Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition

Chen Guo[1] Tianjian Jiang[1] Xu Chen[1,2] Jie Song[1] Otmar Hilliges[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen

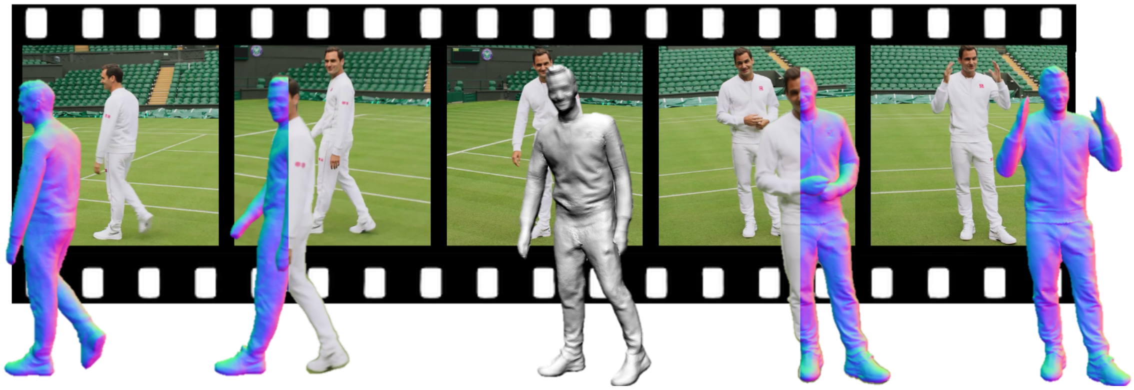

Figure 1. We propose Vid2Avatar, a method to reconstruct detailed 3D avatars from monocular videos in the wild via self-supervised scene decomposition. Our method does not require any groundtruth supervision or priors extracted from large datasets of clothed human scans, nor do we rely on any external segmentation modules.

Abstract

We present Vid2Avatar, a method to learn human avatars from monocular in-the-wild videos. Reconstructing humans that move naturally from monocular in-the-wild videos is difficult. Solving it requires accurately separating humans from arbitrary backgrounds. Moreover, it requires reconstructing detailed 3D surface from short video sequences, making it even more challenging. Despite these challenges, our method does not require any groundtruth supervision or priors extracted from large datasets of clothed human scans, nor do we rely on any external segmentation modules. Instead, it solves the tasks of scene decomposition and surface reconstruction directly in 3D by modeling both the human and the background in the scene jointly, parameterized via two separate neural fields. Specifically, we define a temporally consistent human representation in canonical space and formulate a global optimization over the background model, the canonical human shape and texture, and per-frame human pose parameters. A coarse-to-fine sampling strategy for volume rendering and novel objectives are introduced for a clean separation of dynamic human and static background, yielding detailed and robust 3D human geometry reconstructions. We evaluate our methods on publicly available datasets and show improvements over prior art. Project page: https://moygcc.github.io/vid2avatar/.

1. Introduction

Being able to reconstruct detailed avatars from readily available "in-the-wild" videos, for example recorded with a phone, would enable many applications in AR/VR, in human-computer interaction, robotics and in the movie and sports industry. Traditionally, high-fidelity 3D reconstruction of dynamic humans has required calibrated multi-view systems [9, 10, 19, 27, 31, 45, 49], which are expensive and require highly-specialized expertise to operate. In contrast, emerging applications such as the Metaverse require more light-weight and practical solutions in order to make the digitization of humans a widely available technology. Reconstructing humans that move naturally from monocular in-the-wild videos is clearly a difficult problem. Solving it requires accurately separating humans from arbitrary backgrounds, without any prior knowledge about the scene or the subject. Moreover it requires reconstructing detailed 3D surface from short video sequences, made even more challenging due to depth ambiguities, the complex dynamics of human motion and the high-frequency surface details.

Traditional template-based approaches [15, 16, 58] cannot generalize to in-the-wild settings due to the requirement for a pre-scanned template and manual rigging. Methods that are based on explicit mesh representations are limited to a fixed topology and resolution [3, 8, 14, 34]. Fully-supervised methods that directly regress 3D surfaces from images [17, 18, 21, 41, 42, 55, 68] struggle with difficult out-of-distribution poses and shapes, and do not always predict temporally consistent reconstructions. Fitting neural implicit surfaces to videos has recently been demonstrated [23, 24, 40, 46, 47, 52, 67]. However, these methods depend on pre-segmented inputs and are therefore not robust to uncontrolled visual complexity and are upper-bounded in their reconstruction quality by the segmentation method.

In this paper, we introduce Vid2Avatar, a method to learn human avatars from monocular in-the-wild videos without requiring any groundtruth supervision or priors extracted from large datasets of clothed human scans, nor do we rely on any external segmentation modules. We solve the tasks of scene separation and surface reconstruction directly in 3D. To achieve this, we model both the foreground (i.e., human) and the background in the scene implicitly, parameterized via two separate neural fields. A key challenge is to associate 3D points to either of these fields without reverting to 2D segmentation. To tackle this challenge, our method builds-upon the following core concepts: i) We define a single temporally consistent representation of the human shape and texture in canonical space and leverage the inverse mapping of a parametric body model to learn from deformed observations. ii) A global optimization formulation jointly optimizes the parameters of the background model, the canonical human shape and its appearance, and the pose estimates of the human subject over the entire sequence. iii) A coarse-to-fine sampling strategy for volume rendering that naturally leads to a separation of dynamic foreground and static background. iv) Novel objectives that further improve the scene decomposition and lead to sharp boundaries between the human and the background, even when both are in contact (e.g., around the feet), yielding better geometry and appearance reconstructions.

More specifically, we leverage an inverse-depth parameterization in spherical coordinates [66] to coarsely separate the static background from the dynamic foreground. Within the foreground sphere, we leverage a surface-guided volume rendering approach to attain densities via the conversion method proposed in [60]. Importantly, we warp all sampled points into canonical space and update the human shape field dynamically. To attain sharp boundaries between the dynamic foreground and the scene, we introduce two optimization objectives that encourage a quasi-discrete binary distribution of ray opacities and penalize non-zero opacity for rays that do not intersect with the human. The final rendering of the scene is then attained by differentiable composited volume rendering.

We show that this optimization formulation leads to clean scene decomposition and high-quality 3D reconstructions of the human subject. In detailed ablations, we shed light on the key components of our method. Furthermore, we compare to existing methods in 2D segmentation, novel view synthesis, and reconstruction tasks, showing that our method performs best across several datasets and settings. To allow for quantitative comparison across methods, we contribute a novel semi-synthetic test set that contains accurate 3D geometry of human subjects. Finally, we demonstrate the ability to reconstruct different humans in detail from online videos and hand-held mobile phone video clips.

In summary, our contributions are:

- a method to reconstruct detailed 3D avatars from in-the-wild monocular videos via self-supervised scene decomposition; and
- to achieve robust and detailed 3D reconstructions of the human even under challenging poses and environments without requiring external segmentation methods; and
- a novel semi-synthetic testing dataset that for the first time allows comparing monocular human reconstruction methods on realistic scenes. The dataset contains rich annotations of the 3D surface.

Code and data will be made available for research purposes.

2. Related Work

Reconstructing Human from Monocular Video Traditional works for monocular human performance capture require personalized rigged templates as prior and track the pre-defined human model based on 2D observations [15, 16, 58]. These works require pre-scanning of the performer and post-processing for rigging, preventing such methods from being deployed to real-life applications. Some methods attempt to save the need for pre-scanning and manual rigging [3, 8, 14, 34]. However, the explicit mesh representation is limited to a fixed resolution and cannot represent details like the face. Regression-based methods that directly regress 3D surfaces from images have demonstrated compelling results [4, 12, 17, 18, 21, 41, 42, 55, 68]. However, they require high-quality 3D data for supervision and cannot maintain the space-time coherence of the reconstruction over the whole sequence. Recent works fit implicit neural fields to videos via neural rendering to obtain articulated human models [23, 24, 40, 46, 47, 52, 67]. HumanNeRF [52] extends articulated NeRF to improve novel view synthesis. NeuMan [24] further adds a scene NeRF model. Both methods model the human geometry with a density field, only yielding a noisy, and often low-fidelity human reconstruction. SelfRecon [23] deploys neural surface rendering [61] to achieve consistent reconstruction over the sequence. However, all aforementioned methods rely on pre-segmented inputs and are therefore not robust to uncontrolled visual complexity and are upper-bounded in their reconstruction quality by the external segmentation method. In contrast, our method solves the tasks of scene decomposition and surface reconstruction jointly in 3D without using segmentation modules.

Reconstructing Human from Multi-view/Depth The high fidelity 3D reconstruction of dynamic humans has required calibrated dense multi-view systems [9, 10, 19, 27, 31, 45, 49] which are expensive and laborious to operate and require highly-specialized expertise. Recent works [20, 22, 28, 37, 39, 51, 56, 57] attempt to reconstruct humans from more sparse settings by deploying neural rendering. Depth-based approaches [6, 35, 36] reconstruct the human shape by fusing depth measurements across time. Follow-up work [7, 11, 29, 63, 64] builds upon this concept by incorporating an articulated motion prior and a parametric body shape prior. While the aforementioned methods achieve compelling results, they still require a specialized capturing setup and are hence not applicable to in-the-wild settings. In contrast, our method recovers the dynamic human shape in the wild from a monocular RGB video as the sole input.

Moving Object Segmentation Traditional research in moving object segmentation has been extensively conducted at the image level (i.e. 2D). One line of research relies on motion clues to segment objects with different optical flow patterns [5, 38, 54, 59, 62], while another line of work, termed video matting [25, 30, 43] is trained on videos with human-annotated masks to directly regress the alpha-channel values during inference. Those approaches are not without limitations, as they focus on image-level segmentation and incorporate no 3D knowledge. Thus, they cannot handle complicated backgrounds without enough color contrast between the human and the background. Recent works learn to decompose dynamic objects and the static background simultaneously in 3D by optimizing multiple NeRFs [44, 48, 53, 65]. Such methods perform well for non-complicated dynamic objects but are not directly applicable to articulated humans with intricate motions.

3. Method

We introduce Vid2Avatar, a method for detailed geometry and appearance reconstruction of implicit neural avatars from monocular videos in the wild. Our method is schematically illustrated in Fig. 2. Reconstructing humans from in-the-wild videos is clearly challenging. Solving it requires accurately segmenting humans from arbitrary backgrounds without any prior knowledge about the appearance of the scene or the subject and requires reconstructing detailed 3D surface and appearance from short video sequences. In contrast to prior works that utilize off-the-shelf 2D segmentation tools or manually labeled masks, we solve the tasks of scene decomposition and surface reconstruction directly in 3D. To achieve this, we model both the human and background in the scene implicitly, parameterized via two separate neural fields which are learned jointly from images to composite the whole scene. To alleviate the ambiguity of in-contact body and scene parts and to better delineate the surfaces, we contribute novel objectives that leverage the dynamically updated human shape in canonical space to regularize the ray opacity.

We parameterize the 3D geometry and texture of clothed humans as a pose-conditioned implicit signed-distance field (SDF) and texture field in canonical space (Sec. 3.1). We then model the background using a separate neural radiance field (NeRF). The human shape and appearance fields alongside the background field are learned from images jointly via differentiable composited neural volume rendering (Sec. 3.2). Finally, we leverage the dynamically updated canonical human shape to regularize the ray opacities (Sec. 3.3). The training is formulated as global optimization to jointly optimize the dynamic foreground and static background fields, and the per-frame pose parameters (Sec. 3.4).

3.1. Implicit Neural Avatar Representation

Canonical Shape Representation. We model the human shape in canonical space to form a single, temporally consistent representation and use a neural network f_{sdf}^{H} to predict the signed distance value for any 3D point $\mathbf{x}_c$ in this space. To model pose-dependent local non-rigid deformations such as dynamically changing wrinkles on clothes, we concatenate the human pose $\boldsymbol{\theta}$ as an additional input and model f_{sdf}^{H} as:

$$f_{\text{sdf}}^{H} : \mathbb{R}^{3+n_\theta} \to \mathbb{R}^{1+256}. \tag{1}$$

The pose parameters $\boldsymbol{\theta}$ are defined analogously to SMPL [32], with dimensionality n_θ. Furthermore, f_{sdf}^{H} outputs global geometry features $\mathbf{z}$ of dimension 256. With slight abuse of notation, we also use f_{sdf}^{H} to refer to the SDF value only. The canonical shape $\mathcal{S}$ is given by the zero-level set of f_{sdf}^{H}:

$$\mathcal{S} = \{\, \mathbf{x}_c \mid f_{\text{sdf}}^{H}(\mathbf{x}_c, \boldsymbol{\theta}) = 0 \,\}. \tag{2}$$

Skeletal Deformation. Given the bone transformation matrix $\mathbf{B}_i$ for joint $i \in \{1, ..., n_b\}$ which are derived from the body pose $\boldsymbol{\theta}$, a canonical point $\mathbf{x}_c$ is mapped to the deformed point $\mathbf{x}_d$ via linear-blend skinning:

$$\mathbf{x}_d = \sum_{i=1}^{n_b} w_c^i \mathbf{B}_i \, \mathbf{x}_c. \tag{3}$$

The canonical correspondence $\mathbf{x}_c$ for points $\mathbf{x}_d$ in deformed space is defined by the inverse of Eq. 3:

$$\mathbf{x}_c = (\sum_{i=1}^{n_b} w_d^i \mathbf{B}_i)^{-1} \, \mathbf{x}_d \tag{4}$$

Here, n_b denotes the number of bones in the transformation, and $\mathbf{w}_{(\cdot)} = \{w_{(\cdot)}^1, ..., w_{(\cdot)}^{n_b}\}$ represents the skinning weights for $\mathbf{x}_{(\cdot)}$. Here, deformed points $\mathbf{x}_d$ are associated with the

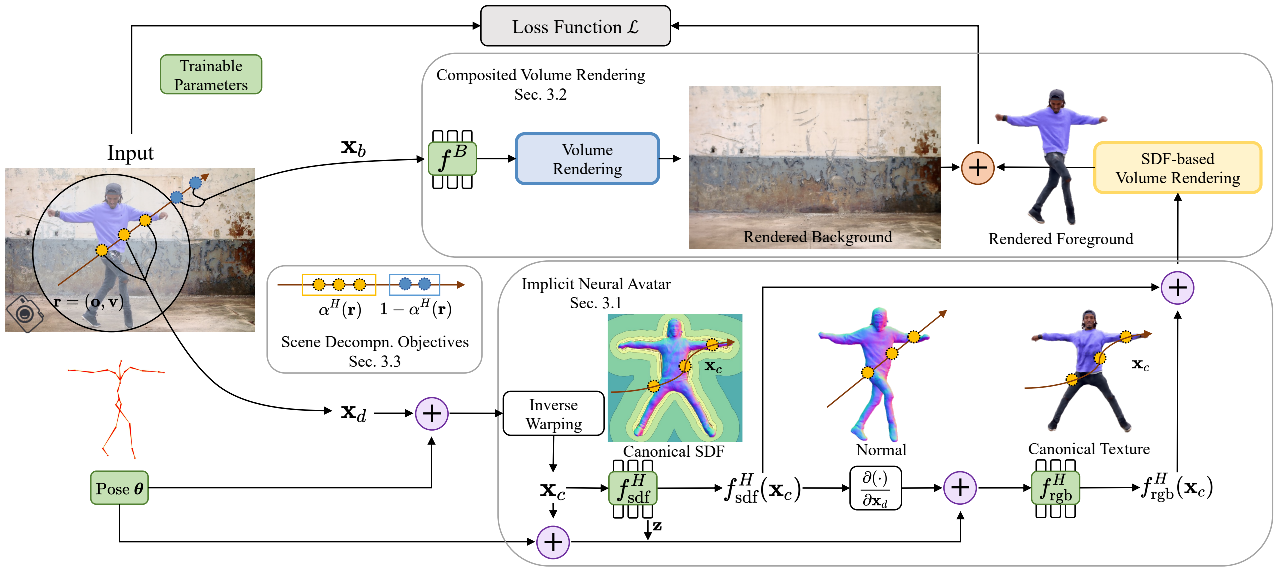
Loss Function $\mathcal{L}$
Trainable Parameters
Composited Volume Rendering
Sec. 3.2
Input
$\mathbf{x}_b$
f^B
Volume Rendering
Rendered Background
Rendered Foreground
SDF-based Volume Rendering
$\mathbf{r} = (\mathbf{o}, \mathbf{v})$
$\alpha^H(\mathbf{r})$
$1 - \alpha^H(\mathbf{r})$
Scene Decompn. Objectives
Sec. 3.3
Implicit Neural Avatar
Sec. 3.1
$\mathbf{x}_c$
Canonical SDF
Normal
Canonical Texture
$\mathbf{x}_d$
Inverse Warping
Pose $\boldsymbol{\theta}$
$\mathbf{x}_c$
f^H_{sdf}
$\mathbf{z}$
$f^H_{\text{sdf}}(\mathbf{x}_c)$
$\frac{\partial(\cdot)}{\partial \mathbf{x}_d}$
f^H_{rgb}
$f^H_{\text{rgb}}(\mathbf{x}_c)$

Figure 2. **Method Overview.** Given a ray $\mathbf{r}$ with camera center $\mathbf{o}$ and ray direction $\mathbf{v}$, we sample points densely ($\mathbf{x}_d$) and coarsely ($\mathbf{x}_b$) along the ray for the spherical inner volume and outer volume respectively. Within the foreground sphere, we warp all sampled points into canonical space via inverse warping and evaluate the SDF of the canonical correspondences $\mathbf{x}_c$ via the canonical shape network f^H_{sdf}. We calculate the spatial gradient of the sampled points in deformed space and concatenate them with the canonical points $\mathbf{x}_c$, the pose parameters $\boldsymbol{\theta}$, and the extracted geometry feature vectors $\mathbf{z}$ to form the input to canonical texture network f^H_{rgb} which predicts color values for $\mathbf{x}_c$. We apply surface-based volume rendering for the dynamic foreground and standard volume rendering for the background, and then composite the foreground and background components to attain the final pixel color. We minimize the loss $\mathcal{L}$ that compares the color predictions with the image observations along with novel scene decomposition objectives.

average of the nearest SMPL vertices' skinning weights, weighted by the point-to-point distances in deformed space. Canonical points $\mathbf{x}_c$ are treated analogously.

Canonical Texture Representation. The appearance is also modeled in canonical space via a neural network f^H_{rgb} that predicts color values for 3D points $\mathbf{x}_c$ in this space.

$$f^H_{\text{rgb}} : \mathbb{R}^{3+3+n_\theta+256} \rightarrow \mathbb{R}^3. \tag{5}$$

We condition the canonical texture network on the normal $\mathbf{n}_d$ in deformed space, facilitating better disentanglement of the geometry and appearance. The normals are given by the spatial gradient of the signed distance field w.r.t. the 3D location in deformed space. Following [67], the spatial gradient of the deformed shape is given by:

$$\begin{aligned}\mathbf{n}_d = \frac{\partial f^H_{\text{sdf}}(\mathbf{x}_c, \boldsymbol{\theta})}{\partial \mathbf{x}_d} &= \frac{\partial f^H_{\text{sdf}}(\mathbf{x}_c, \boldsymbol{\theta})}{\partial \mathbf{x}_c} \frac{\partial \mathbf{x}_c}{\partial \mathbf{x}_d} \\ &= \frac{\partial f^H_{\text{sdf}}(\mathbf{x}_c, \boldsymbol{\theta})}{\partial \mathbf{x}_c} (\sum_{i=1}^{n_b} w^i_d \mathbf{B}_i)^{-1}.\end{aligned} \tag{6}$$

In practice we concatenate the canonical points $\mathbf{x}_c$, their normals, the pose parameters, and the extracted 256-dimensional geometry feature vectors $\mathbf{z}$ from the shape network to form the input to the canonical texture network. For the remainder of this paper, we denote this neural SDF with $f^H_{\text{sdf}}(\mathbf{x}_c)$ and the RGB field as $f^H_{\text{rgb}}(\mathbf{x}_c)$ for brevity.

3.2. Composited Volume Rendering

We extend the inverted sphere parametrization of NeRF++ [66] to represent the scene: an outer volume (i.e., the background) covers the complement of a spherical inner volume (i.e., the space assumed to be occupied by the human) and both are modeled by separate networks. The final pixel value is then attained via compositing.

Background. Given the origin $\mathbf{O}$, each 3D point $\mathbf{x}_b = (x_b, y_b, z_b)$ in the outer volume is reparametrized by the quadruple $\mathbf{x}'_b = (x'_b, y'_b, z'_b, \frac{1}{r})$, where $\|(x'_b, y'_b, z'_b)\| = 1$, $(x_b, y_b, z_b) = r \cdot (x'_b, y'_b, z'_b)$. Here r denotes the magnitude of the vector from the origin $\mathbf{O}$ to $\mathbf{x}_b$. This parameterization of background points leads to improved numerical stability and assigns lower resolution to farther away points. For more details, we refer to [66]. Our method is trained with videos and the background is generally not entirely static. To compensate for dynamic changes in e.g., lighting, we condition the background network f^B on a per-frame learnable latent code t^i.

$$f^B : \mathbb{R}^{4+3+n_t} \rightarrow \mathbb{R}^{1+3}, \tag{7}$$

where f^B takes the 4D representation of the sampled background point $\mathbf{x}'_b$, viewing direction $\mathbf{v}$, and time encoding t^i with dimension n_t as input, and outputs the density and the view-dependent radiance.

Dynamic Foreground. We assume that the inner volume is occupied by a dynamic foreground – the human we seek to reconstruct. This requires different treatment compared to [66] where a static foreground is modeled via a vanilla NeRF. In contrast, we combine the implicit neural avatar representation (Sec. 3.1) with surface-based volume rendering [60]. Thus, we convert the SDF to a density σ by applying the scaled Laplace distribution's Cumulative Distribution Function (CDF) to the negated SDF values $\xi(\mathbf{x}_c) = -f^H_{\text{sdf}}(\mathbf{x}_c)$:

$$\sigma(\mathbf{x}_c) = \alpha \left(\frac{1}{2} + \frac{1}{2} \operatorname{sign}(\xi(\mathbf{x}_c))(1 - \exp(-\frac{|\xi(\mathbf{x}_c)|}{\beta})) \right), \tag{8}$$

where α, $\beta > 0$ are learnable parameters.

Similar to [60], we sample N points on a ray $\mathbf{r} = (\mathbf{o}, \mathbf{v})$ with camera center $\mathbf{o}$ and ray direction $\mathbf{v}$ in two stages – uniform and inverse CDF sampling. We then map the sampled points to canonical space via skeletal deformation and use standard numerical approximation to calculate the integral of the volume rendering equation:

$$C^H(\mathbf{r}) = \sum_{i=1}^{N} \tau_i f^H_{\text{rgb}}(\mathbf{x}^i_c) \tag{9}$$

$$\tau_i = \exp\left(-\sum_{j<i} \sigma(\mathbf{x}^j_c)\delta^j \right) (1 - \exp(-\sigma(\mathbf{x}^i_c)\delta^i)) \tag{10}$$

where $\delta^{(i)}$ is the distance between two adjacent samples. Here, the accumulated alpha value of a pixel, which represents ray opacity, can be obtained by $\alpha^H(\mathbf{r}) = \sum_{i=1}^{N} \tau_i$.

Scene Composition. To attain the final pixel value for a ray $\mathbf{r}$, we raycast the human and background volumes separately, followed by a scene compositing step. Using the parameterization of the background, we sample r along the ray $\mathbf{r}$ to obtain sample points in the outer volume for which we query f^B. The background component of a pixel is then given by the integrated color value $C^B(\mathbf{r})$ along the ray [33]. More details can be found in the Supp. Mat. The final pixel color is then the composite of the foreground and background color.

$$C(\mathbf{r}) = C^H(\mathbf{r}) + (1 - \alpha^H(\mathbf{r}))C^B(\mathbf{r}). \tag{11}$$

3.3. Scene Decomposition Objectives

Learning to decompose the scene into a dynamic human and background by simply minimizing the distance between the composited pixel value and image RGB value is still a severely ill-posed problem. This is due to the potentially moving scene, dynamic shadows, and general visual complexity. To this end, we propose two objectives that guide the optimization towards a clean and robust decoupling of the human from the background.

Opacity Sparseness Regularization. One of the key components of our method is a loss L_{sparse} to regularize the ray opacity via the dynamically updated human shape in canonical space. We first warp sampled points into the canonical space and calculate the signed distance to the human shape. We then penalize non-zero ray opacities for rays that do not intersect with the subject. This ray set is denoted as $\mathcal{R}^i_{\text{off}}$ for frame i.

$$\mathcal{L}^i_{\text{sparse}} = \frac{1}{|\mathcal{R}^i_{\text{off}}|} \sum_{\mathbf{r} \in \mathcal{R}^i_{\text{off}}} |\alpha^H(\mathbf{r})|. \tag{12}$$

Note that we conservatively update the SDF of the human shape throughout the whole training process which leads to a precise association of human and background rays.

Self-supervised Ray Classification. Even with the shape regularization from Eq. 12, we observe that the human fields still tend to model parts of the background due to the flexibility and expressive power of MLPs, especially if the subject is in contact with the scene. To further delineate dynamic foreground and background, we introduce an additional loss term to encourage ray distributions that contain either fully transparent or opaque rays:

$$\begin{aligned} \mathcal{L}^i_{\text{BCE}} = -\frac{1}{|\mathcal{R}^i|} \sum_{\mathbf{r} \in \mathcal{R}^i} & (\alpha^H(\mathbf{r}) \log(\alpha^H(\mathbf{r})) \\ & + (1 - \alpha^H(\mathbf{r})) \log(1 - \alpha^H(\mathbf{r}))), \end{aligned} \tag{13}$$

where $\mathcal{R}^i$ denotes the sampled rays for frame i. This term penalizes deviations of the ray opacities from a binary $\{0, 1\}$ distribution via the binary cross entropy loss. Intuitively this encourages the opacities to be zero for rays that hit the background and one for those that hit the human shape. In practice, this significantly helps separation of the subject and the background, in particular for difficult cases with similar pixel values across discontinuities.

The final scene decomposition loss is then given by L_{dec}:

$$\mathcal{L}_{\text{dec}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}}. \tag{14}$$

3.4. Global Optimization

To train the models that represent the background and the dynamic foreground jointly from videos, we formulate the training as global optimization over all frames.

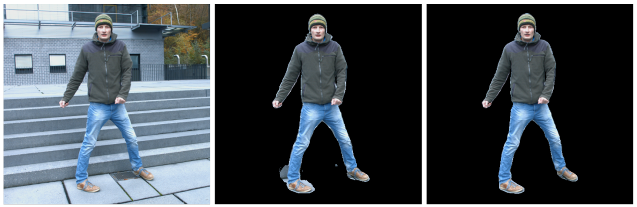

Figure 3. **Importance of scene decomposition loss.** Without the scene decomposition loss, the segmentation includes undesirable background parts due to similar pixel values across discontinuities.

Eikonal Loss. Following IGR [13], we leverage $\mathcal{L}^i_{\text{eik}}$ to force the shape network f^H_{sdf} to satisfy the Eikonal equation in canonical space:

$$\mathcal{L}^i_{\text{eik}} = \mathbb{E}_{\mathbf{x}_c}\left(\|\nabla f^H_{\text{sdf}}(\mathbf{x}_c)\| - 1\right)^2. \tag{15}$$

Reconstruction Loss. We calculate the L_1-distance between the rendered color $C(\mathbf{r})$ and the pixel's RGB value $\hat{C}(\mathbf{r})$ to attain the reconstruction loss $\mathcal{L}^i_{\text{rgb}}$ for frame i:

$$\mathcal{L}^i_{\text{rgb}} = \frac{1}{|\mathcal{R}^i|}\sum_{\mathbf{r}\in\mathcal{R}^i} |C(\mathbf{r}) - \hat{C}(\mathbf{r})|. \tag{16}$$

Full Loss. Given a video sequence with F input frames, we minimize the combined loss function:

$$\mathcal{L}(\mathbf{\Theta}) = \sum_{i=1}^{F} \mathcal{L}^i_{\text{rgb}}(\mathbf{\Theta}^H, \mathbf{\Theta}^B) + \lambda_{\text{dec}}\mathcal{L}^i_{\text{dec}}(\mathbf{\Theta}^H) + \lambda_{\text{eik}}\mathcal{L}^i_{\text{eik}}(\mathbf{\Theta}^H) \tag{17}$$

where $\mathbf{\Theta}^H$ and $\mathbf{\Theta}^B$ are the sets of optimized parameters for the human and background model respectively. $\mathbf{\Theta}^H$ includes the shape network weights $\mathbf{\Theta}^H_{\text{sdf}}$, the texture network weights $\mathbf{\Theta}^H_{\text{rgb}}$ and per-frame pose parameters $\boldsymbol{\theta}_i$. $\mathbf{\Theta}^B$ contains the background density and radiance network weights.

4. Experiments

We first conduct ablation studies on our design choices. Next, we compare our method with state-of-the-art approaches in 2D segmentation, novel view synthesis, and reconstruction tasks. Finally, we demonstrate human reconstruction results on several in-the-wild monocular videos from different sources qualitatively.

4.1. Datasets

MonoPerfCap Dataset [58]: This dataset contains in-the-wild human performance sequences with ground-truth masks. Since our method can provide human segmentation as by-product, we use this dataset to compare our method with other off-the-shelf 2D segmentation approaches to validate the scene decomposition quality of our method.

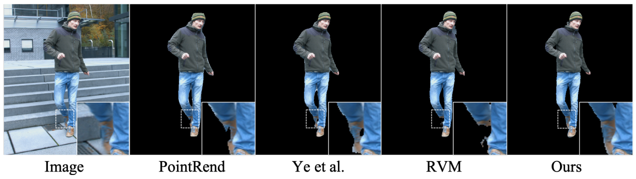

Figure 4. **Qualitative mask comparison.** Our method generates more detailed and robust segmentations compared to 2D segmentation methods by incorporating 3D knowledge.

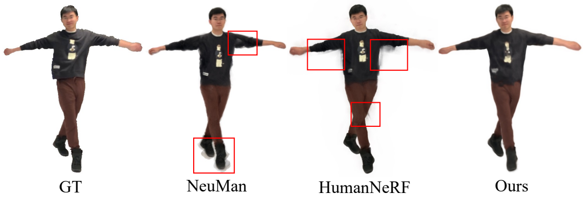

Figure 5. **Qualitative view synthesis comparison.** Our method achieves comparable and even better novel view synthesis results compared to NeRF-based methods (see also Sec. 4.4).

NeuMan Dataset [24]: This dataset includes a collection of videos captured by a mobile phone, in which a single person performs walking. We use this dataset to compare our rendering quality of humans under unseen views with other approaches.

3DPW Dataset [50]: This dataset contains challenging in-the-wild video sequences along with accurate 3D human poses recovered by using IMUs and a moving camera. Moreover, it includes registered static clothed 3D human models. By animating the human model with the ground-truth poses, we can obtain quasi ground-truth scans to evaluate the surface reconstruction performance.

SynWild Dataset: We propose a new dataset called *SynWild* for the evaluation of monocular human surface reconstruction method. We capture dynamic human subjects in a multi-view system and reconstruct the detailed geometry and texture via commercial software [9]. Then we place the textured 4D scans into realistic 3D scenes/HDRI panoramas and render monocular videos from virtual cameras, leveraging a high-quality game engine [2]. This is the first dataset that allows for quantitative comparison in a realistic setting via semi-synthetic data.

Evaluation Protocol: We consider precision, F1 score, and mask IoU for human segmentation evaluation. We report volumetric IoU, Chamfer distance (cm) and normal consistency for surface reconstruction comparison. Rendering quality is measured via SSIM and PSNR.

4.2. Ablation Study

Jointly Pose Optimization: The initial pose estimate from a monocular RGB video is usually inaccurate. To eval-

Method	**Precision** ↑	**F1** ↑	**IoU** ↑
SMPL Tracking	0.829	0.781	0.659
PointRend [26]	0.957	0.960	0.915
Ye et al. [62]	0.945	0.947	0.890
RVM [30]	0.975	0.977	0.950
w/o Scene Dec. Loss	0.979	0.974	0.942
Ours	**0.983**	**0.983**	**0.961**

Table 1. **Quantitative evaluation on MonoPerfCap.** Our method outperforms all 2D segmentation baselines in all metrics.

Method	**SSIM** ↑	**PSNR** ↑
NeuMan [24]	0.958	23.9
HumanNeRF [52]	0.963	24.8
Ours	**0.964**	**25.1**

Table 2. **Quantitative evaluation on NeuMan.** We report the quantitative results on test views. Our method achieves on-par and even better rendering quality compared to NeRF-based methods.

Method	**IoU** ↑	$\mathbf{C}-\ell_2$ ↓	**NC** ↑
ICON [55]	0.718	3.32	0.731
SelfRecon [23]	0.648	3.31	0.675
w/o Joint Opt.	0.810	3.00	0.737
Ours	**0.818**	**2.66**	**0.753**

Table 3. **Quantitative evaluation on 3DPW.** Our method consistently outperforms all baselines in all metrics (*cf*. Fig. 6).

Method	**IoU** ↑	$\mathbf{C}-\ell_2$ ↓	**NC** ↑
ICON [55]	0.764	2.91	0.766
SelfRecon [23]	0.805	2.50	0.776
Ours	**0.813**	**2.35**	**0.796**

Table 4. **Quantitative evaluation on SynWild.** Our method consistently outperforms all baselines in all metrics (*cf*. Fig. 6).

uate the importance of jointly optimizing pose, shape, and appearance, we compare our full model to a version without jointly pose optimization. Tab. 3 shows that the joint optimization significantly helps in global pose alignment and to recover details (normal consistency), this is also confirmed by qualitative results. Please see the Supp. Mat.

Scene Decomposition Loss: To demonstrate the effectiveness of our proposed scene decomposition loss, we conduct an ablation experiment without this term during optimization. Results in Tab. 1 indicate that without the scene decomposition loss, the segmentation tends to be noisy and includes parts of the background as shown in Fig. 3.

4.3. 2D Segmentation Comparisons

We generate human masks by extracting the pixels with ray opacity $\alpha^H(\mathbf{r})$ value of 1. Our produced masks are compared with SMPL Tracking, PointRend [26], Ye *et al.* [62] and RVM [30] on the MonoPerfCap dataset [58]. [26] and [30] are trained on large datasets with human-annotated masks, while [62] rely on optical flow as motion clues to segment objects in an unsupervised manner. SMPL Tracking uses dilated projected SMPL masks as the result. Tab. 1 shows the quantitative results. Our method consistently outperforms other baseline methods on all metrics. Fig. 4 shows that other baselines struggle on the feet since there is no enough photometric contrast between the part of the shoes and the stairs. In contrast, our method is able to generate plausible human segmentation via decomposition from a 3D perspective.

4.4. View Synthesis Comparisons

Though not our primary goal, we also compare with HumanNeRF [52] and NeuMan [24] for the task of novel view synthesis on the NeuMan dataset. Note that both methods require additional human segmentation as input. Overall, we achieve comparable or even better performance quantitatively (*cf*. Tab. 2). As shown in Fig. 5, NeuMan and HumanNeRF have obvious artifacts around feet and arms. This is because, a) off-the-shelf tools struggle to produce consistent masks and b) NeRF-based methods are known to have "hazy" floaters in the space leading to visually unpleasant results. Our method produces more plausible renderings of the human with a clean separation from the background.

4.5. Reconstruction Comparisons

We compare our proposed human surface reconstruction method to several state-of-the-art approaches [23, 55] on both 3DPW [50] and SynWild. ICON (image-based) [55] reconstructs 3D clothed humans by learning a regression model from a large dateset of clothed human scans [1]. SelfRecon (video-based) [23] deploys implicit surface rendering to reconstruct avatars from monocular videos. Both methods rely on additional human masks as input for their methods. Despite this, our method outperforms [23, 55] by a substantial margin on all metrics (*cf*. Tab. 3, Tab. 4). The difference is more visible in qualitative comparison as shown in Fig. 6, where they tend to produce physically incorrect body reconstructions (e.g., missing arms and sunken backs). In contrast, our method generates complete human bodies and recovers more surface details (e.g., cloth wrinkles and facial features). We attribute this to the better decoupling of humans from the background by our proposed modeling and learning schemes.

4.6. Qualitative Results

We demonstrate our results on several in-the-wild monocular videos from different sources: online, datasets, and self-captured video clips (Fig. 7). Our method is able to reconstruct complex cloth deformations and personalized facial features in detail. **Please refer to Supp. Mat for more qualitative results**.

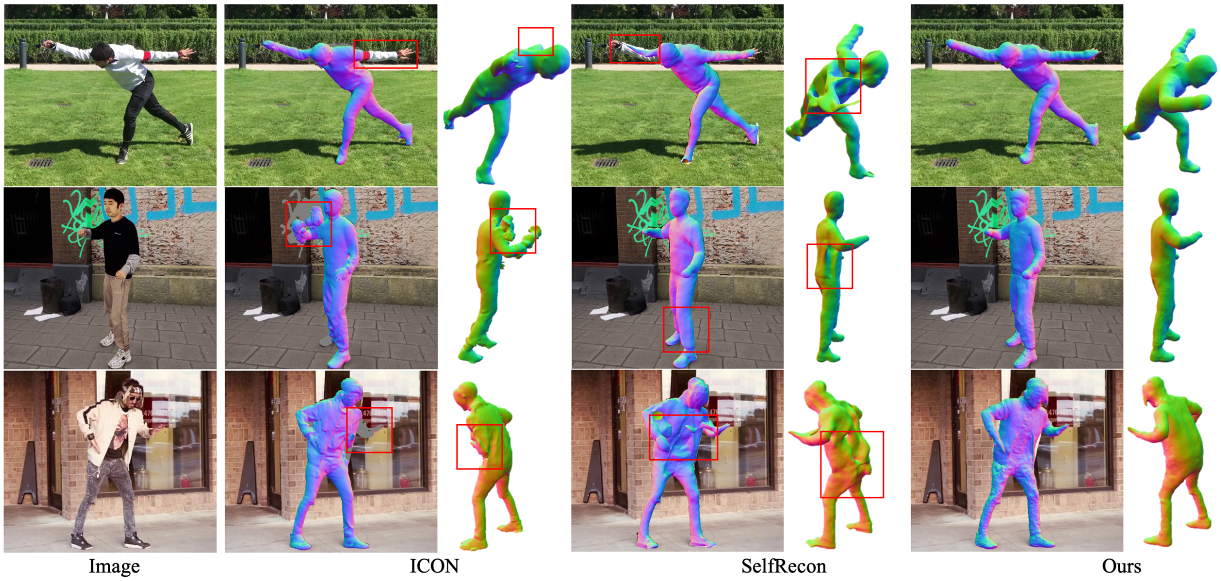

Figure 6. **Qualitative reconstruction comparison.** Data source top to bottom: 3DPW, SynWild, Online. ICON and SelfRecon produce less detailed and even physically implausible reconstructions (incomplete human bodies). In contrast, our method generates complete human bodies and achieves a detailed (e.g., cloth wrinkles) and temporally consistent shape reconstruction.

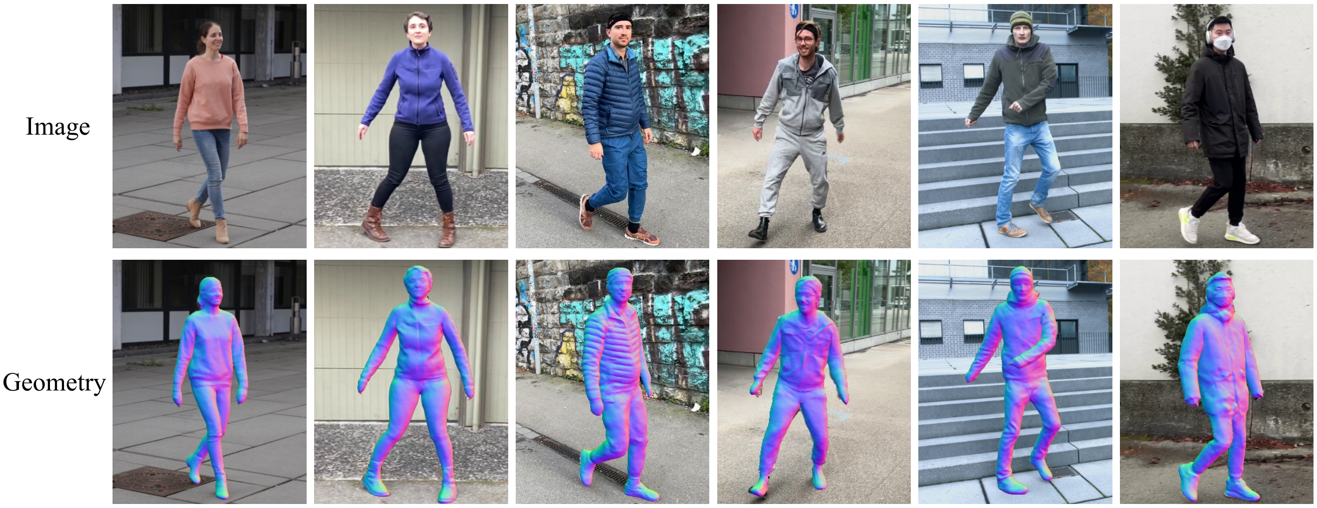

Figure 7. **Qualitative results.** We show qualitative results of our method from monocular in-the-wild videos.

5. Conclusion

In this paper, we present Vid2Avatar to reconstruct detailed 3D avatars from monocular in-the-wild videos via self-supervised scene decomposition. Our method does not require any groundtruth supervision or priors extracted from large datasets of clothed human scans, nor do we rely on any external segmentation modules. With carefully designed background modeling and temporally consistent canonical human representation, a global optimization with novel scene decomposition objectives is formulated to jointly optimize the parameters of the background field, the canonical human shape and appearance, and the human pose estimates over the entire sequence via differentiable composited volume rendering. Our method achieves robust and high-fidelity human reconstruction from monocular videos.

Limitations: Although readily available, Vid2Avatar relies on reasonable pose estimates as inputs. Furthermore, loose clothing such as skirts or free-flowing garments poses significant challenges due to their fast dynamics. We refer to Supp. Mat for a more detailed discussion of limitations and societal impact.

References

[1] *Renderpeople*, 2018. https://www.renderpeople.com. 7

[2] *Unreal*, 2020. https://www.unrealengine.com. 6

[3] 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. 1, 2

[4] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 433–449, Cham, 2016. Springer International Publishing. 3

[6] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Niessner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1450–1459, June 2021. 3

[7] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021. 3

[8] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. PERGAMO: Personalized 3d garments from monocular video. *Computer Graphics Forum (Proc. of SCA), 2022*, 2022. 1, 2

[9] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), jul 2015. 1, 3, 6

[10] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. 27(3):1–10, 2008. 1, 3

[11] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20470–20480, June 2022. 3

[12] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In *NeurIPS*, 2022. 2

[13] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 6

[14] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 1, 2

[15] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1, 2

[16] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 38(2), mar 2019. 1, 2

[17] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9276–9287. Curran Associates, Inc., 2020. 2

[18] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11046–11056, October 2021. 2

[19] A. Hilton and J. Starck. Multiple view reconstruction of people. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 357–364, 2004. 1, 3

[20] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *2022 International Conference on 3D Vision (3DV)*, 2022. 3

[21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2

[22] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 3

[23] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 7

[24] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2, 6, 7

[25] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 3

[26] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. 2019. 7

[27] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Volume Sweeping: Learning Photoconsistency for Multi-View Shape Reconstruction. *International Journal of Computer Vision*, 129:284–299, Feb. 2021. 1, 3

[28] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 3

[29] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14172, 2021. 3

[30] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance, 2021. 3, 7

[31] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Trans. Vis. Comput. Graph.*, 16(3):407–418, 2010. 1, 3

[32] 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. 3

[33] 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. 5

[34] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[35] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 3

[36] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011. 3

[37] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 3

[38] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *2013 IEEE International Conference on Computer Vision*, pages 1777–1784, 2013. 3

[39] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 3

[40] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2

[41] 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

[42] 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. 2

[43] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Computer Vision and Pattern Regognition (CVPR)*, 2020. 3

[44] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Ambrus, Adrien Gaidon, William T. Freeman, Fredo Durand, Joshua B. Tenenbaum, and Vincent Sitzmann. Seeing 3d objects in a single image via self-supervised static-dynamic disentanglement, 2022. 3

[45] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 1, 3

[46] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, 2022. 2

[47] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 2

[48] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021. 3

[49] Vagia Tsiminaki, Jean-Sébastien Franco, and Edmond Boyer. High Resolution 3D Shape Texture from Multiple Videos. In *CVPR 2014 - IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1502–1509, Columbus, OH, United States, June 2014. IEEE. 1, 3

[50] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 6, 7

[51] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 3

[52] 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 (CVPR)*, pages 16210–16220, June 2022. 2, 7

[53] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D^{2}nerf: Self-supervised decoupling of dynamic and static objects from a monocular video, 2022. 3

[54] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9986–9995, 2019. 3

[55] 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 (CVPR)*, pages 13296–13306, June 2022. 2, 7

[56] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-neRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 3

[57] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *CVPR*, 2022. 3

[58] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. 37(2):27:1–27:15, May 2018. 1, 2, 6, 7

[59] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 3

[60] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2, 5

[61] 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, 2020. 2

[62] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3, 7

[63] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. 3

[64] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018. 3

[65] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13144–13152, 2021. 3

[66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 2, 4, 5

[67] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[68] 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. 2