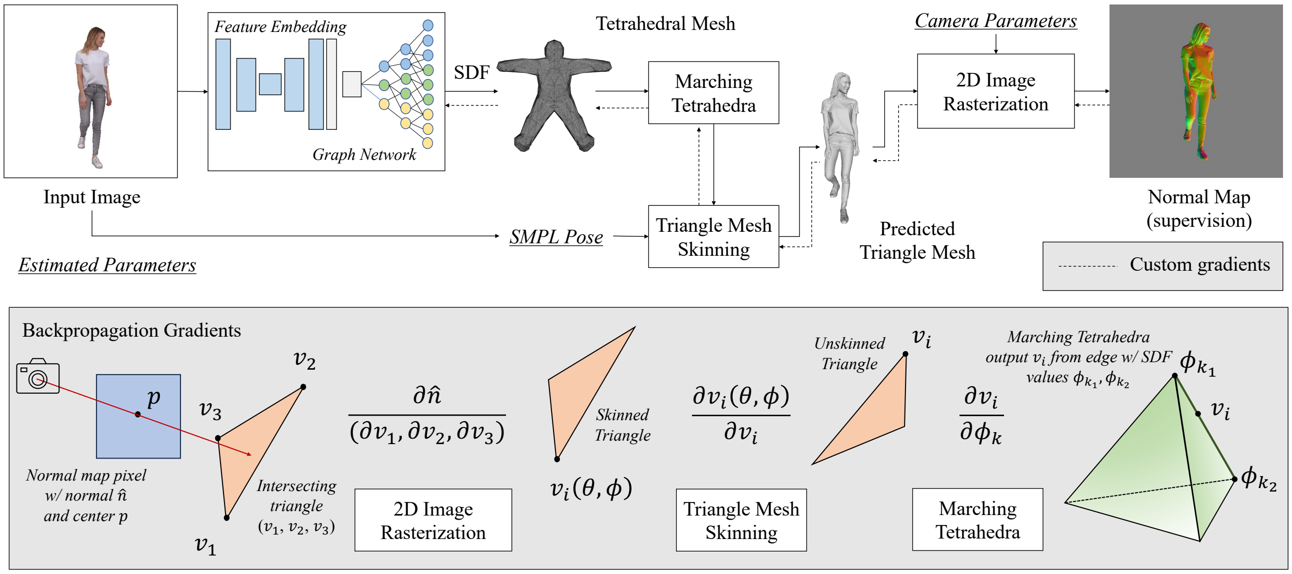

Figure 2. Given an RGB image, our network infers SDF values on a tetrahedral mesh parameterizing a volume surrounding/containing the body. Marching Tetrahedra is used to compute a triangle mesh from this implicit surface, and the resulting mesh is skinned to an estimated pose. A normal map (corresponding to estimated human pose and camera parameters) is generated for the predicted clothed body mesh and compared with ground truth. In order to enable end-to-end optimization, we derive analytical gradients to backpropagate from each pixel in the predicted normal map to the inferred SDF values.

there are no shocks/rarefactions in the SDF isocontours (see e.g. [55]); moreover, Marching Cubes does not move vertices in such a manner, even when it produces a consistent set of vertices under perturbation. Their assumptions also necessitate f_η being a true SDF, even though it is only an approximation. DMTet [69] presents a similar formulation using Marching Tetrahedra. See [46] for more discussion on the problematic assumptions in MeshSDF [62].

Additionally, while our explicit representation of an SDF is similar in spirit to the method used in [46], our method does not require the construction of a velocity in order to capture surface energies with an evolving level set function; thus, we can control mesh based properties (e.g. area and dihedral angles) that would be lost when converting to a velocity field. On the other hand, [46] could be used to alleviate locking concerns in cases where discretizations on the triangle mesh and the tetrahedral mesh do not interact as expected due to differences in the degrees of freedom.

3. Method Overview

We present an optimization-based approach to sparse-view clothed human reconstruction that uses 2D normal maps predicted from the input color images as supervision. Figure 2 shows an overview of our approach. Given input RGB image(s), our network infers SDF values $\hat{\phi}_k$ on a (fixed) set of tetrahedral mesh vertices u_k unskinned in the star pose. The network architecture is composed of a CNN-based stacked hourglass encoder, followed by graph convolutional layers that progressively increase the resolution of the sampled SDF. Due to the large number of tetrahedral mesh vertices, the graph convolutional layers are only partially connected in order to significantly reduce memory usage, as in TetraTSDF [54]. In order to ensure a nonzero number of triangle mesh faces are generated by Marching Tetrahedra (otherwise the predicted normal map is empty), we initialize the network parameters to output a template SMPL body shape [43].

Then, given the network-inferred SDF values $\hat{\phi}_k$, Marching Tetrahedra is used to uniquely generate a triangle mesh with vertices $v_i(\hat{\phi}_k)$. Given estimated human pose $\hat{\theta}$ and camera parameters $\hat{c}$ (e.g. using preprocessing code from PaMIR [96]), a normal map $N(v_i(\hat{\phi}_k), \hat{\theta}, \hat{c})$ can be generated for the skinned clothed human reconstruction. The objective function to be minimized is

$$\mathcal{L}(\hat{\phi}_k, \hat{\theta}, \hat{c}) = \left\| N(v_i(\hat{\phi}_k), \hat{\theta}, \hat{c}) - N_0 \right\| \tag{1}$$

where N_0 is a reference normal map inferred by the pretrained pix2pix network [78] from PIFuHD [66].

Section 4 defines the skinned tetrahedral mesh framework we use to parameterize 3D clothed human geometry. Section 5 details our implementation of Marching Tetrahedra with respect to this framework, and in particular the analytical gradients we derive for correctly backpropagating through the algorithm. Section 6 summarizes our image rasterization method, and Sections 7 and 8 define additional loss functions used during optimization.

4. Tetrahedral Mesh Framework

The 3D space surrounding and including the human body is parameterized via a tetrahedral mesh. First, a Cartesian grid based level set representation is generated for the SMPL template body [43] in the star pose (similar to [54, 81]). Then, a constant value is subtracted from the SDF values in order to inflate the zero level set so that its interior can contain a wide range of clothed body shapes. Subsequently, a tetrahedral mesh is generated for this interior region using red/green refinement [48, 72]. See Figure 3.

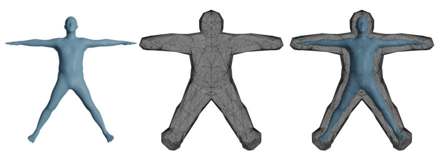

Figure 3. The tetrahedral mesh parameterizes a volume of air surrounding and including the body.

4.1. Skinning

The tetrahedral mesh can be deformed via linear blend skinning (LBS) using per-tetrahedron-vertex, per-joint skinning weights w_{kj}. The skinning weights are assigned in the star pose by first finding the point on the SMPL template body mesh closest to each tetrahedral mesh vertex, and then barycentrically interpolating skinning weights to that point from the vertices of the SMPL template body mesh triangle that contains it.

Given pose parameters θ with joint transformations $T_j(\theta)$, the skinned position of each tetrahedral mesh vertex is

$$u_k(\theta) = \sum_j w_{kj} T_j(\theta) u_k^j \tag{2}$$

where u_k^j is the location of u_k in the untransformed reference space of joint j.

5. Marching Tetrahedra

Given SDF values ϕ_k defined on tetrahedral mesh vertices u_k, Marching Tetrahedra can be implemented to compute a unique (non-ambiguous) triangle mesh with vertices v_i, making differentiation more straightforward as compared to the many cases (and non-uniqueness) that need to be considered for Marching Cubes. In order to avoid triangle vertices v_i coincident with a tetrahedron vertex u_k, all the ϕ_k values are preprocessed (infinitesimally) changing those with $|\phi_k| < \epsilon$ to $\phi_k = \epsilon \text{ sign}(\phi_k)$, e.g. with $\epsilon = 10^{-8}$.

For each tetrahedron mesh edge $e_i = \{u_{k_1}, u_{k_2}\}$ that includes a sign change, i.e. $\text{sign}(\phi_{k_1}) \neq \text{sign}(\phi_{k_2})$, a triangle vertex

$$v_i = \frac{-\phi_{k_2}}{\phi_{k_1} - \phi_{k_2}} u_{k_1} + \frac{\phi_{k_1}}{\phi_{k_1} - \phi_{k_2}} u_{k_2} \tag{3}$$

is defined using linear interpolation. Afterwards, triangles are constructed in a tetrahedron-by-tetrahedron manner by considering the two cases that can occur: either three edges of the tetrahedron contain triangle vertices and one triangle is constructed, or four edges contain triangle vertices and a quadrilateral is constructed and split into two triangles. Note that this typically arbitrary splitting of the quadrilateral can be made consistent for the sake of differentiation. Since the tetrahedral mesh does not change topology, the edges can be numbered in a fixed manner; then, one can consistently split a quadrilateral by connecting the triangle vertex on the lowest numbered edge to the triangle vertex on the highest numbered edge (or via a similar alternative strategy). The resulting triangle mesh is guaranteed to be watertight, and the vertices in each triangle are reordered (when necessary) to ensure that all face normals point outwards. A discussion about alternative implicit surface rendering approaches (e.g. [2, 9, 52, 76, 88]) is in the supplementary material.

5.1. Computing Gradients

According to Equation 3,

$$\frac{\partial v_i}{\partial(\phi_{k_1}, \phi_{k_2})} = \begin{bmatrix} \frac{\phi_{k_2}(u_{k_1} - u_{k_2})}{(\phi_{k_1} - \phi_{k_2})^2} & \frac{\phi_{k_1}(u_{k_2} - u_{k_1})}{(\phi_{k_1} - \phi_{k_2})^2} \end{bmatrix} \tag{4}$$

where dividing by $(\phi_{k_1} - \phi_{k_2})^2$ can be problematic. The preprocess at the beginning of Section 5 guarantees that $|\phi_{k_1} - \phi_{k_2}| \geq 2\epsilon$, which means that the worst possible scenario for Equation 3 (when $|\phi_{k_1}| = |\phi_{k_2}| = \epsilon$) still results in $\mathcal{O}(1)$ coefficients for u_{k_1} and u_{k_2}; however, the ϕ-based coefficients in Equation 4 would be $\mathcal{O}(1/\epsilon)$. Thus, while $\epsilon = 10^{-8}$ is sufficient for Equation 3, a larger value of ϵ might be prudent when considering Equation 4.

6. Image Rasterization

Given a skinned triangulated surface and parameters for a perspective camera model, a camera space normal map is computed using a right-handed coordinate system. We assume that the geometry is centered in the image, since images are cropped and rescaled during preprocessing. Normal maps made using different assumptions, or decoded and stored as RGB values, are readily transformed back into unit normals (in camera space) in order to match our assumptions. The supplementary material contains more details on how normal maps are computed and differentiated.

7. SDF Regularization

Two regularization terms are utilized during neural network training in order to encourage: (1) the inferred $\hat{\phi}$ values to resemble a true SDF and (2) smoothness (similar to [38, 64]). Notably, the smoothness regularizer behaves significantly better when $\hat{\phi}$ is closer to a true SDF.

7.1. Eikonal Regularization

Given a tetrahedron t with vertices $u_k = (x_k, y_k, z_k)$ and inferred $\hat{\phi}_k$ values, $\hat{\phi}$ can be linearly approximated within the tetrahedron by writing

$$\hat{\phi}_k = ax_k + by_k + cz_k + d \tag{5}$$

for each of the four vertices; then, the resulting 4×4 linear system of equations can be solved to obtain the unknown coefficients (a, b, c, d) leading to

$$|\nabla \hat{\phi}_t| = \sqrt{a^2 + b^2 + c^2} \tag{6}$$

as the norm of the gradient. Summing over tetrahedra leads to

$$E_{1a} = \frac{1}{2} \sum_t (|\nabla \hat{\phi}_t| - 1)^2 \tag{7}$$

as the energy to be minimized. The problem with Equation 7 (and similar approaches, such as [4, 19]) is that the chain rule moves the square root in Equation 6 to the denominator, potentially leading to NaNs/overflow; notably, even an exact SDF has $|\nabla \phi| = 0$ at both extrema and pinching/merging saddles, and an inferred $\hat{\phi}$ can have $|\nabla \hat{\phi}| = 0$ elsewhere as well. This can be avoided by instead using

$$E_{1b} = \frac{1}{2} \sum_t (|\nabla \hat{\phi}_t|^2 - 1)^2 \tag{8}$$

which still enforces $|\nabla \hat{\phi}_t| = 1$; alternatively,

$$E_{1c} = \frac{1}{2} \sum_t \text{Volume}(t)(|\nabla \hat{\phi}_t|^2 - 1)^2 \tag{9}$$

scales the penalty on each tetrahedron by its volume.

7.2. Motion by Mean Curvature

In order to encourage smoothness, we define an energy that when minimized results in motion by mean curvature. Following [8, 94], the surface area can be calculated via

$$\int_\Omega |\nabla H(\phi(x, y, z))| dV \tag{10}$$

where H is a Heaviside function and V is the volume; thus, on our tetrahedral mesh, we minimize

$$E_2 = \sum_t |\nabla H(\hat{\phi})| \text{Volume}(t) \tag{11}$$

using a smeared-out Heaviside Function

$$H(\hat{\phi}) = \begin{cases} 0 & \hat{\phi} < -\epsilon_H \\ \frac{1}{2} + \frac{\hat{\phi}}{2\epsilon_H} + \frac{1}{2\pi} \sin\left(\frac{\pi \hat{\phi}}{\epsilon_H}\right) & -\epsilon_H \leq \hat{\phi} \leq \epsilon_H \\ 1 & \hat{\phi} > \epsilon_H \end{cases} \tag{12}$$

where ϵ_H, chosen as 1.5 times the average tetrahedral mesh edge length, determines the bandwidth of numerical smearing (see [70]). $|\nabla H(\hat{\phi})|$ is discretized by linearly approximating $H(\hat{\phi})$ in each tetrahedron along the lines of Equation 5 in order to obtain coefficients (a, b, c, d) for use in the equivalent of Equation 6. In order to avoid division by small numbers, we ignore tetrahedra with $|\nabla H(\hat{\phi})| < 10^{-8}$ in Equation 11 reasoning that $|\nabla H(\hat{\phi})|$ is small enough and thus $\hat{\phi}$ is smooth enough in such tetrahedra.

8. Silhouette Losses

Instead of striving to make the inverse rendering differentiable at silhouette boundaries (as in e.g. [2]), we introduce energies that force the silhouettes to match.

8.1. Shrinking

For pixels that overlap the inferred surface but not the ground truth surface, the interior of the inferred surface needs to shrink so that the corresponding triangles disappear. For each tetrahedron mesh edge containing a vertex of a problematic triangle, the edge's parent tetrahedral mesh vertices are added to the set U_{shrink} if they have negative SDF values; then,

$$\mathcal{L}_{\text{shrink}} = \frac{1}{2} \sum_{k \in U_{\text{shrink}}} (\hat{\phi}_k - \epsilon_s)^2 \tag{13}$$

encourages those negative $\hat{\phi}_k$ values to target a positive $\epsilon_s = 5 \times 10^{-3}$, which is chosen as half the average tetrahedral mesh edge length.

8.2. Expanding

For pixels that overlap the ground truth surface but not the inferred surface, the interior of the inferred surface needs to expand. In order to determine where this expansion should occur, the implicit surface is temporarily inflated by changing the sign of the SDF at every tetrahedral mesh vertex with both $\hat{\phi} > 0$ and a one-ring neighbor with $\hat{\phi} < 0$ (e.g. by setting $\hat{\phi}_{temp} = -\epsilon_s$ at those vertices). Next, the pixels that previously overlapped the ground truth surface but not the inferred surface and now overlap both the ground truth surface and the new inflated surface are identified. For each tetrahedron mesh edge containing a vertex of a triangle corresponding to one of these pixels, the edge's parent tetrahedral mesh vertices are added to the set U_{expand} if they had positive SDF values before inflation. At this point, all of the temporary $\hat{\phi}_{temp}$ values are discarded and the original $\hat{\phi}$ values are restored. Then,

$$\mathcal{L}_{\text{expand}} = \frac{1}{2} \sum_{k \in U_{\text{expand}}} (\hat{\phi}_k + \epsilon_s)^2 \tag{14}$$

encourages the positive $\hat{\phi}_k$ values to target $-\epsilon_s$.

9. Experiments

We first define the metrics and datasets used to evaluate 3D reconstruction in Sections 9.1 and 9.2, respectively. In Section 9.3, we compare our results to available implementations of existing methods for single view and multiview reconstruction on both labeled and in-the-wild datasets. In Section 9.4, we extend our proposed approach to real-world RGB data (with uncalibrated cameras and no ground truth information) in order to demonstrate the ability to reconstruct 3D geometry using only network-inferred normal maps. Finally, in Section 9.5, ablation studies demonstrate that our network can be trained to reconstruct 3D geometry with increasing efficacy as the number of sparse views increases.

9.1. Metrics

For data with 3D labels, we use the same 3D metrics as in prior work [20] for evaluation: volumetric IoU, Chamfer distance (cm), and normal consistency. For all three metrics, we use the publicly available Pytorch3D implementations [61]. For data without 3D labels, we define a normal map error as

$$e_{normal} = \frac{1}{W \times H} \sum_{p} \left(\frac{1}{2} (1 - \hat{n}_p \cdot n_p) \right)^2 \tag{15}$$

where the ground truth and predicted normals at pixel p are n_p and $\hat{n}_p$, respectively, and $\hat{n}_p \cdot n_p \in [-1, 1]$ is replaced with -1 for pixels where the predicted and ground truth silhouettes do not overlap. Note that normal maps do not uniquely determine scale/depth; thus, the reconstructed objects could erroneously move closer/further from the camera becoming smaller/larger in scale (while also undergoing distortion, since this scale variance is not self-similar). In order to monitor this, we define a depth map error as

$$e_{depth} = \frac{1}{W \times H} \sum_{p} \left(\hat{d}_p - d_p \right)^2 \tag{16}$$

where $(\hat{d}_p - d_p)$ is replaced with the thickness of the tetrahedral mesh (0.2 meters) for pixels where the predicted and ground truth silhouettes do not overlap.

9.2. Datasets

SynWild: The SynWild dataset [20] was generated by placing captured 4D scans of dynamic human subjects into realistic 3D scenes/HDRI panoramas. Because the scene is rendered from virtual cameras, accurate 3D ground truth is available in a realistic setting. The dataset contains five videos.

People Snapshot: The People Snapshot Dataset [1] contains 24 in-the-wild video sequences of 11 subjects. Each video was captured with a fixed camera, and the subjects were asked to rotate while holding an A-pose.

Method	IoU ↑	CD (cm) ↓	NC ↑
ICON [84]	0.764	2.91	0.766
SelfRecon [27]	0.805	2.50	0.776
Vid2Avatar [20]	0.813	2.35	0.796
Our Method	**0.923**	**1.53**	**0.889**

Table 1. **Quantitative evaluation on SynWild.** 3D metrics averaged over all SynWild examples. For each method, we compute IoU, Chamfer Distance (CD), and normals consistency.

Method	Normals error ↓	SD
ECON [83]	0.1918	0.376
PIFuHD [66]	0.1680	0.358
VideoAvatar [1]	0.1342	0.325
Vid2Avatar [20]	0.0345	0.179
SelfRecon [27]	0.0213	0.137
Our Method	**0.0207**	**0.111**

Table 2. **Quantitative evaluation on People Snapshot.** Normal map errors (computed over predicted foreground pixels) and standard deviations (SD) averaged over four test views for each example shown in Figure 4.

9.3. Comparisons

We quantitatively compare our reconstruction method to existing single-view and multiview reconstruction approaches using the two aforementioned datasets: SynWild [20] (3D labels) and People Snapshot [1] (no 3D labels).

Table 1 compares our method to existing work on the SynWild dataset, using the same 3D metrics reported in Vid2Avatar [20]. As usual, our method uses PIFuHD inferred normal maps [66] as training supervision. Our method achieves state-of-the-art performance on all three metrics.

Table 2 compares our method to official implementations of existing approaches using the People Snapshot dataset, and we compute the normal map error metric defined in Section 9.1. Our method outperforms all existing methods, despite only using four frames per video (front, back, and two side views), whereas Vid2Avatar [20], SelfRecon [27], VideoAvatar [1] were trained on all video frames. For the single view approaches [66,83], we took the mesh predicted using the front or back-facing frame (whichever is closer to the test view) and scaled/rigidly aligned it using ICP to fit the corresponding SelfRecon mesh. Note that for People Snapshot, we consider "ground truth" as the PIFuHD inferred normal map because there are no 3D labels. Figure 4 shows the corresponding qualitative results using PeopleSnapshot. Vid2Avatar and SelfRecon produce significantly less detail compared to our approach (particularly around the face and wrinkles in the clothing).

Notably, the runtime of our approach on a single

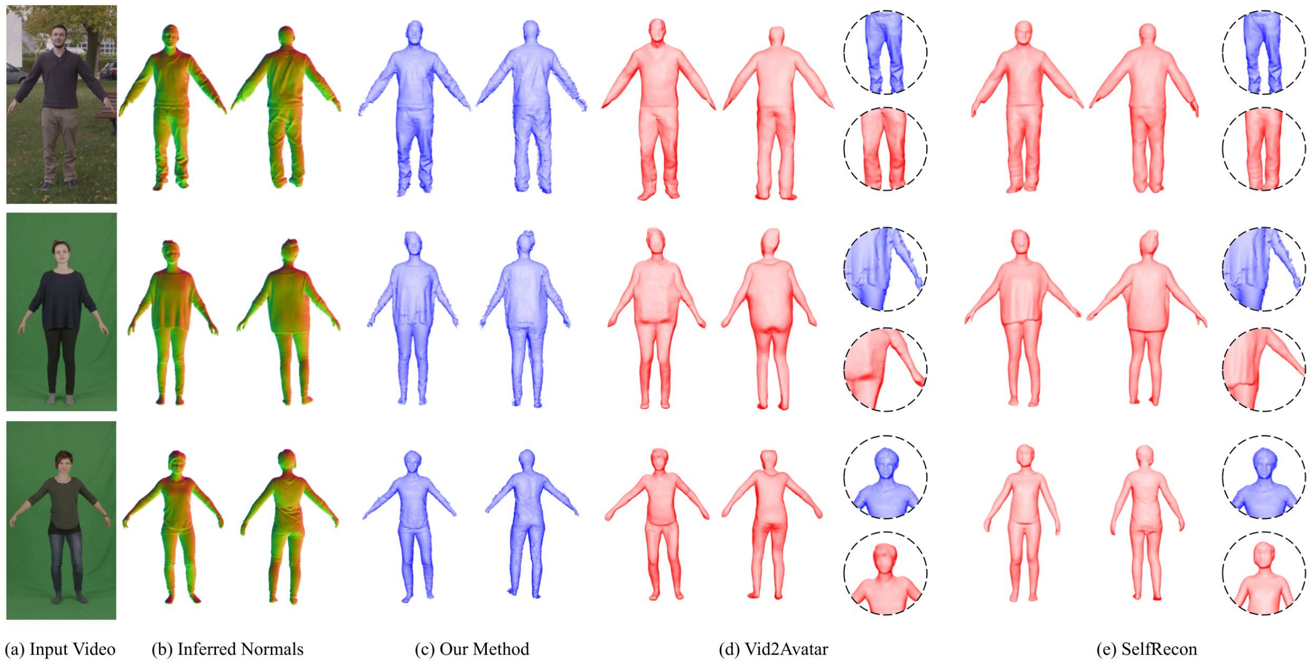

Figure 4. Predicted 3D geometry using our method (c) for videos from the PeopleSnapshot dataset (one frame is shown in (a)). The results from Vid2Avatar [20] are shown in (d), and the results from SelfRecon [27] are shown in (e). Note that the geometry is shown from novel views.

NVIDIA 3090 GPU is at least 50× faster than SelfRecon, which takes over a day of training (per video) to achieve their published results (our network is optimized for about 20 minutes).

9.4. In-the-wild Reconstruction from RGB Images

We further illustrate that our approach can be used to reconstruct 3D geometry from a sparse set of uncalibrated RGB images, without requiring any pretraining on scanned data (or any other informed initialization of the network parameters). As usual, we utilize PIFuHD [66] to infer ground truth normal maps and note that pix2pix was trained on 3D ground truth geometry.

First, we captured monocular video footage of a person in a static pose; then, a sparse number of frames were extracted and preprocessed by removing the background using [6] and cropping to a square image. See Figure 5, row 1. The resulting images were then passed into pix2pix to obtain "ground truth" normal maps. Since estimated camera parameters will be prone to error, we refine a rough initialization iteratively. At each iteration, we optimize the network and use Marching Tetrahedra to create a mesh inferenced off of the image for (and overfit to) each view; then, we use ICP [3] to rigidly align all the meshes. Although one could delete all the triangles and remesh the point cloud, we obtained better results by updating each camera to match the ICP rigid transform of its corresponding mesh. The updated camera positions are then used to iteratively repeat the entire process. Once the camera parameters converge, the network can be optimized with an additional loss that encourages 3D consistency. For a given camera view c_0, this loss is defined as

$$\mathcal{L}(\hat{\phi}_k) = \sum_{c \neq c_0} \left\| \hat{\phi}_k - \hat{\phi}_k(c) \right\| \tag{17}$$

where $\hat{\phi}_k(c)$ refers to the inferred SDF values obtained from using view c's image.

The network obtained from the aforementioned process (to improve camera extrinsics) will tend to be less detailed on the back side of the mesh, since only the front side can be seen in any given input image; thus, after improving camera extrinsics, we proceed as follows. Each view is fine-tuned with a regularizer that aims to keep ϕ close to that which was obtained using Equation 17; then, we delete any visible triangles that are not consistent with the normal map (within some tolerance). Since these are (actual) triangle meshes, it is trivial to load them into a suitable computer graphics application and align/resize the meshes in order to combine them into a single unified mesh. See Figure 5, row 2.

9.5. Ablation Study

Given ground truth 3D data from [63], we show how our network reconstructs 3D geometry with increasing efficacy as the number of sparse views increases. Figure 6 shows

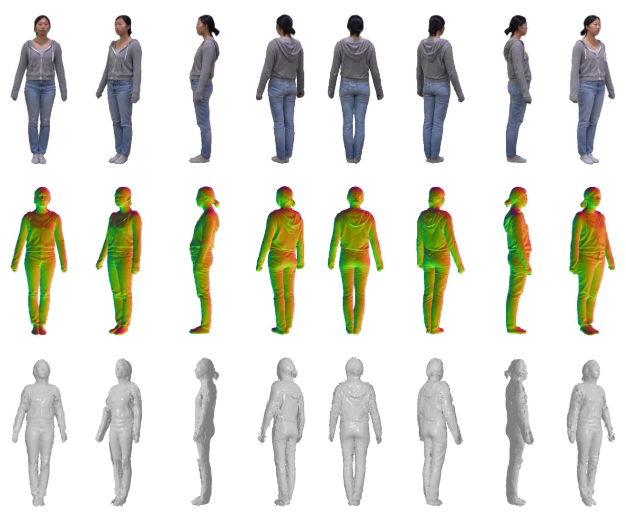

Figure 5. RGB images, inferred normal maps, and the final triangle mesh shown from training (odd) and novel (even) views.

# Views	Normals Error ↓	Depth Error (m) ↓
1	0.0317	0.0012
3	0.0299	0.0011
5	0.0060	0.0002

Table 3. Quantitative metrics for three unseen test views spaced between the training views.

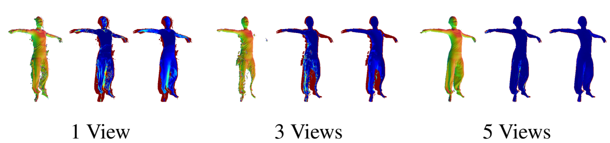

Figure 6. Models trained on an increasing number of camera views (inference from a novel view is shown). Each set of images show: (left) predicted triangle mesh and its normal map, (middle) normal errors (blue is zero, red is max), and (right) depth errors. See also Figure 7.

the inferred 3D geometry from a novel view, and Table 3 shows how per-pixel normal and depth errors decrease as the number of training views increases. When the network is trained on only one view, there are no constraints on the side/back of the person; hence, the predicted geometry has a high degree of noise when rendered from novel views. When trained with 5 views, the ground truth geometry is recovered with high accuracy.

10. Conclusion

Although image-based reconstruction can be solved as an inverse problem, regularization is required in order to address issues with noise. Parameterized models (such as SMPL [43] or 3DMM [14]) provide for such regularization.

Figure 7. Using the 5-view network from Figure 6, we illustrate inference for four more novel views.

We choose a neural network to parameterize our reconstruction where regularization is provided by having a limited number of network parameters. Our network aims to convert images from any view direction into a unique implicit surface, regardless of the view direction (similar in spirit to how the human brain process visual input); in fact, our eyes discern relative distance (similar to normal maps) more proficiently than they discern raw distance.

In summary, we present a weakly-supervised method for clothed human reconstruction by leveraging 2D normal maps as the supervisory signal during neural network training. In order to train a learned model that can infer high-frequency cloth and body geometry without any ground truth 3D data, our proposed approach builds on strong geometric priors for modeling and rendering. Our results reinforce the notion that less training data is required to train networks that infer normal maps than to train networks that infer 3D geometry (in agreement with ECON [83]). This means that working to improve the efficacy of network-inferred normal maps (and using the results for 3D reconstruction, as in Section 9.4) is likely to be more productive than working to obtain (via expensive 3D scanning) the excessive amount of ground truth data required to train a network to inference 3D geometry directly. Moreover, the process outlined in Section 9.4 provides an alternative mechanism (significantly cheaper than 3D scanning) for acquiring the ground truth data required to train a network to inference 3D geometry directly.

Acknowledgements

Research supported in part by ONR N00014-19-1-2285, ONR N00014-21-1-2771. We would like to thank Reza and Behzad at ONR for supporting our efforts into machine learning. This work was also supported by JSPS KAKENHI Grant Number JP23H03439 and AMED JP24wm06254. J. W. was supported in part by the Gerald J. Lieberman Graduate Fellowship, the NSF Mathematical Sciences Postdoctoral Fellowship, and the UC President's Postdoctoral Fellowship.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 6

[2] Sai Praveen Bangaru, Michaël Gharbi, Fujun Luan, Tzu-Mao Li, Kalyan Sunkavalli, Milos Hasan, Sai Bi, Zexiang Xu, Gilbert Bernstein, and Fredo Durand. Differentiable rendering of neural sdfs through reparameterization. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 4, 5

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 7

[4] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. *arXiv preprint arXiv:2206.15258*, 2022. 5

[5] Yukang Cao, Kai Han, and Kwan-Yee K Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4647–4657, 2023. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 7

[7] Kennard Yanting Chan, Guosheng Lin, Haiyu Zhao, and Weisi Lin. Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 328–344. Springer, 2022. 2

[8] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001. 5

[9] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20427–20437, 2022. 2, 4

[10] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. *arXiv preprint arXiv:2212.06820*, 2022. 2

[11] Luca De Luigi, Ren Li, Benoît Guillard, Mathieu Salzmann, and Pascal Fua. Drapenet: Garment generation and self-supervised draping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1451–1460, 2023. 2

[12] Junting Dong, Qi Fang, Yudong Guo, Sida Peng, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. *Advances in Neural Information Processing Systems*, 35:13654–13667, 2022. 2

[13] Zheng Dong, Ke Xu, Ziheng Duan, Hujun Bao, Weiwei Xu, and Rynson Lau. Geometry-aware two-scale pifu representation for human reconstruction. *Advances in Neural Information Processing Systems*, 35:31130–31144, 2022. 2

[14] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models: past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 8

[15] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: learning fourier occupancy field for monocular real-time human reconstruction. *Advances in Neural Information Processing Systems*, 35:7397–7409, 2022. 2

[16] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. *arXiv preprint arXiv:2210.01868*, 2022. 2

[17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2232–2241, 2019. 2

[18] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[19] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[20] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2, 6, 7

[21] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2

[22] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12869–12879, 2023. 2

[23] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 535–545, 2021. 2

[24] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. *arXiv preprint arXiv:2308.08545*, 2023. 2

[25] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for

humans in motion. *arXiv preprint arXiv:2305.06356*, 2023. 2

[26] Vinoj Jayasundara, Amit Agrawal, Nicolas Heron, Abhinav Shrivastava, and Larry S Davis. Flexnerf: Photorealistic free-viewpoint rendering of moving humans from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21118–21127, 2023. 2

[27] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2, 6, 7

[28] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 2

[29] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022. 2

[30] Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and PJ Narayanan. Sharp: Shape-aware reconstruction of people in loose clothing. *International Journal of Computer Vision*, 131(4):918–937, 2023. 2

[31] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1

[33] Hyomin Kim, Hyeonseo Nam, Jungeon Kim, Jaesik Park, and Seungyong Lee. Laplacianfusion: Detailed 3d clothed-human body reconstruction. *ACM Transactions on Graphics (TOG)*, 41(6):1–14, 2022. 2

[34] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is matter: Point-guided 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12880–12889, 2023. 2

[35] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021. 2

[36] Minjae Lee, David Hyde, Michael Bao, and Ronald Fedkiw. A skinned tetrahedral mesh for hair animation and hair-water interaction. *IEEE transactions on visualization and computer graphics*, 25(3):1449–1459, 2018. 1

[37] Minjae Lee, David Hyde, Kevin Li, and Ronald Fedkiw. A robust volume conserving method for character-water interaction. In *Proceedings of the 18th annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–12, 2019. 1

[38] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 5

[39] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 322–341. Springer, 2022. 2

[40] Yiyi Liao, Simon Donne, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2916–2925, 2018. 2

[41] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 210–228. Springer, 2022. 2

[42] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective. *ACM Computing Surveys*, 55(4):1–41, 2022. 2

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3, 4, 8

[44] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2

[45] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10974–10984, 2021. 2

[46] Ishit Mehta, Manmohan Chandraker, and Ravi Ramamoorthi. A level set theory for neural implicit evolution under explicit flows. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 711–729. Springer, 2022. 2, 3

[47] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1

[48] Neil Molino, Robert Bridson, Joseph Teran, and Ronald Fedkiw. A crystalline, red green strategy for meshing highly deformable objects with tetrahedra. In *IMR*, pages 103–114. Citeseer, 2003. 4

[49] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 184–200. Springer, 2022. 2

[50] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular

video for 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14829–14839, 2023. 2

[51] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019. 2

[52] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2, 4

[53] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[54] Hayato Onizuka, Zehra Hayirci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-ichiro Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6011–6020, 2020. 1, 3, 4

[55] Stanley Osher, Ronald Fedkiw, and K Piechor. Level set methods and dynamic implicit surfaces. *Appl. Mech. Rev.*, 57(3):B15–B15, 2004. 3

[56] Anqi Pang, Xin Chen, Haimin Luo, Minye Wu, Jingyi Yu, and Lan Xu. Few-shot neural human performance rendering from sparse rgbd videos. *arXiv preprint arXiv:2107.06505*, 2021. 2

[57] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[58] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[59] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

[60] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 1

[61] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6

[62] Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoît Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Meshsdf: Differentiable iso-surface extraction. *Advances in Neural Information Processing Systems*, 33:22468–22478, 2020. 2, 3

[63] Renderpeople, 2018. 7

[64] Radu Alexandru Rosu and Sven Behnke. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2023. 5

[65] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[66] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1, 2, 3, 6, 7

[67] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 702–720. Springer, 2022. 2

[68] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023. 2

[69] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3

[70] Mark Sussman, Peter Smereka, and Stanley Osher. A level set approach for computing solutions to incompressible two-phase flow. *Journal of Computational physics*, 114(1):146–159, 1994. 5

[71] Gusi Te, Xiu Li, Xiao Li, Jinglu Wang, Wei Hu, and Yan Lu. Neural capture of animatable 3d human from monocular video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 275–291. Springer, 2022. 2

[72] Joseph Teran, Neil Molino, Ronald Fedkiw, and Robert Bridson. Adaptive physics based tetrahedral mesh generation using level sets. *Engineering with computers*, 21(1):2–18, 2005. 4

[73] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[74] Graham M Treece, Richard W Prager, and Andrew H Gee. Regularised marching tetrahedra: improved iso-surface extraction. *Computers & Graphics*, 23(4):583–598, 1999. 2

[75] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings*

of the European Conference on Computer Vision (ECCV), pages 20–36, 2018. 2

[76] Delio Vicini, Sébastien Speierer, and Wenzel Jakob. Differentiable signed distance function rendering. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022. 2, 4

[77] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems*, 34:2810–2822, 2021. 2

[78] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[79] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 1, 2

[80] Chung-Yi Weng, Pratul P Srinivasan, Brian Curless, and Ira Kemelmacher-Shlizerman. Personnerf: Personalized reconstruction from photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 524–533, 2023. 1, 2

[81] Jane Wu, Zhenglin Geng, Hui Zhou, and Ronald Fedkiw. Skinning a parameterization of three-dimensional space for neural network cloth. *arXiv preprint arXiv:2006.04874*, 2020. 1, 4

[82] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 1

[83] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 1, 2, 6, 8

[84] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13296–13306, 2022. 1, 2, 6

[85] Yuxuan Xue, Bharat Lal Bhatnagar, Riccardo Marin, Nikolaos Sarafianos, Yuanlu Xu, Gerard Pons-Moll, and Tony Tung. Nsf: Neural surface fields for human modeling from monocular depth. In *ICCV*, October 2023. 2

[86] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018. 2

[87] Xihe Yang, Xingyu Chen, Shaohui Wang, Daiheng Gao, Xiaoguang Han, and Baoyuan Wang. Have-fun: Human avatar reconstruction from few-shot unconstrained images. *arXiv preprint arXiv:2311.15672*, 2023. 2

[88] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2, 4

[89] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[90] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[91] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, pages 465–481. Springer, 2020. 2

[92] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *arXiv preprint arXiv:2309.13524*, 2023. 2

[93] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 1, 2

[94] Hong-Kai Zhao, Tony Chan, Barry Merriman, and Stanley Osher. A variational level set approach to multiphase motion. *Journal of computational physics*, 127(1):179–195, 1996. 5

[95] Ruichen Zheng, Peng Li, Haoqian Wang, and Tao Yu. Learning visibility field for detailed 3d human reconstruction and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 216–226, 2023. 2

[96] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1, 2, 3

[97] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 2

[98] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *arXiv preprint arXiv:2305.04789*, 2023. 2

[99] Tiansong Zhou, Tao Yu, Ruizhi Shao, and Kun Li. Hdhuman: High-quality human performance capture with sparse views. *CoRR*, 2022. 2