SOAR: Self-Occluded Avatar Recovery from a Single Video In the Wild

Zhuoyang Pan[1,2,*], Angjoo Kanazawa[1], and Hang Gao[1,*]

[1]UC Berkeley
[2]ShanghaiTech University

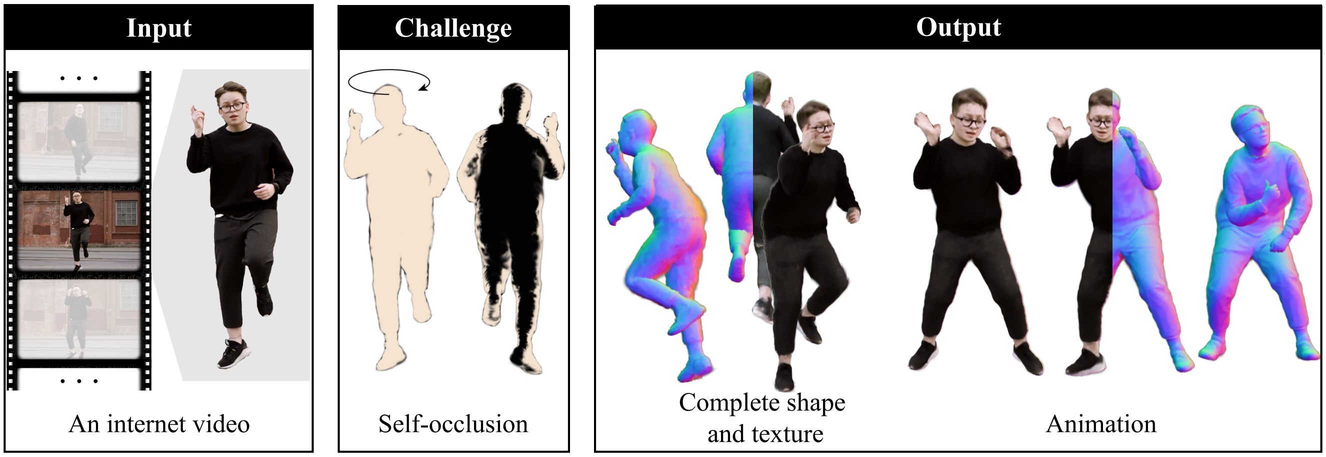

Figure 1. **Complete human reconstruction from partial observations in the wild.** We present **SOAR**: **S**elf-**O**ccluded **A**vatar **R**ecovery. Given a video of a moving human where parts of the body are entirely unobserved (left), SOAR recovers a photo-realistic avatar with complete texture and shape (right), by leveraging structural human normal prior and generative diffusion prior.

Abstract

Self-occlusion is common when capturing people in the wild, where the performer do not follow predefined motion scripts. This challenges existing monocular human reconstruction systems that assume full body visibility. We introduce Self-Occluded Avatar Recovery (SOAR), a method for complete human reconstruction from partial observations where parts of the body are entirely unobserved. SOAR leverages structural normal prior and generative diffusion prior to address such an ill-posed reconstruction problem. For structural normal prior, we model human with an reposable surfel model with well-defined and easily readable shapes. For generative diffusion prior, we perform an initial reconstruction and refine it using score distillation. On various benchmarks, we show that SOAR performs favorably than state-of-the-art reconstruction and generation methods, and on-par comparing to concurrent works. Additional video results and code are available at https://soar-avatar.github.io/.

1. Introduction

Recovering life-like human avatar from a single in-the-wild video, such as internet footage or smartphone capture, is crucial for advancing virtual reality, robotics, and content creation. This task is challenging due to dynamic modeling and the lack of effective multi-view signals [12]. Despite tremendous progress [23, 30, 40, 50, 58, 59] in recent years, success in human reconstruction methods in the wild remains limited. One key reason is that existing approaches often assume full visibility of the human body, which fails in most of unscripted casual captures. For this ill-posed problem, reconstruction alone is insufficient.

We present SOAR, a general system for human avatar recovery from a single self-occluded video in the wild. In Figure 1, we demonstrate our setting and results. We tackle this challenging problem with two key insights. First, to optimize with ill constraints, we need stronger data terms and more parsimonious representations. Second, we need to combine reconstruction with generation based on how many observations we have. With more observations, we prioritize

* Equal Contribution.

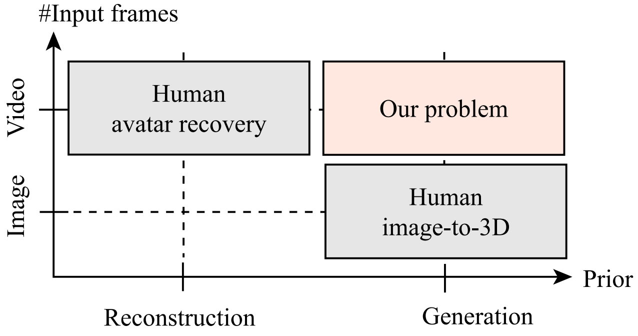

Figure 2. **Relation to existing problems.** Our problem requires combining human reconstruction from video frames and human generation for occluded regions.

reconstruction to preserve details like identity. With fewer observations, generation becomes crucial. A successful system should seamlessly integrate these two components.

Motivated by these two insights, we model the human avatar as a globally consistent set of Gaussian surfels [9] with well-defined and easily readable normals. We model articulation between different poses using a simple forward mapping with linear blend skinning [35]. We fit this compact, dynamic human representation to a general self-occluded video in the wild by incorporating two additional sources of supervision on top of the input RGB data: structural human normal prior [51, 52] and generative diffusion prior [48]. They provide strong shape and texture constraints for unobserved regions, crucial in our challenging problem setup. To this end, our approach is able to recover complete photo-realistic avatar with highly detailed geometry, which can be used for real-time rendering and animation.

To investigate the effectiveness and robustness of our approach, we compare against reconstruction-based [18, 29] and generation-based approaches [17] as baselines. We also compare with concurrent work HAVE-FUN [53] that reconstructs from partial observations on its own experimental protocols using the official open-source implementation. Extensive experiments show that SOAR performs favorably than state-of-the-art reconstruction and generation methods, and on-par comparing to concurrent works.

2. Related work

2.1. 3D Gaussian and surfel splatting

Neural rendering has advanced significantly since the introduction of NeRF [36]. 3D Gaussian Splatting [25] is particularly notable for its efficiency in high-resolution synthesis and real-time rendering. It represents scenes as explicit 3D Gaussians, allowing direct rasterization in pixel space that is much faster than volume integration. However, 3D Gaussians struggle with accurate scene geometry recovery. Various attempts [15, 24, 32] have been made to solve this problem. Recently, 2D Gaussian Splatting [19] and Gaussian surfels [9] propose to flatten 3D Gaussians into surfels, making geometry easier to readout and coupled with RGB rendering. Our work builds on these advancements and use surfel model for precise human shape recovery while preserving effective appearance modeling.

2.2. Neural rendering for human reconstruction

Neural rendering significantly advances template-based human reconstruction [5, 10, 11, 45] by allowing 3D avatar recovery from 2D images. Recent works have focused on dynamic modeling, out-of-distribution reposing, and runtime efficiency. NeuralBody [40] and Vid2Avatar [16] are canonical frameworks in the first category. For reposing, Animatable NeRF [39], TAVA [30], and InstantAvatar [23] use inverse blend skinning or root finding [7] to ensure consistency. Recent methods [46, 59] employ 3D Gaussians for efficient rendering. We select GART [29] and GaussianAvatar [18] as baseline in our experiments. We also compare with concurrent work HAVE-FUN [53] that aims to recover complete avatar from partial observations on its own benchmarks. We found that existing benchmarks all assume full-body visibility, even for our concurrent works, and thus test our methods on a new evaluation split from DNA-Rendering [8].

2.3. Diffusion prior for human generation

Score distillation sampling [41] has shown that 2D diffusion models are effective 3D priors for content creation. Since then, significant progress has been made in making 3D generation more stable [44, 48], realistic [49] and efficient [47, 54]. In the human modeling community, this paradigm has also been adopted, with notable works [28, 33, 55] incorporating predefined SMPL templates [35] to bias the generation process. Works on human image-to-3D [17, 20, 57] aim to recover human avatars from a single photo. For example, TeCH [20] optimizes a differentiable tetrahedron representation [43] through score distillation, while SiTH [17] directly optimizes an SDF field from diffused images. We include SiTH as a baseline in our work. However, these approaches struggle with video input, resulting in temporally inconsistent prediction when applied frame-by-frame, investigated in Section 4.3. Our work fuses pose-conditioned, noisy diffusion priors into a single, globally consistent avatar model.

3. Method

We aim to recover photo-realistic human avatar from a single self-occluded in-the-wild video, where parts of the human body remain unobserved. This highly ill-posed problem necessitates stronger priors and better avatar representation.

Existing human reconstruction methods [16, 18, 23, 29–31, 40] require the performer to reveal 360 views of their body, which does not often occur in internet videos. Conversely, existing human image-to-3D methods [2, 17, 20, 57]

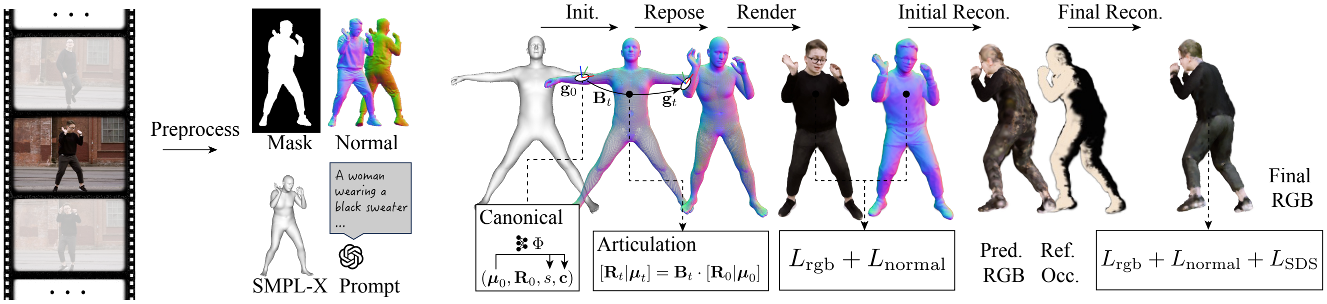
. . .
. . .
Preprocess
Mask
Normal
A woman wearing a black sweater ...
SMPL-X
Prompt
Init.
Repose
Render
Initial Recon.
Final Recon.
$\mathbf{g}_0$
$\mathbf{B}_t$
$\mathbf{g}_t$
Canonical
Φ
$(\boldsymbol{\mu}_0, \mathbf{R}_0, s, \mathbf{c})$
Articulation
$[\mathbf{R}_t|\boldsymbol{\mu}_t] = \mathbf{B}_t \cdot [\mathbf{R}_0|\boldsymbol{\mu}_0]$
$L_{\text{rgb}} + L_{\text{normal}}$
Pred. RGB
Ref. Occ.
Final RGB
$L_{\text{rgb}} + L_{\text{normal}} + L_{\text{SDS}}$

Figure 3. **System overview.** Given an input video, we preprocess for frame-wise mask, front and back normal, SMPL-X parameters, as well as video-level text prompt description (Section 3.1). Our model consists of a canonical Gaussian surfel representation and an articulation representation (Section 3.2). We perform initial reconstruction while estimating occlusion, producing partially completed avatar due to the lack of observation (Section 3.3), which is then refined by generative diffusion priors (Section 3.4).

can only condition on one input view, producing inconsistent results across frames. Our method bridges the gap between reconstruction and generation, addressing these challenges to produce consistent and accurate human avatars with self-occlusion.

The rest of this section is organized as follows. First, we talk about the preprocessing step given a single in-the-wild video (Section 3.1) Next, we discuss our avatar model, represented as a globally consistent set of 3D Gaussian surfels [9] that transforms from a canonical space to each pose configuration (Section 3.2). Then, we fuse RGB and structural normal supervision through an initial reconstruction while estimating occlusion in 3D (Section 3.3). Finally, we refine this initial reconstruction using score distillation (Section 3.4). Our whole pipeline is illustrated in Figure 3.

3.1. Preprocessing

Given a sequence of video frames capturing a moving person, we prepare a set of estimates using off-the-shelf methods. Specifically, for each frame $\mathbf{I}_t$, we estimate the foreground mask $\mathbf{M}_t$ using SAM [27], generate a video-level text prompt $\mathbf{p}$ using GPT-4o [1], obtain front and back normal maps $(\mathbf{N}_t, {}^{B}\mathbf{N}_t)$ using ICON [51], and infer 2D keypoints $\mathbf{k}_t \in \mathbb{R}^{137\times 2}$ with confidence $\psi_t \in \mathbb{R}^{137}$ using OpenPose [4] including body, hands and facial landmarks. Additionally, we extract SMPL-X body shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and body pose $\boldsymbol{\theta}_t \in \mathbb{R}^{52\times 3}$, as well as camera parameters $\boldsymbol{\pi}_t = [\mathbf{K}_t \in \mathbb{R}^{3\times 3}, \mathbf{E}_t \in \mathbb{SE}(3)]$ using SMPLer-X [3].

We find that high quality alignment between the reprojected SMPL-X model and human pixels is crucial to final results. Indeed, most previous works [6, 18, 23, 29] jointly refine SMPL/SMPL-X parameters along reconstructing the human avatar. However, we find it sufficient to refine SMPL-X in the preprocessing step, akin to SMPLify-X [38], without joint optimizing the avatar. Concretely, we seek to solve the following optimization problem that balances between pixel alignment and temporal smoothness:

$$\min_{\boldsymbol{\beta},\{\boldsymbol{\theta_t}\},\{\mathbf{b}_t\}} \lambda_{\text{data}} E_{\text{data}} + \lambda_{\text{smooth}} E_{\text{smooth}} + \lambda_{\text{preserve}} E_{\text{preserve}},$$
$$E_{\text{data}} = \psi_t \rho(\mathbf{k}_t - \hat{\mathbf{k}}_t),$$
$$E_{\text{smooth}} = \|\boldsymbol{\Theta}_{t-1}^T \boldsymbol{\Theta}_t\|, \quad \boldsymbol{\Theta}_t = \texttt{Rodrigues}(\boldsymbol{\theta}_t),$$
$$E_{\text{preserve}} = \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\| + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^{(0)}\|, \tag{1}$$

where ρ is the robust Geman-McClure function [14], $\hat{\mathbf{k}}_t$ is the reprojected SMPL-X keypoints from current estimates, and $\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}_t^{(0)}$ are the initial SMPL-X prediction. We set $\lambda_{\text{data}} = 100.0, \lambda_{\text{smooth}} = 10000.0, \lambda_{\text{preserve}} = 60.0$ throughout our experiments. We optimize with the second order LBFGS optimizer [34] with a learning rate $\eta = 1.0$ for a total epochs $K = 40$.

3.2. Globally-consistent surfel avatar

We encode the human appearance and geometry with a global set of 3D Gaussian surfels [9] that allows expressive differentiable rendering and surface modeling. Similar to existing Gaussian-based avatars [18, 29, 59], we define surfels in a single canonical space, which can be reposed using forward skinning as opposed to backward root-finding used in previous NeRF-based avatars [7, 23, 30]. A pictorial illustration is shown in Figure 3.

Canonical representation. For each surfel $\mathbf{g}_0$ that lives in the canonical frame t_0, we define their attributes as

$$\mathbf{g}_0 \equiv (\boldsymbol{\mu}_0, \mathbf{R}_0, s, \mathbf{c}, \tau), \tag{2}$$

where the position $\boldsymbol{\mu}_0 \in \mathbb{R}^3$ and the orientation $\mathbf{R}_0 \in \mathbb{SO}(3)$ can be reposed, the scale $s \in \mathbb{R}$ and the color $\mathbf{c} \in \mathbb{R}^3$ are constant across poses. Additionally, we assign the occlusion $\tau \in [0, 1]$ to each canonical surfel, with 1 indicating full occlusion, for evaluation. Similar to GaussianAvatar [18], we treat each surfel as an oriented round disk with isotropic scale to prevent needle-like artifacts after reposing. We keep surfels constantly opaque, *i.e.* $o = 1$, to avoid semi-transparent surfaces after alpha compositing. The surfel normal $\mathbf{n}_0$ can be read out trivially as the last column component in $\mathbf{R}_0$.

We find that explicit parameters tend to have large variance after convergence given sparse supervision, which leads to high frequency artifacts when applying score distillation sampling [41]. To this end, we employ a hybrid parameterization of surfel attributes. Specifically, we define $\boldsymbol{\mu}_0$ and $\mathbf{R}_0$ as explicit parameters and use a hash-based MLP network Φ to predict s and $\mathbf{c}$:

$$\Phi : \boldsymbol{\mu}_0 \mapsto s, \mathbf{c}, \tag{3}$$

where each attribute has its own shallow MLP network, taken as input a shared hash grid encoding [37]. We ablate over this design choice in Section 4.5.

We initialize surfels in a predefined virtruvian pose by subdividing corresponding SMPL-X mesh. Concretely, we subdivide SMPL-X mesh twice and obtain $N = 167333$ oriented vertices, which is used to initialize $\boldsymbol{\mu}_0, \mathbf{R}_0$. We compute the initial s as the average point-to-point distance between each surfel and its 3-nearest neighbors, as per [25]. Since we adopt implicit parameterization Φ for s, we supervise Φ with our pre-computed $(\boldsymbol{\mu}_0, s)$ labels for proper initialization.

Articulation representation. Given the SMPL-X parameters $\boldsymbol{\beta}, \boldsymbol{\theta}_t$, we can compute their corresponding bone transformations $\{\mathbf{B}_{t,j}\}$ for each joints j. We then articulate each canonical surfel $\mathbf{g}_0$ to posed surfel $\mathbf{g}_t \equiv (\boldsymbol{\mu}_t, \mathbf{R}_t, s, \mathbf{c})$ by linear blend skinning

$$[\mathbf{R}_t|\boldsymbol{\mu}_t] = \mathbf{B}_t \cdot [\mathbf{R}_0|\boldsymbol{\mu}_0], \quad \text{where } \mathbf{B}_t = \sum_j w_j \mathbf{B}_{t,j}. \tag{4}$$

w_j is the average skinning weight of the nearest $K = 30$ SMPL-X vertices, weighted by the point-to-point distances in canonical space, similar to [16, 18].

We note that, this articulation formulation is much simpler than previous NeRF-based approach that uses backward root-finding [7, 23, 30]. By adopting forward skinning, our method naturally supports out-of-distribution reposing.

Rendering. Each posed surfel $\mathbf{g}_t$ can be efficiently rasterized onto the image plane based on camera parameters $\boldsymbol{\pi}_t$. For example, RGB image $\hat{\mathbf{I}}_t$ can be rendered by

$$\hat{\mathbf{I}}_t(\mathbf{x}) = \sum_{i \in \mathcal{H}_t(\mathbf{x})} T_i \alpha_i \cdot \mathbf{c}_i, \tag{5}$$

where T_i and α_i are the transmittance and opacity of each projected 2D Gaussian surfel. $\mathcal{H}_t(\mathbf{x})$ is the set of surfels that intersect the ray originated from pixel $\mathbf{x}$. We can render mask $\hat{\mathbf{M}}_t$, depth map $\hat{\mathbf{D}}_t$, normal map $\hat{\mathbf{N}}_t$, and occlusion map $\hat{\mathbf{O}}_t$ similarly. This process is fully differentiable and allows end-to-end training from 2D observations.

3.3. Initial reconstruction

We start our optimization process by initial reconstruction, while reasoning about 3D occlusion of our model with respect to the input views.

We adopt both image supervision and structural priors from our preprocessed data for optimization. During each training iteration, we randomly sample a training view with its corresponding camera and SMPL-X parameters. Concretely, we seek to solve the following optimization problem:

$$\begin{aligned}
\min_{\{\boldsymbol{\mu}_0\},\{\mathbf{R}_0\},\Phi} & L_{\text{rgb}} + \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{normal}} L_{\text{normal}} + L_{\text{reg}}, \\
L_{\text{rgb}} &= 0.2 \cdot \|\mathbf{I}_t - \hat{\mathbf{I}}_t\|_1 + 0.8 \cdot \texttt{SSIM}(\mathbf{I}_t, \hat{\mathbf{I}}_t) + \texttt{LPIPS}(\mathbf{I}_t, \hat{\mathbf{I}}_t), \\
L_{\text{mask}} &= \|\mathbf{M}_t - \hat{\mathbf{M}}_t\|_1, \\
L_{\text{normal}} &= l_{\text{normal}}(\mathbf{N}_t, \hat{\mathbf{N}}_t) + l_{\text{normal}}({}^B\mathbf{N}_t, {}^B\hat{\mathbf{N}}_t) \\
l_{\text{normal}}(\mathbf{N}, \hat{\mathbf{N}}) &= 0.2 \cdot \mathbf{N}^T \hat{\mathbf{N}} + \texttt{LPIPS}(\mathbf{N}, \hat{\mathbf{N}}).
\end{aligned} \tag{6}$$

Similar to TeCH [20], we find that LPIPS [56] works with normal supervision and encourages crisp geometry over overly smoothed solution. To render back normal ${}^B\hat{\mathbf{N}}_t$, we rasterize by sorting surfels in descending depth order as opposed to usual ascending order. Using back normal supervision, the geometry of our avatar is constrained in unobserved regions.

We set $\lambda_{\text{mask}} = 1.0$ and $\lambda_{\text{normal}} = 1.0$ throughout our experiments.

Our regularization term L_{reg} consists of normal-depth consistency loss and curvature loss from [9], as well as an offset and scale regularization from [18] that penalizes irregular solution. This reconstruction process is trained with an Adam optimizer [26] for a total steps $K = 500$. The whole process takes about 5 minutes to finish.

As a side task, we are interested in estimating occlusion of our human model during the optimization process for quantify the portion of human body that has been observed from input video.

This is achieved by optimizing occlusion map $\hat{\mathbf{O}}_t$ in each training view per iteration, *i.e.*,

$$\min_{\{\tau\}} \|\hat{\mathbf{O}}_t\|_1. \tag{7}$$

Note that we detach all gradient from this objective towards other surfel properties such that we are only estimating the self-occlusion of our *current* geometry with respect to training views, without affecting the reconstruction process. We find that it is necessary to perform back-face culling [9] when rendering occlusion map. Without this operation, the occlusion signal "leaks" onto the back of the human figure.

3.4. Generative refinement

After initial reconstruction and occlusion estimation, we have a partially completed avatar. We next refine the initial result by score distillation sampling (SDS) a diffusion model [41].

In this work, we use ImageDream [48] as our diffusion prior. Empirically, we find this image-conditional multi-view diffusion model to be much more reliable compared to other alternatives, such as MVDream [44] or SD [42]. These alternatives rely heavily on text prompt and often produce overly saturated textures that are inconsistent with the original video. For example, TeCH [20] needs to finetune a SD model.

In addition to the set of losses in Equation 6, we sample a novel view camera $\tilde{\boldsymbol{\pi}}$ and render novel view $\tilde{\mathbf{I}}_t, \tilde{\mathbf{N}}_t$ during each training iteration for SDS supervision using the SMPL-X parameter in the current batch. The diffusion process is conditioned on image prompts $\mathbf{I}_t$, $\mathbf{N}_t$ and text prompt $\mathbf{p}$,

$$
\begin{aligned}
&\min_{\{\boldsymbol{\mu}_0\},\{\mathbf{R}_0\},\Phi} \lambda^{\text{sds}}_{\text{rgb}} L^{\text{sds}}_{\text{rgb}} + \lambda^{\text{sds}}_{\text{normal}} L^{\text{sds}}_{\text{normal}}, \\
L^{\text{sds}}_{\text{rgb}} &= \mathbb{E}_{i,\epsilon}\left[\left\|\tilde{\mathbf{I}}_t - \texttt{Denoise}_{\Psi}(\tilde{\mathbf{I}}_t; \mathbf{I}_t, \mathbf{p}, i, \epsilon)\right\|_2^2\right], \\
L^{\text{sds}}_{\text{normal}} &= \mathbb{E}_{i,\epsilon}\left[\left\|\tilde{\mathbf{N}}_t - \texttt{Denoise}_{\Psi}(\tilde{\mathbf{N}}_t; \mathbf{N}_t, \mathbf{p}, i, \epsilon)\right\|_2^2\right],
\end{aligned}
\tag{8}
$$

where $\texttt{Denoise}_{\Psi}$ denotes the a full denoising step from timestep i to 0 using noise ϵ with pretrained parameter Ψ. Please refer to ImageDream [48] for more detail. We first refine the shape for $K = 500$ iterations by setting $\lambda^{\text{sds}}_{\text{rgb}} = 0, \lambda^{\text{sds}}_{\text{normal}} = 10^{-4}$. We then refine the texture for $K = 1000$ iterations by setting $\lambda^{\text{sds}}_{\text{rgb}} = 10^{-4}, \lambda^{\text{sds}}_{\text{normal}} = 0$. The whole process takes about 20 minutes to finish.

4. Experiments

Our method is unique in being able to reconstruct self-occluded human from a single video. We first compare with HAVE-FUN [53] on its own benchmark. After carefully examining the actual occlusion in its evaluation, we find that large portion (~90%) of human body is observed during training. We then devise our own experimental setup to rigorously evaluate the performance of our approach on self-occluded videos, both quantitatively and qualitatively, discussed next.

4.1. Experimental setup

Datasets. While we primarily focus on reconstructing self-occluded humans in-the-wild, it is unpractical to quantitatively evaluate solely based on internet footage. Therefore, we show results on three types of dataset: FS-XHuman used by HAVE-FUN [53], a re-purposed multi-view human dataset and a set of internet footage of moving people. Concretely, we follow HAVE-FUN's evaluation split, which consists of 20 different subjects. We evaluate over few-view reconstruction given 2 views, 4 views and 8 views, respectively. After closer look, we find that FS-XHuman has very few occlusion and thus propose our own evaluation on DNA-Rendering dataset [8] due to its diversity and capture quality. DNA-Rendering comes with ground truth camera and SMPL-X annotations, making it suitable for fair comparisons between different approaches. Finally, we experiment with a set of in-the-wild videos from internet. These videos feature severe self-occlusion, fast motion and motion blur, making them much harder to reconstruct compared to the ones captured from a light stage. We use them to demonstrate the robustness of our method in the real-world scenario.

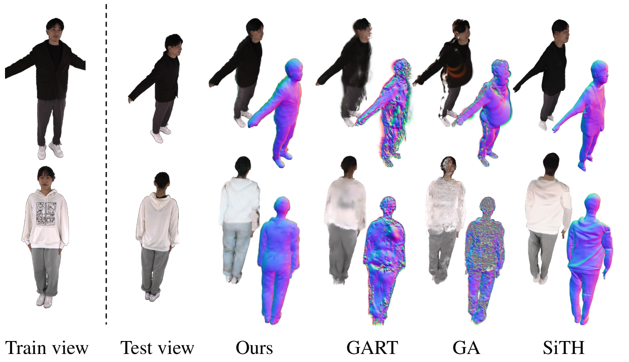

Figure 4. **Qualitative results on DNA-Rendering dataset.** For each training view, we visualize the ground-truth novel view along with predicted RGB rendering and normal map from different approaches. Our method recovers photo-realistic and geometrically plausible avatars comparing to baselines. For GART and GA, we read out their normals by depth gradient [9, 19, 24].

Metrics. For quantitative assessment, we evaluate novel view rendering on the FS-XHuman, DNA-Rendering and show qualitative comparisons on in-the-wild videos. Concretely, we evaluate standard PSNR, SSIM and LPIPS metrics from neural rendering literatures. In addition to that, we evaluate the rendering quality in occluded regions using mPSNR and mLPIPS [13, 21]. This evaluation shed light on how different approaches balance between reconstruction and generation. Since no ground-truth occlusion map exists in DNA-Rendering dataset, we use the inferred occlusion from our model for this task. We also propose a new metric called Body Occlusion Ratio (BOR) to quantify the portion of human body being seeing during training. It is computed by averaging inferred per-surfel occlusion τ for each training sequence.

In practice we sample 4 views for [48] and discuss one-view rendering for simplicity.

Baselines. We consider both reconstruction-based and generation-based methods as baseline. For reconstruction-based method, we evaluate against state-of-the-art Gaussian-based avatars including GART [29] and GaussianAvatar [18] (GA). For generation-based method, there exists no method that can handle video input. We therefore compare against recent human-specific image-to-3D approach SiTH [17] out of all candidates [2, 20, 57] due to its efficiency and code availability. For these methods, we run baselines on three randomly selected frame independently and repose using the input SMPL-X parameters. As one can expect, they are temporally inconsistent and have severe reposing artifacts due to the inability to properly handle articulation, which we show in Figure 5. The final quantitative metrics are averaged across these independently generated avatars.

4.2. Results on FS-XHumans dataset

To compare with concurrent work HAVE-FUN [53] that combines reconstruction and generation, we evaluate our method on FS-XHumans. Quantitative results are shown in Table 2 and qualitative results are included in supplement. Our method consistently outperforms HAVE-FUN in terms of PSNR and SSIM while on-par with it for LPIPS metric. We evaluate BOR for occlusion assessment in Table 3. As demonstrated, around ~90% human body is observed in FS-XHumans, even though its evaluation tries to work in few-view setting. Comparing to it, we propose a testing split from DNA-Rendering, which aligns closer to in-the-wild videos that have self-occlusion.

4.3. Results on DNA-Rendering dataset

DNA-Rendering dataset contains 500 human captures in a light stage setup. We choose 7 sequences without object interaction and loose clothing for our experiments. For each video, we train from a single camera and evaluate novel view rendering from 4 unseen cameras. We select the training camera as the one that the human is facing to in the first frame. With this simple rule, we already make sure that there are parts of human body remaining self-occluded throughout the video because the actors rarely orient on this dataset. For validation camera selection, we uniformly sample from the provided 60 cameras such that unobserved regions have ground-truth pixels.

We report our quantitative results in Table 1. Our proposed method outperforms all baselines on all metrics by a substantial margin. We separate our baselines into two categories: reconstruction-based methods GART and GA, generation-based method SiTH. When evaluating against the full image, denoted as "Full", our method improves significantly over baselines, with +1.2 PSNR and -10% LPIPS improvements comparing to reconstruction-based methods. This improvement is even larger comparing to generation-based baseline SiTH in terms of PSNR by +5.3, with similar -10% LPIPS improvement. It is perhaps not surprising giving that PSNR favors exactness over realism while LPIPS does the other way around. It is also notable that our approach dramatically improves over the existing approaches in both visible and occluded regions (noted as "Visible" and "Occlusion" in the table).

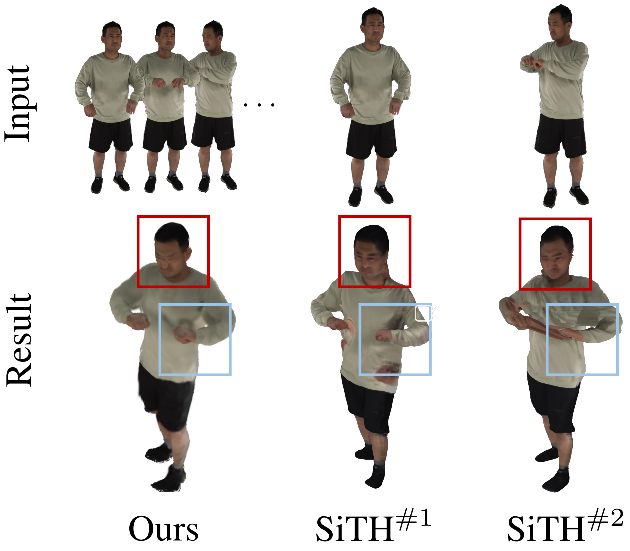

Figure 5. **Comparison between our globally consistent avatar and image-to-3D baseline.** Our method is able to fuse all observations from a video and allow natural reposing.

We visualize our qualitative comparisons in Figure 4 with both novel view RGB rendering and normal map predictions. For GART and GA, we read out their normals by depth gradient [9, 19, 24]. We want to emphasize our improvements in three aspects. First, our method produces crisp geometry details as suggested by predicted normal maps. Second, it is clear to see that our method produces highly realistic synthesis on the unobserved regions, *e.g.*, around the back regions in the second row. Reconstruction-based methods struggle in these under-constrained areas. Third, our reconstruction component helps prevent unnatural shapes – this is evident in the first row where SiTH produces very thin arms.

Finally, we demonstrate that our approach benefits from a globally consistent representation for avatar creation such that we can fuse observations from multiple video frames. We visualize our reposed avatar in Figure 5, comparing to two different runs by SiTH, which can only condition on a single input frame. The note three observations. First, fusing observations from multiple frames by reconstruction resolves ambiguity in seen regions (in blue bounding box), where SiTH strugle to produce accurate shape solely based on image prior. Second, SiTH produces inconsistent results between runs as shown in the red bounding box. Third, our globally consistent avatar representation allow much more

Method	Type	Full		Visible		Occlusion	
		PSNR↑	LPIPS↓	mPSNR↑	mLPIPS↓	mPSNR↑	mLPIPS↓
SiTH [17]	Gen.	17.84	0.065	14.55	0.282	9.63	0.271
GA [18]	Recon.	21.97	0.068	15.28	0.456	10.06	0.492
GART [29]		21.89	0.067	16.33	0.290	12.86	0.283
Ours	Gen. + Recon.	**23.16**	**0.055**	**19.96**	**0.227**	**14.54**	**0.253**

Table 1. **Quantitative results on DNA-Rendering dataset**. We evaluate the novel view rendering performance of all approaches in full image ("Full"), visible regions ("Visible") and occluded regions ("Occlusion"). Our method consistently out-performs different baselines in all metrics by a significant margin. Best metrics are marked as bold.

Method	#Input	PSNR↑	SSIM↑	LPIPS↓
SelfRecon [22]	8-shots	19.90	0.927	0.065
	100-shots	20.70	0.943	0.063
TeCH [20]	1-shot	21.00	0.924	0.065
HaveFun [53]	2-shots	24.00	0.955	**0.042**
	4-shots	25.60	0.963	**0.035**
	8-shots	26.80	0.967	**0.030**
Ours	2-shots	**25.18**	**0.958**	0.049
	4-shots	**26.86**	**0.964**	0.043
	8-shots	**27.94**	**0.968**	0.039

Table 2. **Comparison on FS-XHumans dataset**. We compare our methods with few-shot human reconstruction methods. Our method consistently outperforms different baselines in PSNR and SSIM metrics by a significant margin. Best metrics in 2-view/4-view/8-view settings are marked as bold.

Dataset	BOR
FS-XHumans [53]	0.202/0.138/0.095
DNA-Rendering [8]	0.460
In-the-wild	0.341

Table 3. **BOR value of different datasets**. FS-XHumans have 2-view/4-view/8-view evaluation. However, only around ~10% of human body remains unobserved. In comparison, our proposed DNA-Rendering split aligns closer to self-occluded videos in-the-wild.

Method	Full		Visible		Occlusion	
	PSNR↑	LPIPS↓	mPSNR↑	mLPIPS↓	mPSNR↑	mLPIPS↓
W/o sds	21.47	0.065	17.14	0.275	9.78	0.441
W/o implicit	22.86	0.066	17.74	0.275	11.01	0.455
Ours w/ occ	**23.58**	**0.055**	17.91	0.256	**12.15**	**0.324**
Ours	23.51	**0.055**	**18.28**	**0.247**	11.93	**0.324**

Table 4. **Ablation results on DNA-Rendering dataset**. We ablate the novel view rendering performance of all approaches in full image ("Full"), visible regions ("Visible") and occluded regions ("Occlusion"). Our method consistently out-performs different baselines in all metrics by a significant margin. Best metrics are marked as bold.

natural reposing comparing to SiTH mesh skinning.

4.4. Results on in-the-wild videos

We report the qualitative results of our method applied to in-the-wild videos, as shown in Figure 6. Our dataset comprises single human-centered internet videos with severe self-occlusion. The first row of the figure displays the novel view rendering results. We conducted comparisons with methods GART and GA, which often fail to accurately reconstruct human shape and texture under these challenging conditions. In contrast, our method consistently produces high-detail normal maps and realistic textures. Additionally, we include our reposing results to further demonstrate the robustness of our approach. Our method again is able to produce photo-realistic rendering and accurate shape prediction.

4.5. Ablation

We conducted ablation study about our design choices and consider following baselines: (1) our model with occlusion masking in SDS loss as discussed in Section 3.4 ("Occ. SDS"); (2) our model without the generation component ("No SDS"); (3) our model without implicit parameterization Φ ("No Φ"). Qualitative results are shown in Figure 7 and quantitative results are included in Table 4. Each component is beneficial and contributes to our full model.

5. Discussion and Conclusion

While we present promising steps towards robust human avatar recovery from in-the-wild videos several limitations remain. It inherits the issue of generating saturated colors from SDS-based methods, remains a test-time optimization approach limiting interactive use, and lacks a comprehensive in-the-wild dataset with ground-truth multi-view annotations

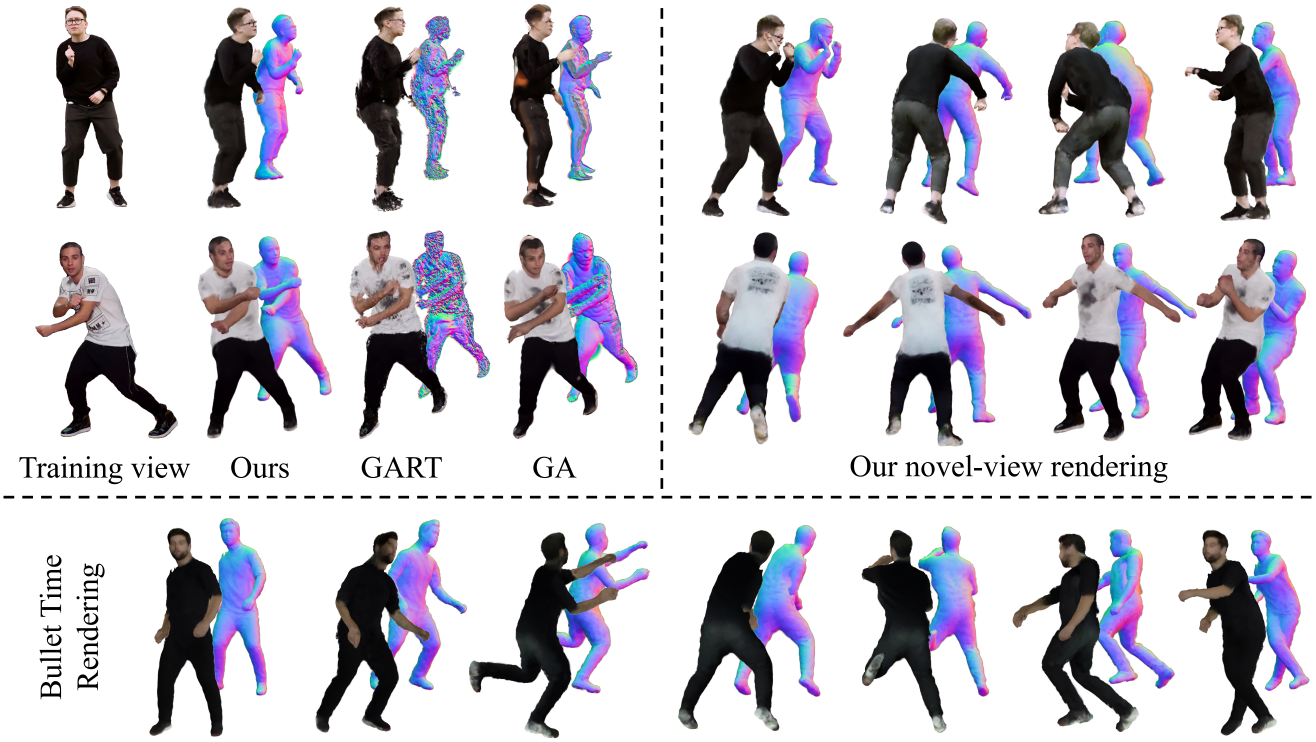

Figure 6. **Qualitative results on in-the-wild videos.** We visualize novel-view rendering comparison in top left, our 360 rendering on top right, and our bullet time rendering on the bottom. We visualize both the RGB rendering and normal map rendering in each result.

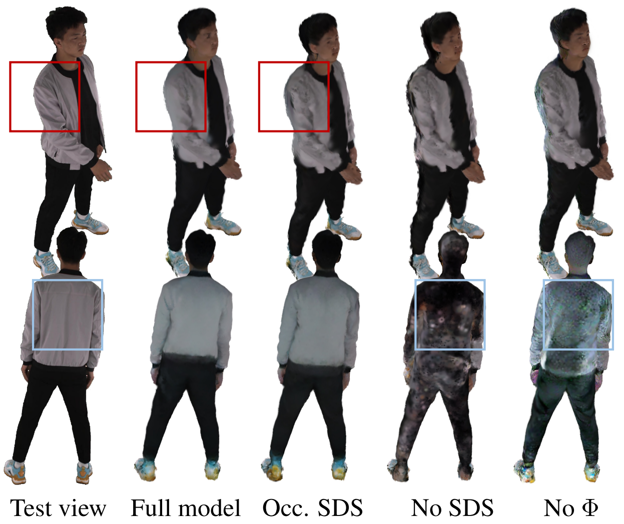

Figure 7. **Ablation.** We ablate over occlusion masking in SDS, generation itself and our implicit parameterization Φ.

for better evaluation. Future work includes training human-specific multi-view diffusion models on large-scale human capture data and creating an in-the-wild human dataset with multi-view validation. Despite these limitations, we presented SOAR for self-occluded avatar recovery from a single in-the-wild video, employing a globally-consistent surfel model for fusing noisy supervision and reposing, and leveraging structural human normal priors and generative diffusion priors. Our method recovers photo-realistic avatar models with plausible shapes, significantly improving over existing methods. Experiments on multi-view datasets and in-the-wild videos demonstrate that our method achieves state-of-the-art performance compared to purely reconstruction-based and generation-based methods.

6. Acknowledgement

This project is supported in part by DARPA No. HR001123C0021, IARPA DOI/IBC No. 140D0423C0035, NSF:CNS-2235013, Bakar Fellows, and Bair Sponsors. The views and conclusions contained herein are those of the authors and do not represent the official policies or endorsements of these institutions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d hu-

man digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2, 6

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[5] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM TOG*, 2003. 2

[6] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 3

[7] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 2, 3, 4

[8] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 2, 5, 7

[9] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. *arXiv preprint arXiv:2404.17774*, 2024. 2, 3, 4, 5, 6

[10] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM TOG*, 2008. 2

[11] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009. 2

[12] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. 2022. 1

[13] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3985–3993, 2017. 5

[14] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 4:5–21, 1987. 3

[15] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, 2024. 2

[16] 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 (CVPR)*, 2023. 2, 4

[17] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. *arXiv preprint arXiv:2311.15855*, 2023. 2, 6, 7

[18] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 6, 7

[19] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2403.17888*, 2024. 2, 5, 6

[20] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *3DV*, 2023. 2, 4, 5, 6, 7

[21] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 17–34. Springer, 2020. 5

[22] 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. 7

[23] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 1, 2, 3, 4

[24] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. *arXiv preprint arXiv:2311.17977*, 2023. 2, 5, 6

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[28] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[29] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. *arXiv preprint arXiv:2311.16099*, 2023. 2, 3, 6, 7

[30] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 1, 2, 3, 4

[31] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 2

[32] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. *arXiv preprint arXiv:2311.16473*, 2023. 2

[33] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 2

[34] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 3

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[36] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 4

[38] 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. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[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 *CVPR*, 2021. 1, 2

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 4, 5

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

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

[44] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 5

[45] Carsten Stoll, Juergen Gall, Edilson De Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. *ACM TOG*, 2010. 2

[46] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. *arXiv preprint arXiv:2404.01053*, 2024. 2

[47] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[48] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 5

[49] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[50] 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

[51] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2, 3

[52] 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. 2

[53] Xihe Yang, Xingyu Chen, Daiheng Gao, Shaohui Wang, Xiaoguang Han, and Baoyuan Wang. Have-fun: Human avatar reconstruction from few-shot unconstrained images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 742–752, 2024. 2, 5, 6, 7

[54] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 2

[55] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. *arXiv preprint arXiv:2312.11461*, 2023. 2

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 586–595, 2018. 4

[57] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. *arXiv preprint arXiv:2312.06704*, 2023. 2, 6

[58] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. Physavatar: Learning the physics of dressed 3d avatars from visual observations. *arXiv preprint arXiv:2404.04421*, 2024. 1

[59] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 1, 2, 3