Rendering Humans from Object-Occluded Monocular Videos

Tiange Xiang*, Adam Sun, Jiajun Wu, Ehsan Adeli, Li Fei-Fei
Stanford University

Abstract

*3D understanding and rendering of moving humans from monocular videos is a challenging task. Despite recent progress, the task remains difficult in real-world scenarios, where obstacles may block the camera view and cause partial occlusions in the captured videos. Existing methods cannot handle such defects due to two reasons. First, the standard rendering strategy relies on point-point mapping, which could lead to dramatic disparities between the visible and occluded areas of the body. Second, the naive direct regression approach does not consider any feasibility criteria (*i.e.*, prior information) for rendering under occlusions. To tackle the above drawbacks, we present* ***OccNeRF****, a neural rendering method that achieves better rendering of humans in severely occluded scenes. As direct solutions to the two drawbacks, we propose surface-based rendering by integrating geometry and visibility priors. We validate our method on both simulated and real-world occlusions and demonstrate our method's superiority. Project page:* `https://cs.stanford.edu/~xtiange/projects/occnerf/`

1. Introduction

Rendering 3D human bodies from a sequence of observations is of great interest in various communities, including robotics [70], motion analysis [16], and healthcare [19]. This task is challenging, since one must recover the complete human body with complex textures and poses from sparse partial observations. It is usually cumbersome to acquire images of the same human object from multiple camera angles simultaneously; hence, capturing a monocular video from a single camera is more common and feasible.

The task of rendering humans from a monocular video is not new. Progress so far mainly focuses on rendering quality [51, 66] and rendering efficiency [49, 28]. However, most existing neural rendering methods assume that the human object is placed in a scene with a clear view of the entire body without any external interference. In contrast, real-world environments often contain undesired obstacles that contaminate training data and impact the rendering quality (See Figure 1). These real-world occlusions pose significant challenges for training when using only monocular videos, where no other camera angles can be used to provide complementary information. As a result, a direct application of previous neural rendering methods on object-occluded videos leads to subpar performance. Optimizing a neural radiance field is difficult under occlusions. There is often no ground truth associated with the occluded area. Additionally, radiance fields are typically optimized in a scene-specific manner; that is, no external information can and should be used to fill in the occluded areas.

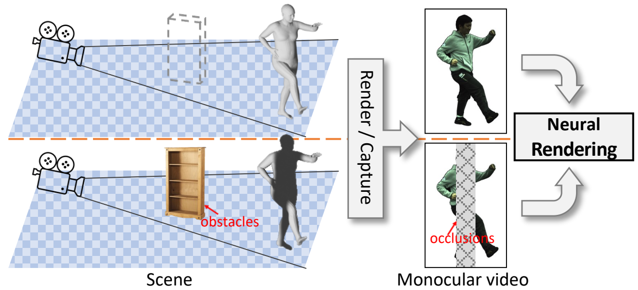

Figure 1. Object obstacles in the scene may cause severe occlusions in the rendered/captured videos, imposing additional challenges into the rendering process. **Top row:** Ideal scene with no defects and clear view of the body; **Bottom row:** Real-world scene with undesired obstacles and occluded body parts.

Two major drawbacks of previous methods impair their capabilities to train on object-occluded videos. First, the prior work does not account for local geometry cues in their rendering process. Following the point-based rendering paradigm as in NeRF [43], most previous methods render color and density values of a ray sample by only looking at a single 3D coordinate. However, we explain in section 3.2 that this basic strategy may lead to dramatically different rendering results even in very close positions. Second, methods suffer from not properly incorporating priors. In the monocular video setting, geometry (*e.g.*, SMPL [40]) and visibility priors can describe a complete human geometry and indicate which body parts are visible to the camera.

In this work, we propose novel methods for dealing with the above drawbacks, allowing us to accurately render occluded humans from monocular video. We first present a surface-based rendering strategy that determines the radiance of each 3D ray sample by conditioning it on a wide re-

*Correspondence to xtiange@stanford.edu

gion of the human body's surface. A geometry prior is used to discretely parameterize the surface segments. We then collect visibility frequencies on the human body through training frames and formulate them as attention maps for better aggregation of the surface regions. Finally, we design a loss function to encourage the network to output high-density values for positions within the human body.

In summary, our contributions are three-fold: ***(i)*** We are the first to study dynamic human rendering under real-world settings with severe occlusions. ***(ii)*** We propose novel methods that include surface-based rendering, a reformulation of body part visibility frequency as attention, and a completeness loss to enable human rendering from object-occluded monocular videos. ***(iii)*** We empirically demonstrate that our methods achieve significant quantitative and qualitative improvements compared to the previous state-of-the-art, yielding the first baseline in this topic.

2. Related Work

3D Human Modeling. Reconstructing the appearance and geometry of humans has always been challenging. From [42, 12, 15], techniques have been consistently designed for high-quality and efficient human modeling. Traditional methods mainly relied on SMPLify [6] or Video-avatars [1] to regress SMPL [40] to parameterize a structured human body. More complex networks were subsequently designed that can model 3D humans based on temporal priors [31, 33], based on depth [27, 55, 23], or multiple human instances simultaneously [29, 59, 60, 71]. Although this line of methods can generate a reasonable human body mesh fast, using parametric SMPL models limit their ability to achieve photo-realistic view synthesis.

Neural Radiance Field for Human Rendering. Since the emergence of Neural Radiance Fields (NeRF) [43], different extensions have been recently developed to enable high-quality rendering of static scenes [22, 57, 2, 3, 63, 61, 58, 44], moving objects [18, 36, 47, 48, 52, 46], and dynamic humans [51, 4, 7, 11, 9, 13, 14, 17, 20, 21, 26, 27, 35, 37, 45, 50, 62, 64, 68, 49, 28, 30]. NeRF predicts the color and density of each ray sample point in a 3D space and aggregates them together through volume rendering (more details are in section 3.1). This approach enables the capture of intricate lighting effects and textural details that are typically difficult to model in traditional methods.

Our work is built upon HumanNeRF [66] due to its state-of-the-art rendering quality for monocular videos. HumanNeRF maintains a static T-pose human body as the canonical space and learns a motion field [65] that maps the canonical representation to every frame of the video in the observation space (more details are in section 3.1). We note that a concurrent work, SelfNeRF [49], shares a similar regression schema as ours. However, their method is designed particularly for fast rendering and compromises rendering quality. Moreover, all of the above approaches were developed on clean training data only, where body parts are assumed to be clearly demonstrated in the monocular video without any occlusions. On the other hand, our work aims to render humans under occlusions.

Occluded Human Modelling. Rendering objects under all kinds of real-world defects, especially partial occlusions, is a long-standing research problem. Early works sought to estimate human poses from occluded images and videos [56, 73, 34, 5], while more recent works [54, 59, 69, 34] learn both SMPL shape and pose priors directly from occluded images and videos. The generated SMPL parameters from these robust methods can be used as good geometric prior for a subsequent rendering process.

However, optimizing a NeRF from occluded images is still an unsolved problem. There are very few works that were specifically designed for rendering under occlusions. NeuRay [38] was proposed to regress not only the radiance but also a feature vector of every ray sample to indicate visibility. This enables the optimization of the radiance field to focus on visible features and reduce interference from occlusions. Ha-NeRF [10] presents an appearance hallucination module to handle time-varying appearances and an anti-occlusion module to decompose the static subjects for visibility accurately. Unfortunately, these existing methods are not capable of handling dynamic objects, and the multi-view inputs used in past work actually make it easier to learn under occlusions. In this work, we consider visibility as an additional prior to assist in rendering under occlusions. Our work is the very first in this field to handle occlusions for rendering dynamic objects from only a monocular video.

3. Methods

In this section, we first review preliminaries and background in NeRF [43] and HumanNeRF [66] (section 3.1). We then present our OccNeRF by introducing a new rendering strategy (section 3.2), a formulation of visibility into attention (section 3.3), and a novel loss function (section 3.4) to ensure high rendering quality as well as geometry completeness under occlusions. An overview of our OccNeRF is shown in Figure 2.

3.1. Preliminaries and Background

Neural Radiance Field [43]. Consider a (bounded) 3D scene. NeRF learns a regression function $\mathcal{F}$ (usually an MLP) that takes the encoded coordinates of a 3D point $\mathbf{x} \in \mathbb{R}^3$ in the scene as input, and outputs the corresponding color $\mathbf{c}$ and density σ at that position:

$$\mathbf{c}, \sigma = \mathcal{F}(\gamma(\mathbf{x})), \tag{1}$$

where $\gamma(\cdot)$ is an encoding function. We refer to the above point-to-point mapping as **point-based rendering**. Instead

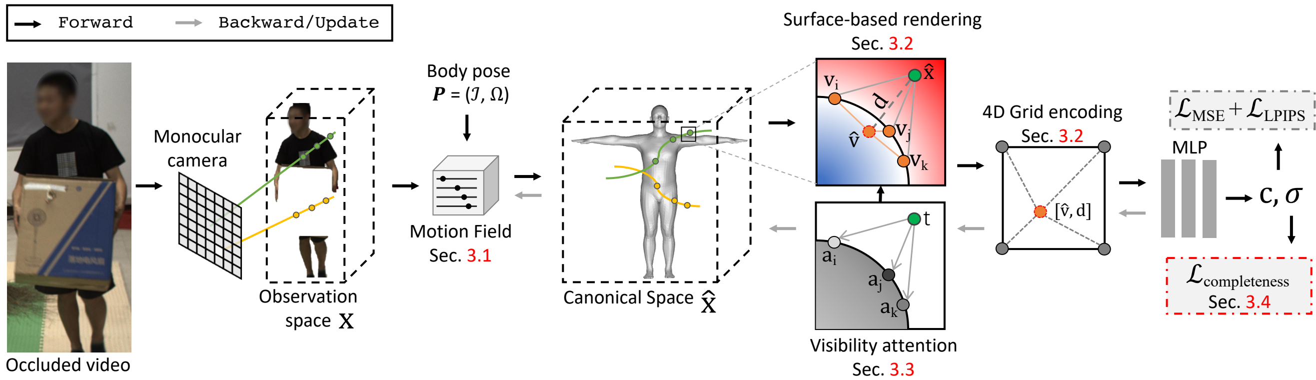

Figure 2. **OccNeRF** functions on video frames and optimizes a neural radiance field for synthesizing novel views of an object-occluded human. With a pre-computed body pose, we first adopt the motion field to map observable ray samples $\mathbf{x}$ into coordinates $\hat{\mathbf{x}}$ in a canonical space. Nearest parameterization vertices $\{\mathbf{v}_i\}$ of every $\hat{\mathbf{x}}$ are searched and conditioned by our surface-based rendering method. During training, we iteratively update the attention scores $\{\mathbf{a}\}$ for all $\{\mathbf{v}\}$ as indications of their visibility. This ensures more attention on frequently visible vertices to improve rendering quality. The blended vertex $\hat{\mathbf{v}}$ along with its signed distance to $\hat{\mathbf{x}}$ are jointly encoded via a 4D hash grid before being fed into the regression MLP along with the encoded vertices. Photometric and perceptual constraints are enforced against visible pixels, while an additional loss function is designed to encourage geometry completeness in occluded areas.

of sampling points $\mathbf{x}$ randomly in the scene, NeRF casts rays $\mathbf{r}$ towards the directions π from the camera origin $\mathbf{o}$ to every pixel, and sample $\mathbf{x}$ on the rays uniformly. Then, NeRF renders the pixel by aggregating the regressed color and density at each $\mathbf{x}$ via volume rendering [39]:

$$\sum_i \alpha(\mathbf{x}_i) \prod_{j<i} (1-\alpha(\mathbf{x}_j))\mathbf{c}, \tag{2}$$

where $\alpha(\mathbf{x}_i) = 1 - \exp(-\sigma_i \delta_i)$, $\mathbf{x}_i = \mathbf{o} + z_i \pi$, z_i is the z-axis position of ray samples, and $\delta_i = z_{i+1} - z_i$ is the distance between two samples along the ray.

HumanNeRF [66]. HumanNeRF is a method based on NeRF that can render humans from monocular videos by representing them as neural fields. The method first defines a moving human in a static canonical space with 3D coordinates $\hat{\mathbf{x}}$, and warps the human in different dynamic poses by warping the canonical body pose $\mathbf{p}$ to the observation space. This warping process also defines the transformation of 3D coordinates in the two spaces:

$$\hat{\mathbf{x}} = \mathcal{T}(\mathbf{p}, \mathbf{x}), \tag{3}$$

where $\mathcal{T}$ is a network that maps $\mathbf{x}$ in the observation space to corresponding coordinates $\hat{\mathbf{x}}$ at the canonical space, denoted as the **motion field**. The motion field achieves the mapping by performing a weighted sum of a set of K motion bases defined by rotations R_i and translation t_i of the i_{th} bone of the human body:

$$\hat{\mathbf{x}} = \sum_i^K w_i(\mathbf{x})(R_i \mathbf{x} + t_i), \tag{4}$$

where R_i and t_i can be directly computed from $\mathbf{p}$. w_i serves as the weights in the observation space, which can be approximated using the weights defined in the canonical space. Similar to [67], we removed both the non-rigid motion and the pose correction part of the motion field.

3.2. Surface-based Rendering

Motivation. Although HumanNeRF and its variants can already achieve decent rendering quality in an occlusion-free scene, they fail to excel when obstacles block the view of the camera that causes severe occlusions. We attribute this failure to the point-based rendering strategy (reviewed in section 3.1). Given the ray samples at discrete 3D coordinates $\mathbf{x}$ in a continuous 3D space, even a mild variation between two coordinates $\mathbf{x}_a$ and $\mathbf{x}_b$ can lead to dramatic disparities on the outputs. Let there be no overlaps between the input distributions $\{\mathbf{x}_a\}$ and $\{\mathbf{x}_b\}$:

$$\{\mathbf{x}_a\} \cap \{\mathbf{x}_b\} = \emptyset \mid \mathbf{x}_a \not\equiv \mathbf{x}_b. \tag{5}$$

Then, in an occluded scene, when only $\mathbf{x}_a$ is visible to the camera, non-overlapping inputs may yield huge output differences, even at very close locations. This is because $\mathbf{x}_a$ has visible supervisions while $\mathbf{x}_b$ does not, which leads to unexpected artifacts and unstable rendering quality at occluded regions.

This motivates us to enlarge the range of the inputs to cover a wider range of 3D space rather than a single 3D coordinate. We expect that a new rendering strategy with range-to-point mapping will be able to reduce the output difference at adjacent locations:

$$\int_{\mathbb{R}^3} \min[\mathcal{N}(\mathbf{x}_a), \mathcal{N}(\mathbf{x}_b)] d\mathbf{x} \gg 0, \tag{6}$$

where $\mathcal{N}(\mathbf{x}_a)$ and $\mathcal{N}(\mathbf{x}_b)$ are 3D sub-regions corresponding to the target coordinates $\mathbf{x}_a$ and $\mathbf{x}_b$. With a focus on

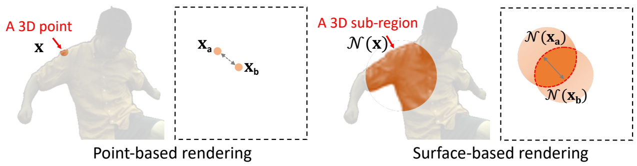

Figure 3. **Left:** Point-based rendering takes as input a single 3D point $\mathbf{x}$ that has no overlap with nearby (but not identical) points, and is poorly conditioned at occluded areas; **Right:** Our surface-based rendering approach takes as input a 3D sub-regions $\mathcal{N}(\mathbf{x})$ at location $\mathbf{x}$ that yields a large overlap at adjacent locations for better conditioning at occluded areas.

human rendering, we define the sub-regions as continuous segments on the body surface. We name this rendering strategy **surface-based rendering**. A high-level comparison to the standard point-based rendering is outlined in Figure 3.

Parameterization. It is difficult to process continuous variables, especially ones with irregular distributions, which is the case with human surfaces. We approach this challenge by using a discretized parameterization of the continuous 3D sub-regions. Specifically, we use the pre-computed SMPL [40] mesh as a geometry prior to roughly outline the surface of the human body. The surface segments are then parameterized by the k nearest mesh vertices $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ when using the target coordinates $\mathbf{x}$ as queries. We denote these discrete neighboring coordinates as parameterization vertices. With our surface-based rendering approach, we now reformulate Equation 1 as a hybrid combination of both the target coordinate and the parameterized surface:

$$\mathbf{c}, \sigma = \mathcal{F}(\underbrace{\gamma(\hat{\mathbf{x}})}_{\text{point term}} \parallel \underbrace{\phi(\{\gamma(\mathbf{v}_1), \cdots, \gamma(\mathbf{v}_k)\})}_{\text{surface term}}), \tag{7}$$

where ϕ is a function that aggregates all $\{\mathbf{v}_i\}$ of a query $\mathbf{x}$ and $\parallel$ denotes concatenation. The above formulation requires all parameterization points $\mathbf{v}$ to be as accurately laid on the human body as possible. However, this is difficult for the coarsely structured SMPL mesh with potential approximation errors. Therefore, we rely on the inaccurate SMPL mesh only as an initialization and enable the positions of $\mathbf{v}$ to be optimized jointly with the network. This formulation is analogous to area sampling [41] for ray tracing, which not only integrates samples along the ray but in vicinity area.

Multi-Scale Representations. Choosing the area of surface segments and the number of parameterization vertices k per query is another issue. A small area leads to less overlap and more unstable results, while a large area leads to more overlap of $\{\mathbf{v}_i\}$ at two query locations but a more inefficient search of nearest neighbors. Taking inspiration from the multi-scale mechanism used in I-NGP [44], we construct the set of parameterization vertices by finding the nearest neighbors on the SMPL mesh at multiple scales. We define the default SMPL mesh at the finest scale and iteratively down-sample the mesh with sparse vertices through furthest point sampling [53] with a ratio of 0.25 for 3 iterations. In practice, we set a small $k = 5$ at all 4 scales, which enables an efficient span over a large surface area.

Surface-Aware Regression. The additional operations introduced above inevitably slow down network training. Similar to [49, 28], we adopt a hash grid [44] in the canonical space as our encoding function $\gamma(\cdot)$ instead of the standard frequency-based positional encoding [43]. Furthermore, for better awareness of the human body surface, we represent a query point in the canonical space $\hat{\mathbf{x}}$ by the combination of its closest parameterization vertex $\hat{\mathbf{v}}$ and their signed distance $\mathbf{d}$. For simplicity, we reuse the previously searched k nearest vertices and blend them through normal similarities to form the closest vertex $\hat{\mathbf{v}}$:

$$\hat{\mathbf{v}} = \frac{\sum_i^k |\cos(\hat{\mathbf{x}}, \mathbf{v}_i)| \mathbf{v}_i}{\sum_i^k |\cos(\hat{\mathbf{x}}, \mathbf{v}_i)|}, \tag{8}$$

where $\cos(\hat{\mathbf{x}}, \mathbf{v}_i)$ denotes the cosine similarity between the vector $\hat{\mathbf{x}} - \mathbf{v}_i$ and the normal vector at $\mathbf{v}_i$. After obtaining $\hat{\mathbf{v}}$, we can easily determine $\mathbf{d}$ between $\hat{\mathbf{v}}$ and $\hat{\mathbf{x}}$ via a multiplication with the normal vectors at $\mathbf{v}_i$. Inspired by [49], we then rely on a 4D hash grid to encode the combination $[\hat{\mathbf{v}}, \mathbf{d}]$. Note that our formulation differs from [49], which encodes a 4D feature vector for every nearest neighbor first and then blends the feature vectors afterward. Our implementation encodes every $\hat{\mathbf{x}}$ only once.

With the above formulation, we can rewrite the point term $\gamma(\hat{\mathbf{x}})$ in Equation 7 into $\gamma([\hat{\mathbf{v}}, \mathbf{d}])$. The surface term is formulated with visibility priors, as discussed below.

3.3. Visibility Attention

In occluded videos, some parts of the human body may be more frequently visible by the camera than others. As a result, more supervision is provided for these highly visible parts which makes $\mathcal{F}$ fit on these *visible areas* much better. When conditioning on a wide range of surfaces, we hope to pay more attention to the highly visible vertices than the hardly visible ones. We achieve this through an attentive aggregation of the neighbor vertices $\{\mathbf{v}_i\}$ via the function ϕ (Equation 7) based on their visibility frequency.

Specifically, for each of the vertices $\mathbf{v}_i$, we maintain a separate attention score $\mathbf{a}_i$ to be updated on-the-fly as the training proceeds. Instead of recording the visibility frequency of all sample points in the camera rays, only the termination point $\mathbf{t}$ per ray should be considered. However, it is computationally expensive to find the exact intersection point between the camera rays and the human body. We approximate $\mathbf{t}$ as the sample point with the highest α along each of the rays, such that $\mathbf{t} = \hat{\mathbf{x}}_{\arg\max\{\alpha\}}$. For each $\mathbf{t}$, we again rely on the k nearest vertices $\{\mathbf{v}_i\}$ found earlier to determine the visible area on the body. At each training step, for all neighbors $\{\mathbf{v}_i\}$ of every $\mathbf{t}$, we increment their

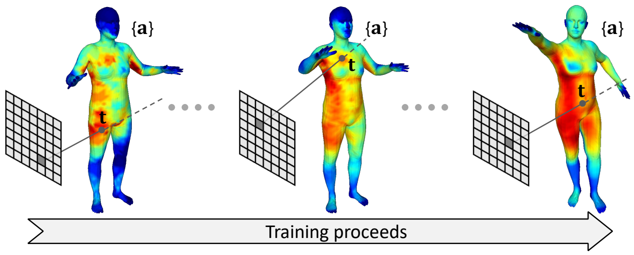

Figure 4. **Formulating visibility as attention**. Highly visible body parts along with associated parameterization vertices are expected to correspond to more attention.

corresponding attention scores $\{\mathbf{a}_i\}$ by 1. Taking visibility into account, Equation 7 can be reformulated as:

$$\mathbf{c}, \sigma = \mathcal{F}(\underbrace{\gamma([\hat{\mathbf{v}}, \mathbf{d}])}_{\text{point term}} \parallel \underbrace{\frac{\sum_i^k \mathbf{a}_i \gamma([\mathbf{v}_i, \mathbf{d}^{(\mathbf{v}_i)}])}{\sum_i^k \mathbf{a}_i}}_{\text{surface term}}), \quad (9)$$

where the hash grid encoding $\gamma(\cdot)$ is shared for both point and surface. Recall that all vertices $\mathbf{v}$ have learnable coordinates, and we compute their signed distance $\mathbf{d}^{(\mathbf{v}_i)}$ w.r.t the vertices on the initial SMPL mesh. The updating process of our visibility attention is demonstrated in Figure 4.

3.4. Loss Functions

Following HumanNeRF, we mainly supervise the training of OccNeRF through pixel-wise photometric loss $\mathcal{L}_{\text{MSE}}$ and LPIPS [72] loss $\mathcal{L}_{\text{LPIPS}}$ to encourage high-quality rendering at the visible parts. Unfortunately, these constraints do not apply to the occluded parts, where supervisions are hardly available. Hence we design another constraint to explicitly penalize renderings with incomplete geometry and encourage high-density values within the human body. The previously computed signed distances $\mathbf{d}$ are good approximations of the position of ray samples w.r.t the SMPL mesh. Instead of only enforcing the samples near the body surface, we apply the constraint to all samples with negative $\mathbf{d}$. Our completeness loss $\mathcal{L}_{\text{comp}}$ is therefore defined as:

$$\mathcal{L}_{\text{comp}} = m \cdot \exp(\texttt{ReLU}(-\texttt{ReLU}(\sigma) + \beta) - \beta), \quad (10)$$

where $m = 1$ if $\mathbf{d} < 0$ and 0 otherwise, and $\beta = 10$ is a hyper-parameter. Intuitively, it is designed to penalize incompleteness inside the human body. We use ReLU to clip negative σ in the range of $[-\beta, 0]$ and use exponential trick to decrease penalty for high densities. OccNeRF is supervised by a weighted combination of the three losses:

$$\lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{LPIPS}} + \lambda_3 \mathcal{L}_{\text{comp}}. \quad (11)$$

4. Experiments

4.1. Datasets

ZJU-MoCap [51]. This dataset contains humans performing a wide variety of activities. Following HumanNeRF [66], we mainly evaluate our methods on the 6 subjects (377, 386, 387, 392, 393, 394) for direct comparisons. Videos captured by *camera 1* are used as training data, and the other 22 cameras are used for evaluation. Since the MoCap data was captured in a lab environment without the interference of any obstacles, we simulate occlusions to be applied to the training videos. Without losing generality, we simulate the presence of a box-like obstacle right in front of the camera that causes a rectangular area at the center of the frame to be occluded. To do so, we first determine a center point of the valid pixels from video frames, and then mask out 50% of these pixels (demonstrated in Figure 1). Our simulated obstacle and the occluded area are not intended to be moving along with the subject. Since there is no obvious horizontal movement of subjects, we further expect that they can move out of the occluded area for a short time and therefore only apply the mask to 80% of the frames.

OcMotion [24]. This dataset contains humans interacting with various objects, subject to real-world occlusions. There are a total of 48 videos, and each video was captured at 6 different camera poses. We evaluated on 2 videos with different extents of occlusions. Specifically, we selected 540 frames from *video 14, camera 4* and 500 frames from *video 11, camera 2* as benchmarks for **mild** and **severe** real-world occlusions respectively. For both benchmarks, we use the camera matrices, human body poses, and SMPL parameters provided by the dataset, which were computed by [25] directly on the occluded videos. We provide more results in supplementary materials. We also show the robustness of our method to inaccurately estimated priors.

4.2. Comparison and Metrics

We mainly compare our method with HumanNeRF [66], the state-of-the-art human rendering method. We also compare against a baseline method Neural Body [51] in supplementary materials. Note that all methods use identical prior information, including pre-computed binary human mask and SMPL/camera parameters. The extra visibility prior used in OccNeRF can be calculated from the videos.

Methods are compared qualitatively and quantitatively. For qualitative evaluations, we directly visualize novel views to assess the quality of the renderings. For quantitative evaluations, we rely on the commonly used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics [51, 66, 49]. Previous methods computed these metrics on full-scale images, which contain a majority of transparent backgrounds. These regions are identical between predictions and references, which inflate the overall metrics. To focus on the quality of rendered humans, we compute metrics on the pixels with non-zero accumulated α. For OcMotion, since there is no ground truth for real-world occlusions, we compute the metrics on the visible area only. We refer to the standard metrics as $\text{PSNR}_{\text{full}}$/$\text{SSIM}_{\text{full}}$ and

Figure 5. Qualitative results on **simulated occlusions** in the ZJU-MoCap dataset [51].

modified metrics as PSNR_{vis}/SSIM_{vis}.

4.3. Implementation Details

Using the loss formulated in Equation 11, we optimize OccNeRF with the Adam optimizer [32]. We set the learning rate to 5×10^{-4} for the regression MLP $\mathcal{F}$, 1×10^{-4} for the parameterization vertices $\mathbf{v}$, and 5×10^{-5} for the rest. λ_1, λ_2, and λ_3 were set to 0.2, 1.0, and 10.0 respectively. We adopted patch-wise sampling of rays, each with 128 sample points. Due to the usage of the hash grid, OccNeRF converges faster than HumanNeRF. As a result, we trained our models for only 10K iterations while Hu-

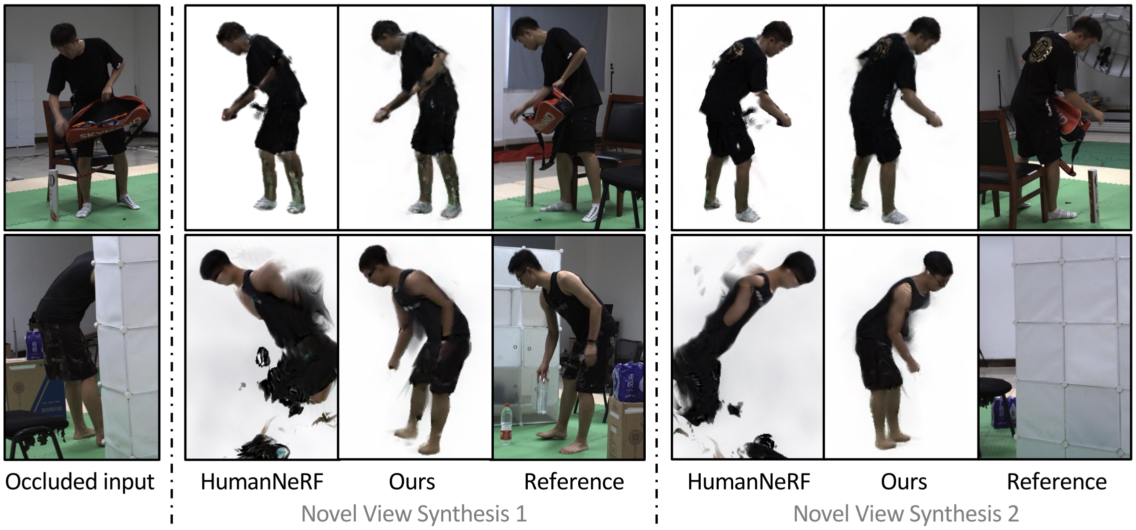

Figure 6. Qualitative results on **real-world** occlusions in OcMotion dataset [24].

manNeRF models for 40K iterations.

4.4. Results on Simulated Occlusions

Qualitative comparison on ZJU-MoCap videos with simulated occlusions between HumanNeRF and OccNeRF is shown in Figure 5. OccNeRF is capable of rendering a mostly completed body geometry with sensible details filled in at occluded areas. On the contrary, HumanNeRF fails to recover occluded body parts and produces significant artifacts in the occluded areas. Additionally, the quantitative results in Table 1 show that OccNeRF surpasses HumanNeRF for all subjects and under both metrics by a great margin. Note that this straightforward simulation of occlusions is in fact uncommon in real-world settings, where obstacles should have various shapes and humans are able to move across the entire scene with interactions with obstacles. *More comparisons against Neural Body [51] can be found in supplementary materials.*

4.5. Results on Real-world Occlusions

For better validating on real-world scenes, we present the rendering results on OcMotion videos in Figure 6. For the video with mild occlusions (top row), OccNeRF outperforms HumanNeRF with a higher fidelity rendering of texture details and much fewer artifacts at non-human regions. For the video with severe occlusions (bottom row), OccNeRF is still able to generate novel views with high-level rendering quality. However, HumanNeRF fails completely in such challenging cases when most body parts are occluded. This validates the superiority of OccNeRF in real-world scenes. OccNeRF also exceeds HumanNeRF on quantitative benchmarks as indicated in Table 1. Note that the metrics were computed on visible pixels in training images only, which ignored most of the artifacts HumanNeRF generated. *More comparisons on real-world scenes can be found in supplementary materials.*

4.6. Ablation Studies

In this section, we conduct additional experiments by simply removing each of the proposed components from the OccNeRF framework to prove their effectiveness. Quantitative metrics are also presented in the figures. *More ablation studies can be found in supplementary materials.*

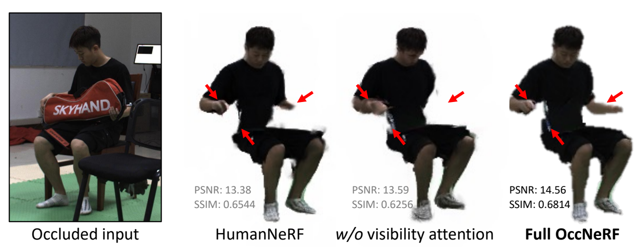

Figure 7. Our visibility attention improves rendering quality with more confident predictions at occluded areas with fewer blurs.

Impact of Visibility Attention. Our ablation studies start by proving the benefits of reformulating visibility priors as attention maps to be applied during surface-based rendering. Figure 7 shows that when disabling the attentive aggregation from Equation 9, the model becomes less confident in occluded areas, resulting in more blurs.

Impact of $\mathcal{L}_{\text{comp}}$. The proposed completeness loss $\mathcal{L}_{\text{comp}}$ is designed to encourage high-density values at locations inside the SMPL mesh. When removing this loss, Figure 8 shows that our method cannot render a complete geometry anymore. However, with our surface-based rendering, we still yield better results than HumanNeRF.

Impact of Surface-based Rendering. As discussed in section 3.2, we claimed that our proposed rendering strategy

ZJU-MoCap	Subject **377**				Subject **386**			
	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$
HumanNeRF [66]	11.29	0.5649	22.15	0.9612	9.491	0.4877	19.89	0.9531
OccNeRF	13.23	0.6097	23.43	0.9642	13.44	0.5974	23.66	0.9639

ZJU-MoCap	Subject **387**				Subject **392**			
	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$
HumanNeRF [66]	9.551	0.4140	19.47	0.9408	11.04	0.5290	21.01	0.9543
OccNeRF	13.27	0.5243	22.26	0.9513	13.00	0.5692	22.13	0.9575

ZJU-MoCap	Subject **393**				Subject **394**			
	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$
HumanNeRF [66]	10.86	0.4483	20.92	0.9476	10.55	0.4764	20.56	0.9489
OccNeRF	12.00	0.4655	21.58	0.9489	13.12	0.5317	22.06	0.9532

OcMotion	Video **Mild**				Video **Severe**			
	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$	PSNR_{vis}	SSIM_{vis}	$\text{PSNR}_{\text{full}}$	$\text{SSIM}_{\text{full}}$
HumanNeRF [66]	13.38	0.6544	21.18	0.9680	11.40	0.4545	17.96	0.9470
OccNeRF	14.56	0.6814	21.50	0.9695	14.95	0.5998	21.16	0.9692

Table 1. Quantitative comparison on the ZJU-MoCap and OcMotion datasets. We color cells that have the best metric values.

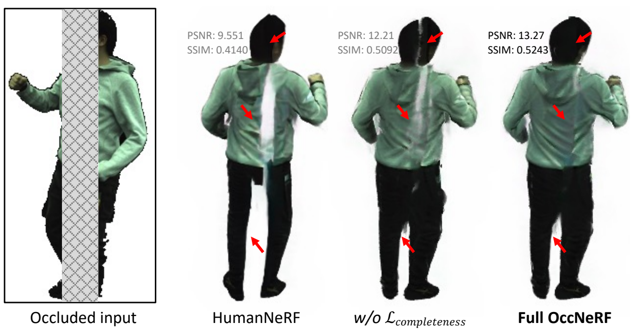

Figure 8. Our $\mathcal{L}_{\text{comp}}$ improves geometry completeness a step further when combined with the proposed rendering strategy.

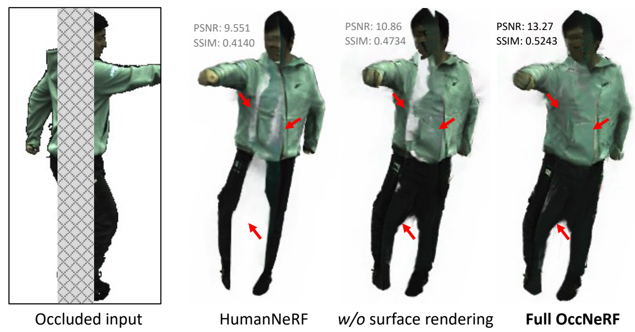

Figure 9. Our surface-based rendering method fills in the occluded parts with both accurate geometry and appropriate appearance.

enables $\mathcal{F}(\cdot)$ to condition on inputs better with more overlaps. Here we validate the necessity of such a design by removing it from the framework. We, however, still keep the hash grid encoding to see its impact. According to Figure 9, the hash grid encoding alone is not able to achieve comparable performance to our full OccNeRF. It has to be equipped together with the proposed rendering strategy. This validates that major performance improvements do come from surface-based rendering.

5. Discussions and Conclusion

Discussions. It is difficult to optimize scene-specific neural radiance fields under occlusions. There is neither a ground truth for the occluded parts nor external information from different scenes to inpaint the missing area. OccNeRF achieves rendering of the occluded regions by referring to nearby visible correspondences and enforcing complete geometry. However, OccNeRF can yield subtle artifacts. This is because we have more parameters to optimize and fewer training data due to occlusions. Since no external information is accessible, OccNeRF is not capable of inpainting an area that has never been seen in the video. The above limitations can be overcome with a better geometry prior [27] and a cross-scene training strategy [8]. Although the hash grid encoding accelerates the convergence at training, OccNeRF runs relatively slower than HumanNeRF at inference.

Conclusion. We proposed OccNeRF for rendering humans from object-occluded monocular videos. Most existing methods assume clear views of the entire human body without any interference, which is not feasible in real-world scenes. We designed a surface-based rendering strategy that incorporates geometry and visibility priors to assist rendering under occlusions. Moreover, our novel loss function is also able to help maintain geometry completeness. In our experiments, we compared OccNeRF against the state-of-the-art method under both simulated and real-world video occlusions. Our state-of-the-art results set up a new benchmark in this field of research.

Acknowledgments. This work was partially funded by the Gordon and Betty Moore Foundation, Panasonic Holdings Corporation, NSF RI #2211258, and Stanford HAI.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[4] Alexander W Bergman, Petr Kellnhofer, Yifan Wang, Eric R Chan, David B Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *arXiv preprint arXiv:2206.14314*, 2022. 2

[5] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in Neural Information Processing Systems*, 33:20496–20507, 2020. 2

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 2

[7] 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. 2

[8] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *ECCV*, 2022. 8

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

[10] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12952, 2022. 2

[11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 2

[12] 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 Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 2

[13] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11875–11885, 2021. 2

[14] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 2

[15] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 2

[16] Mark Endo, Kathleen L Poston, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, and Ehsan Adeli. Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 130–139. Springer, 2022. 1

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

[18] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[19] Beerend GA Gerats, Jelmer M Wolterink, and Ivo AMJ Broeders. 3d human pose estimation in multi-view operating room videos using differentiable camera projections. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–9, 2022. 1

[20] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 2

[21] 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*, pages 11046–11056, 2021. 2

[22] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 2

[23] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed

human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 535–545, 2021. 2

[24] Buzhen Huang, Yuan Shu, Jingyi Ju, and Yangang Wang. Occluded human body capture with self-supervised spatial-temporal motion prior. *arXiv preprint arXiv:2207.05375*, 2022. 5, 7

[25] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 710–720. IEEE, 2021. 5

[26] 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

[27] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2, 8

[28] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *arXiv preprint arXiv:2212.10550*, 2022. 1, 2, 4

[29] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2020. 2

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

[31] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[33] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[34] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[35] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 419–436. Springer, 2022. 2

[36] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[37] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2

[38] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 2

[39] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3

[40] 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, Oct. 2015. 1, 2, 4

[41] Guillaume Loubet, Nicolas Holzschuch, and Wenzel Jakob. Reparameterizing discontinuous integrands for differentiable rendering. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 38(6), Dec. 2019. 4

[42] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 369–374, 2000. 2

[43] 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. 1, 2, 4

[44] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2, 4

[45] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 2

[46] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 2

[47] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[48] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[49] Bo Peng, Jun Hu, Jingtao Zhou, and Juyong Zhang. Selfnerf: Fast training nerf for human from monocular self-rotating video. *arXiv preprint arXiv:2210.01651*, 2022. 1, 2, 4, 5

[50] 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

[51] 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. 1, 2, 5, 6, 7

[52] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[53] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[54] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 522–539. Springer, 2020. 2

[55] 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

[56] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? *arXiv preprint arXiv:1808.09316*, 2018. 2

[57] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[58] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *CVPR*, 2022. 2

[59] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 2

[60] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 2

[61] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2

[62] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 2

[63] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[64] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 1–19. Springer, 2022. 2

[65] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 2

[66] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1, 2, 3, 5, 8

[67] Chung-Yi Weng, Pratul P Srinivasan, Brian Curless, and Ira Kemelmacher-Shlizerman. Personnerf: Personalized reconstruction from photo collections. *arXiv preprint arXiv:2302.08504*, 2023. 3

[68] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2

[69] Kaibing Yang, Renshu Gu, Maoyu Wang, Masahiro Toyoura, and Gang Xu. Lasor: Learning accurate 3d human pose and shape via synthetic occlusion-aware data and neural mesh rendering. *IEEE Transactions on Image Processing*, 31:1938–1948, 2022. 2

[70] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13284–13293, 2021. 1

[71] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 546–556, 2021. 2

[72] 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. 5

[73] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single

color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 2