One-shot Implicit Animatable Avatars with Model-based Priors

Yangyi Huang[1,4*] Hongwei Yi[2*] Weiyang Liu[2,3] Haofan Wang[4]
Boxi Wu[5] Wenxiao Wang[5] Binbin Lin[5,6†] Debing Zhang[4] Deng Cai[1]
[1] State Key Lab of CAD & CG, Zhejiang University
[2] Max Planck Institute for Intelligent Systems, Tübingen [3] University of Cambridge
[4] Xiaohongshu Inc. [5] School of Software Technology, Zhejiang University [6] Fullong Inc.
huangyangyi@zju.edu.cn hongwei.yi@tuebingen.mpg.de

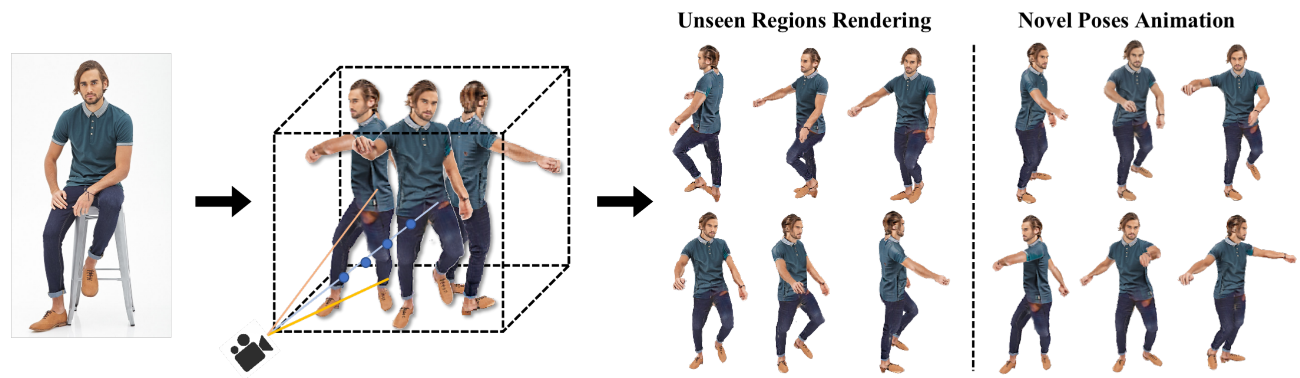

Figure 1: Our method creates free-viewpoint motion videos from a single image by constructing an animatable NeRF representation in one-shot learning.

Abstract

Existing neural rendering methods for creating human avatars typically either require dense input signals such as video or multi-view images, or leverage a learned prior from large-scale specific 3D human datasets such that reconstruction can be performed with sparse-view inputs. Most of these methods fail to achieve realistic reconstruction when only a single image is available. To enable the data-efficient creation of realistic animatable 3D humans, we propose ELICIT, a novel method for learning human-specific neural radiance fields from a single image. Inspired by the fact that humans can effortlessly estimate the body geometry and imagine full-body clothing from a single image, we leverage two priors in ELICIT: 3D geometry prior and visual semantic prior. Specifically, ELICIT utilizes the 3D body shape geometry prior from a skinned vertex-based template model (i.e., SMPL) and implements the visual clothing semantic prior with the CLIP-based pretrained models. Both priors are used to jointly guide the optimization for creating plausible content in the invisible areas. Taking advantage of the CLIP models, ELICIT can use text descriptions to generate text-conditioned unseen regions. In order to further improve visual details, we propose a segmentation-based sampling strategy that locally refines different parts of the avatar. Comprehensive evaluations on multiple popular benchmarks, including ZJU-MoCAP, Human3.6M, and DeepFashion, show that ELICIT outperforms strong baseline methods of avatar creation when only a single image is available. The code is public for research purposes at https://huangyangyi.github.io/ELICIT

1. Introduction

Creating realistic 3D contents of animatable human avatars from readily available camera inputs is of great significance for AR/VR applications, such as telepresence, virtual fitness, and so on. It is quite a challenging task and requires disentangled reconstruction of 3D geometry, the appearance of a clothed human, and accurate modeling of complex body poses for animation.

Current human-specific neural rendering methods have achieved promising performance when dense and well-

*Equal contribution.
†Corresponding author.

controlled inputs are available, e.g., multi-view videos captured by well-calibrated multi-camera systems [39, 62, 58, 37, 68], or long monocular videos [57] where almost all parts of the human body are visible. Despite their excellent performance, it is inconvenient (sometimes impossible) for ordinary users to obtain such high-quality dense inputs. Various methods have been proposed to address this data inefficiency. For example, ARCH [16] and ARCH++ [13] train reconstruction models with a single image input on large 3D scans datasets, but they do not generalize well to in-the-wild data. Neural radiance fields (NeRF) [32] based human-specific methods [11, 27, 22] train conditional models on multi-view images or video datasets to improve generalizability. However, when only sparse-view inputs are available, they also fail to generate realistic results under extreme settings, e.g., single monocular images.

Instead of learning conditional models from large-scale datasets [66, 6], recent work introduces various regularizations for geometry [36] and appearance [19, 60] to avoid degeneration, which makes it possible to synthesize visually plausible views in a semi-supervised framework without extra training data. However, due to the missing information about the occluded areas of the subject, they can hardly synthesize unseen views that barely overlap with the input views. To address these limitations, we propose a novel method, ELICIT, to learn human-specific neural radiance fields from a single image. We explicitly take advantage of the body shape geometry prior and the visual clothing semantic prior to guide the optimization and achieve free-view rendering from single images.

In summary, our contributions are listed below:

- We present ELICIT, a novel approach that can train an animatable neural radiance field from a single image without relying on extra training data.
- We propose two effective model-based priors to achieve an animatable 3D free-view rendering digital avatar from single image: 1) the visual clothing semantic prior. Specifically, we leverage the power of large pretrained vision-language models (i.e., CLIP) to hallucinate the unseen parts of the clothed body. 2) the human shape prior from the SMPL model. We use the estimated SMPL body shape and pose to constrain our reconstructed clothed 3D avatar be consistent with it.
- To create more realistic and consistent body part details, we propose a novel sampling strategy conditioned on the SMPL semantic segmentation and body rotation.

We conduct both quantitative and qualitative comparisons with recent human-specific neural rendering methods in the setting of single image input. We observe that ELICIT can consistently outperform existing methods in both free-view rendering and avatar animation, and simultaneously demonstrate promising performance on in-the-wild images.

Method	Subject data	Extra training data	Invisible area completion	Animatable
NeuralBody [39] Ani-NeRF [37] HumanNeRF [57]	multi-view images, monocular videos	**data-free**	✗	✓
PiFU [47] PaMIR [72] ARCH [16] ARCH++ [13] PHORHUM [1]	**monocular images**	3D scans	✓	✓
MPS-NeRF [11] NHP [22]	sparse videos, multi-view images	multi-view videos	✗	✓
MonoNHR [7]	**monocular images**	multi-view images	✓	✗
EVA3D [14]	**monocular images**	monocular images	✓	✓
ELICIT (ours)	**monocular images**	**data-free**	✓	✓

Table 1: **Recent human rendering methods that are most relevant to our work.** ELICIT is the first work that satisfies these four characteristics together: 1) only requires a single monocular image as an input. 2) doesn't need extra training data of the subject person. 3) supports recovering body areas that are invisible from the given input view. 4) animatable.

2. Related Work

Animatable human neural rendering. Existing methods of animatable human-specific neural rendering can be divided into 2D-based methods and 3D-based methods. 2D-based methods are mostly derived from image-based human pose transfer methods [49, 30, 35, 2], leveraging explicit temporal constraints [3, 63], optical flow estimation [56], and warping field [51, 65] to create temporally consistent pose-guided videos from input videos or images. Most single-image-based 3D methods [47, 48, 34, 24] learn encoder-decoder models from high-quality human 3D scans data. Among these works, ARCH [16], ARCH++ [13], and PHORHUM [1] are some of the most promising methods for reconstructing animation-ready 3D representations. However, data-driven methods are limited by the diversity of their training data distribution and may struggle with generalization issues when dealing with unseen clothing styles and complex body poses.

Recent works about human-specific neural radiance fields reconstruct animatable 3D human NeRF representation from multi-view or single-view video (For NeRF, see [32]). Most of them do per-subject optimization on an implicit model, using the whole video sequence as training data. Among which [39] learns structured latent codes on SMPL [29] mesh vertices, other methods construct the representation in a canonical space by modeling pose-driven deformation [57, 62, 52, 71, 38, 37]. While these methods produce impressive results, they require dense inputs that cover most areas of the human body. In contrast, our approach can generate an animatable realistic character from a single image, making it more user-friendly and flexible for a wider range of applications.

Single-view-based NeRF. The setting of novel view synthesis from only a single image is challenging for NeRF-based methods because incomplete geometric information can lead to degeneration results. Also, it is difficult for the model to synthesize regions in the novel view which is not visible for the input due to occlusion. Some existing methods utilize learned prior about scene geometry and appearance in a data-driven manner, e.g., generative adversarial models [50, 53], supervised learning [66, 6, 21, 55], and unsupervised learning [31] for conditional NeRF. However, most of these methods only focus on simple 3D shapes [5]. Eg3D [4] and CG-Nerf [20] are two representative methods that work on specific types of objects, such as human faces, using conditional generative NeRF.

There are also non-data-driven methods introducing priors from off-the-shelf models, including depth cues [9, 23] and other knowledge such as object geometry [36]. SinNeRF [60] and DietNeRF [19] use pre-trained image encoders to introduce semantic prior and produce semantically consistent novel view synthesis results from sparse inputs. Similarly, our work utilizes an SMPL-based human body prior and a CLIP-based visual semantic prior available in the task setting of single image-based human rendering and generates photo-realistic free-view renderings.

CLIP-driven radiance fields. CLIP [42] is a cross-modality representation learning method that has recently been applied to text-driven image generation [44, 46, 43]. Several works have incorporated CLIP and radiance fields for 3D-aware synthesis tasks. DietCLIP [19] synthesizes view-consistent novel views from sparse view input with a CLIP-based loss as a regularization on NeRF. CLIP-NeRF [54] applies joint image-text latent space in a conditioned NeRF for manipulation with multi-model inputs. LaTeRF [33] uses CLIP loss to extract objects of interest from the scene, similar to texture cues. AvatarCLIP [15] and Dream Fields [18] apply CLIP to the optimization process for text-driven 3D generation. NeuralLift-360 [61] enables lifting a 3D object from a single image based on CLIP-based image similarity. In our work, we extend the use of CLIP-driven NeRF by leveraging it for human-specific rendering from a single image, exploring its potential in generating photo-realistic free-view renderings.

Most relevant works. Recently, there have been several related works in the field of single-image-based human rendering. MonoNHR [7] proposes a data-driven approach using a conditional NeRF to render free-viewpoint images of a character from a single image input. EVA3D [14], on the other hand, learns an unconditional 3D human generative model on the DeepFashion dataset [28] and can reconstruct 3D humans from a single image by GAN inversion [45, 8]. However, its generalizability is largely limited by the biased distribution of the training datasets. A comparison of our method and related works is summarized in Table 1. ELICIT only requires a single image as input without using extra training data, and yet supports both invisible area completion and body animation.

3. Method

3.1. Problem Specification

We formulate the task of creating free-view videos for a character in novel poses as follows. The input includes a single-view image I_s of the character with camera parameters $\mathbf{e}_s$, SMPL-parameters (β, θ_s), where β describes the body shape of the character, and θ_s describes the body pose of the character in the input image. We also input a motion sequence of length n by SMPL pose parameters $\Theta_t = \{\theta_t^i\}_{i=1}^L$ and camera parameters of each frame $\boldsymbol{E}_t = \{\mathbf{e}_t^i\}_{i=1}^L$ for animation. The output is n video frames $\{I_t^i\}_{i=1}^L$ rendered by pose-conditioned NeRF model under the given camera parameter,

$$I_t^i = \Gamma[\boldsymbol{F}(\mathbf{x}, \theta_t^i), \mathbf{e}_t^i], \tag{1}$$

where $\boldsymbol{F}$ is the pose-conditioned radiance field function and Γ represents volume rendering.

3.2. Preliminaries

SMPL [29], Skinned Multi-Person Linear model, is a skinned vertex-based template model driven by large-scale aligned human surface scans. SMPL encodes posed body shape by a pose parameter $\theta_t^i \in \mathbb{R}^{72}$ and a shape parameter $\beta \in \mathbb{R}^{10}$, and outputs a blend shape sculpting the human body with 6890 vertices. We use SMPL parameters to represent the input character's body shape, posture, and pose sequence input of the target motion.

HumanNeRF [57] is a human-specific variant of neural radiance field(NeRF), which supports free-view rendering of a moving character from monocular video inputs. In particular, HumanNeRF represents a moving character with a canonical appearance volume F_c warped to an observed pose to produce output appearance volume F_o:

$$F_o(\boldsymbol{x}, \boldsymbol{p}) = F_c(\mathcal{T}(\boldsymbol{x}, \boldsymbol{p})), \tag{2}$$

where $F_c :\longrightarrow (\boldsymbol{c}, \sigma)$maps position $\boldsymbol{x}$ to color $\boldsymbol{c}$ and volume density σ. Notice that HumanNeRF uses a simplified version of NeRF without considering viewing directions. The motion field $\mathcal{T} : (\boldsymbol{x}_o, \boldsymbol{p}) \rightarrow \boldsymbol{x}_c$ maps positions in observed space back to the canonical space, conditioned by pose parameters $\boldsymbol{p} = (J, \Omega)$, where J represents 3D joint locations and Ω represents local joint rotations. The novel views are synthesized by NeRF-based volume rendering:

$$\boldsymbol{C}(\boldsymbol{r}) = \int_{t_n}^{t_f} T(t)\sigma(\boldsymbol{r}(t))\boldsymbol{c}(\boldsymbol{r}(t), \boldsymbol{d})dt, \tag{3}$$

where the $T(t)$ is the transmittance of the light at position t, $T(t) = \exp(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(s))ds)$. And r is the pixel ray cast

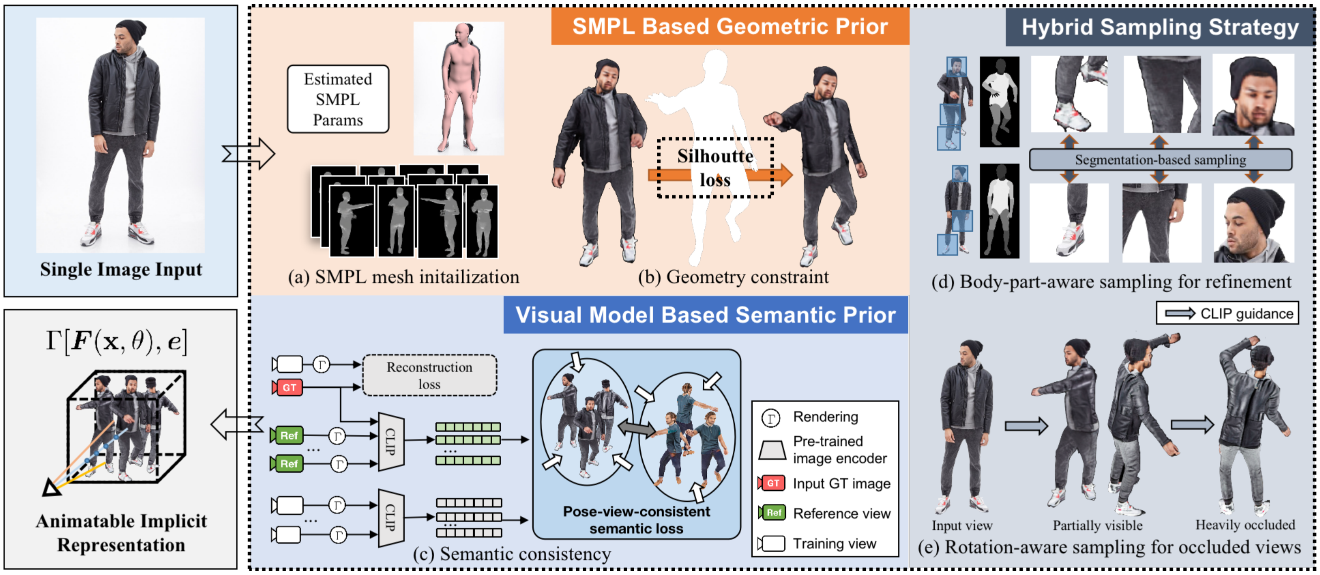

Figure 2: **Method overview.** Our method generates an animatable avatar from a single source image of a person, which can be used to create pose-guided free-view renderings of the person with any target motion in SMPL format. ELICIT train an animatable implicit human representation called HumanNeRF using one-shot prior-based learning. We use two model-based priors to guide the optimization process: the SMPL-based Geometric Prior and the Visual-Model-based Semantic Prior. The Human Body Prior is (a) initialized with multi-view video frames rendered by SMPL meshes and (b) uses a silhouette loss to constrain synthesized geometry and body poses during training. The Semantic Prior provides (c) pose-view-consistent semantic supervision for novel views of novel poses using a powerful pre-trained visual model. Additionally, we propose a Hybrid Sampling Strategy that includes (d) body-part-aware sampling to refine body-part details and (e) rotation-aware sampling to better recover heavily occluded views.

from the observer ,$\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$. The original data-driven optimization of HumanNeRF requires monocular video input where most of the regions of the character are visible. We use it as the basic model of implicit neural representation for free-view motion rendering.

3.3. Prior-driven One-shot Learning for Single-image Human Rendering

Figure 2 illustrates our overall pipeline. ELICIT obtains the animatable implicit human representation by per-subject optimization with a single image input. We formulate this one-shot learning process as follows:

For each iteration, a training view with a character pose and camera parameters, $V_{\text{train}} = (\theta_{\text{train}}, \mathbf{e}_{\text{train}})$, is sampled from the input view $V_s = (\theta_s, \mathbf{e}_s)$ and target views $V_t \in \{(\theta_i, \mathbf{e}_j)\}_{i=1,j=1}^{L,M}$, where $\{\mathbf{e}_j\}_{j=1}^{M}$ are preset cameras around the character. We supervise the training view rendering with a respective reference view V_{ref}. The reference view could be the ground-truth view or rendered results of a sampled neighboring view.

On the one hand, to get realistic synthesis, rendering a consistent input view is the fundamental goal to be guaranteed. When the sampled view V_{train} is identical to V_s, we select $V_{\text{ref}} = V_s$ and use the input image I_s as the training target of rendered $\hat{I}_s$. We formulate our reconstruction loss the same as HumanNeRF [57].

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{LPIPS}}(I_s, \hat{I}_s) + \lambda \mathcal{L}_{\text{MSE}}(I_s, \hat{I}_s), \qquad (4)$$

where $\mathcal{L}_{\text{MSE}}$ is a pixel-wise mean square error loss, and $\mathcal{L}_{\text{LPIPS}}$ is a VGG-based perception loss that is robust to slight misalignment and improves reconstruction details.

On the other hand, we need to supervise V_t for novel view synthesis and pose synthesis. We expect the synthesis results to have: (1) a consistent appearance with the input character, (2) a plausible geometry that approximates the actual clothed body shape, and (3) a body pose that matches the target motion. Obtaining such 3D-aware synthesis from incomplete input requires utilizing prior knowledge. In contrast to using a learned prior from multi-view images [11, 22, 7] or 3D scans training data [59, 48, 47, 13], we introduce two model-based prior to guide the optimization. One is *visual model-based semantic prior*, which supervises the synthesis of consistent visual contents. The other is *SMPL-based human-specific prior* that provides knowledge about human body shape and posture.

3.3.1 Visual model-based semantic prior

Recent works [19, 18, 61] show that novel view synthesis from a single image or sparse inputs can be done with the guidance of an embedding loss, which enforces semantic consistency between unseen views and the reference view. Such optimization-based methods are also applicable to our task of synthesizing 3D-aware content for a clothed human. To achieve this, we need a powerful vision model to embed the images from different views of 3D humans in a semantically meaningful latent space.

Among different models, we find that the CLIP [42] visual encoders pre-trained on diverse image-text pairs data are suitable for this task. In Figure 3, we carry out a similar evaluation of CLIP-NeRF [54], demonstrating view-pose-consistency of CLIP embeddings on human images. On the other hand, the CLIP models can also capture detailed visual semantics such that rich supervision signals can be utilized for a vivid generation [12].

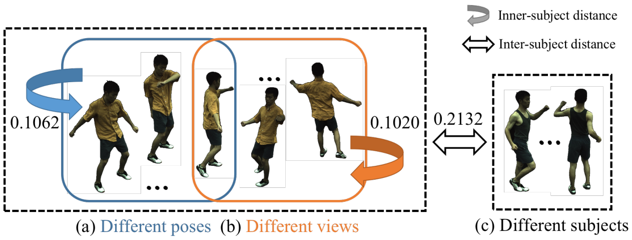

Figure 3: **View-pose-consistency of the CLIP embeddings.** The embedding distance of the same character under different views and poses is significantly smaller than the distance between two different characters.

Figure 4: **Enhancing semantic prior with text prompts.** Combining text guidance with image guidance in the semantic loss helps recover the garment structure of the avatar's backside. It is worth noticing that using only text guidance leads to false facial appearances.

By comparing the performance of different model-based embedding losses, we select the CLIP ViT-based cosine distance as the semantic loss, formulated as follows:

$$\mathcal{L}_{\text{CLIP}} = \phi(I_{\text{ref}})^T \phi(\hat{I}_{\text{train}}), \tag{5}$$

where $\hat{I}_{\text{train}}$ is the rendered image of sampled training view, I_{ref} is the reference view, and ϕ is the normalized embedding function of the CLIP ViT.

Notably, the joint embedding space of CLIP is widely applied in the latest text-driven image generation works [54, 44, 10, 15]. It also enables our method to support the use of user prompts P_{ref} to guide the optimization of novel views. We can use the CLIP text embedding $\phi_{\text{text}}(P_{\text{ref}})$ as a reference in $\mathcal{L}_{\text{CLIP}}$. Figure 4 demonstrates that detailed text prompts can aid in synthesizing invisible garments, but using only text guidance is insufficient for preserving the identity of the avatar. On the other hand, image-based semantic loss can recover crucial visual attributes like facial appearance and texture details. By utilizing both semantic losses, we can enhance the performance in challenging cases, such as those involving complex garments.

3.3.2 SMPL-based geometric prior

By incorporating off-the-shelf pose estimation models [25], we can obtain information about approximate body shapes from SMPL. Our method utilizes this human-specific prior as a geometric clue for 3D human reconstruction and animation by introducing an SMPL-based NeRF initialization and a soft geometry constraint in training.

SMPL-based NeRF initialization. It is difficult for a NeRF model to recover the exact body shape because of occlusions and depth ambiguity. Thus directly optimizing a NeRF with a single image is likely to result in representation degeneration. Inspired by AvatarCLIP [15], we initialize our HumanNeRF implicit representation by SMPL meshes renderings. More specifically, we use detected body shape parameters along with pose parameters of the target motion sequence to construct corresponding animated SMPL meshes. Then the multi-view renderings of the meshes are used as pseudo ground truth for initialization.

Given the estimated parameterized body shape β and target motion sequence $\Theta_t = \{\theta_t^i\}_{i=1}^L$, we render image views $\{I_{\text{SMPL}_i^{(j)}}\}_{i=1,j=1}^{L,m}$ with pre-defined m-view camera poses $\boldsymbol{E}_s = \{\mathrm{e}_s^i\}_{i=1}^m$ and template meshes generated by SMPL model $M_i = M_{\text{SMPL}}(\beta, \theta_t^i; \Phi)_{i=1}^L$. We also use a template texture to avoid body part occlusion ambiguity. We initialize HumanNeRF with a multi-view setup of its training process. Each iteration samples an image view $I_{\text{SMPL}_i^{(j)}}$ for training with a reconstruction loss on the result.

Soft geometry constraint. In the initialization stage, We utilize the human body geometry with SMPL meshes to help the model render approximate shapes for the target character in specific poses. However, we empirically find that optimizing the model with only the semantic loss may lead to degenerated results, including inconsistent rendered poses and missing body parts, despite the similarity of the CLIP embedding across various views and poses.

For this issue, we introduce a soft geometry constraint based on the assumption that the estimated SMPL meshes are close to the geometry of the naked body of the target character. Therefore, the estimated SMPL meshes should be covered by the actual shape of the clothed character. This loss function can be viewed as a masked version of the silhouette loss in [69], which consists of an MSE loss and a one-way Chamfer distance loss for the silhouette boundary. We only compute this loss for the rendered alpha pixels that are covered by the SMPL silhouette. Given the SMPL silhouette mask S and the rendered alpha map A, we compute the following loss function:

$$\mathcal{L}_{\text{sil}} = \sum_{p \in S} \|A(p) - S(p)\|_2^2 + \min_{\hat{p} \in \text{Edge}(S)} A(p) \, \|p - \hat{p}\|_1 \,, \quad (6)$$

where $\circ$ denotes the element-wise product, $\text{Edge}(S)$ computes the edge of mask S, $A(p)$ is the pixel value of A at p, and $S(p)$ is the pixel-wise mask of S at p. This constraint maintains the character's body structure in target motion during training. Also, it allows the creation of detailed outside geometries that better match the target avatar.

3.3.3 Hybrid sampling strategy with appearance prior

Only enforcing global semantic consistency among novel views and poses can possibly lead to unrealistic artifacts on body parts. To tackle this issue, we proposed a body-part-aware sampling to refine body-part details and a rotation-aware sampling to better recover heavily occluded views.

Body-part aware sampling. To improve the quality of synthesized details and overcome the resolution constraint of pre-trained CLIP, SinNeRF [60] proposes using semantic feature loss between extracted features of randomly sampled local patches rather than complete global views. However, in human-specific rendering tasks, appearance and semantic features vary significantly across different body parts. To address this issue, we introduce a body-part-aware patch sampling strategy to synthesize the well-aligned visual details of the human body.

For each sampled view $V_{\text{train}} = (\theta_{\text{train}}, \mathbf{e}_{\text{train}})$, our method randomly selects a body part p ,including the whole body, to refine. The rendered segmentation of SMPL can determine the corresponding region, $S^p_{\text{SMPL}}(\theta_{\text{train}}, \mathbf{e}_{\text{train}})$, explicitly defined by groupings of SMPL meshes. Accordingly, we adjust the training camera to render a local patch $V^p_{\text{train}} = (\theta_{\text{train}}, \mathbf{e}^p_{\text{train}})$ for this body part. We can also crop a corresponding reference patch V^p_s from the input image by the SMPL segmentation $S^p_{\text{SMPL}}(\theta_s, \mathbf{e}_s)$. Similarly, we can also render the patch $V^p_{\text{ref}} = (\theta_{\text{train}}, \mathbf{e}^p_{\text{ref}})$ when a neighboring view $\mathbf{e}^p_{\text{ref}}$ is sampled as the reference.

Rotation-aware sampling for occluded views. As discussed in [18, 15], using a global semantic loss for 3D generation can result in multi-faced appearances on different sides of the object, which are against realism for 3D avatars. To address this issue, we propose an orientation-aware sampling to recover heavily occluded regions.

Specifically, for a sampled pose $\theta^i_t \in \mathbf{\Theta_t}$, we calculate the body orientations (relative to the input image) $\{\psi(\mathbf{e}^j_s)\}^m_{j=1}$ on the horizontal plane of defined camera $\{\mathbf{e}^j_s\}^m_{j=1}$ and divide the cameras into pre-defined ranges of front cameras $\mathbf{E}_{\text{front}}$, side cameras $\mathbf{E}_{\text{side}}$, and rear cameras $\mathbf{E}_{\text{rear}}$ according to $\{\psi^j_i\}^m_{j=1}$. Since body regions of rear views are heavily occluded in the input image, when a rear camera $\mathbf{e}_{\text{train}} \in \mathbf{E}_{\text{rear}}$ is sampled for training, we use the nearest camera $\mathbf{e}_{\text{ref}} \in \mathbf{E}_{\text{side}}$ to render a reference view $V_{\text{ref}} = (\theta_{\text{train}}, \mathbf{e}^p_{\text{ref}})$ instead of the input view V_s. Additionally, since the head region of the avatar is more susceptible to multi-faced artifacts, we also use $\mathbf{e}_{\text{ref}} \in \mathbf{E}_{\text{front}}$ for $\mathbf{e}_{\text{train}} \in \mathbf{E}_{\text{side}}$. Such a strategy infers the appearance of totally occluded views with partially visible views, based on the assumption of visual continuity.

3.3.4 Overall loss function

ELICIT constructs animatable avatars through a two-stage optimization process. In the SMPL-based initialization stage, we optimize the model using only the reconstruction loss in Eq. (4). In the one-shot training stage, based on the input image and target motion, we optimize the model using the overall loss function consisting of L_{recon}, L_{CLIP}, and L_{sil}. The detailed loss function is defined as follows:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{recon}}, & \text{if } V_{\text{train}} = V_s \\ \lambda_{\text{CLIP}} \mathcal{L}_{\text{CLIP}} + \lambda_{\text{sil}} \mathcal{L}_{\text{sil}}, & \text{otherwise} \end{cases} \quad (7)$$

where λ_{CLIP}, λ_{sil} are hyperparameters for different losses. See Sup. Mat. for more details of the optimization.

4. Experiments and Results

4.1. Datasets

We conducted evaluations on two multi-view human video datasets: ZJU-MoCap [39] and Human3.6M [17], as well as a 2D human image dataset, DeepFashion [28]. We selected all nine subjects from ZJU-MoCap and the "Posing" video of all seven subjects from Human3.6M to evaluate free-view animation. To obtain input images, we sampled frames from the first camera of ZJU-MoCap and the third camera of Human3.6M, along with annotated SMPL parameters, camera matrices, and segmentation masks. We applied the annotated motion sequence of each video clip for animation. Additionally, we used high-resolution full-body photos from DeepFashion [28] to evaluate our model's performance on human avatars with various clothing styles.

4.2. Comparison to Existing Methods

To the best of our knowledge, NeRF-based human-specific novel view synthesis can be classified into per-subject optimization methods and generalizable methods. We selected three state-of-the-art methods as baselines: Neural Body [39] (NB) and Animatable NeRF [37] (Ani-NeRF) from per-subject optimization methods, and Neural Human Performer [22] (NHP) from generalizable methods. All three methods employ SMPL-based human body priors, with NB and Ani-NeRF supporting novel pose synthesis for

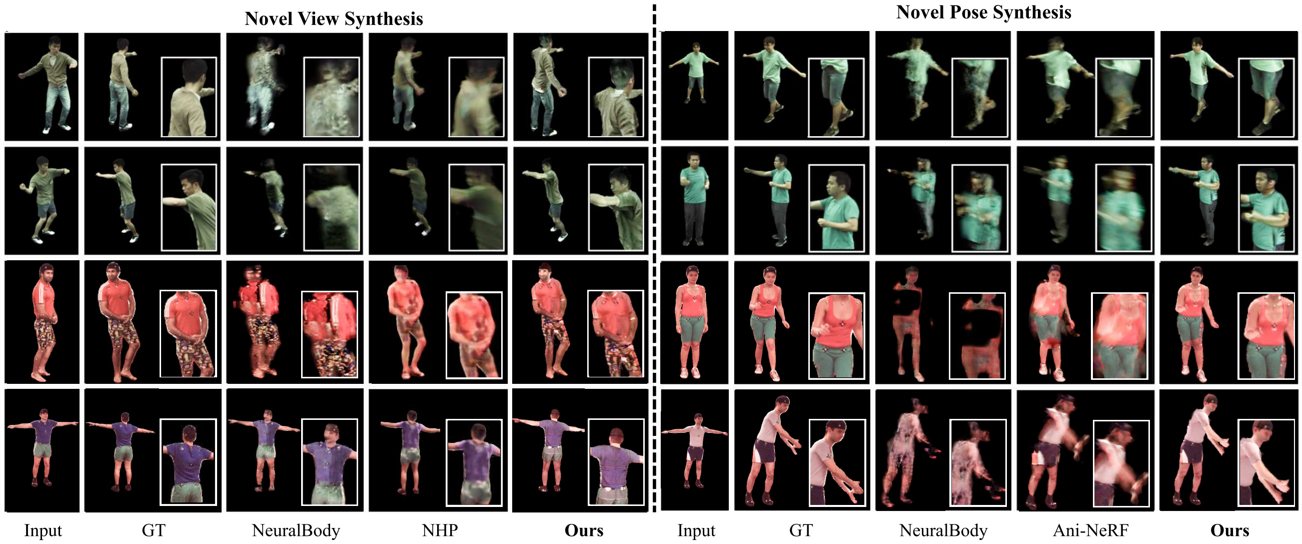

Figure 5: **Overall comparison.** Compared with state-of-the-art NeRF based methods[39, 22, 37] on novel view synthesis and novel pose synthesis, ELICIT generates human 3D renderings with more consistent appearance and realistic details from a single image. We adjust the exposure for better visualization.

animation. For a fair comparison, we adapted these baselines to take single-image inputs. We compared the performance of these methods in two different task settings: novel view synthesis for free-view rendering and novel pose synthesis for character animation.

Metrics. As in previous works [39, 22, 27], we evaluated our results using two standard metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). To account for perceptual similarity, we also calculated the Learned Perceptual Image Patch Similarity (LPIPS) metric [70], which has been used in recent NeRF-based human rendering works [57, 71]. We followed the evaluation protocol of [39] and calculated the metrics only on the bounding box region, rather than the entire image.

Comparison on novel view synthesis. We evaluated the performance of our method and the baselines on the task of novel view synthesis using two multi-view human video datasets, ZJU-MoCap and Human3.6M. We uniformly sampled 10 frames from each subject video and evaluated the results on all available camera views, except for the input view. For per-subject optimization methods NB [39] and Ani-NeRF [37], we optimized one model for each frame. For the generalizable method NHP [22], we sampled three subjects from each dataset as the testing set S_{test} and pre-trained the model only on the remaining subjects of each dataset to ensure a fair comparison.

As shown in Table 2, our ELICIT outperformed the baselines in terms of PSNR, SSIM and LPIPS on both datasets. Notably, our method's superior performance on the SSIM and LPIPS metrics highlights its advantage in producing perceptually high-quality rendering results. Overall, these results demonstrate the effectiveness of our method in synthesizing high-quality novel views of human subjects.

		ZJU-MoCAP			Human 3.6m		
Subjects	Methods	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS ↓
S_{all}	NB[39]	20.2	0.811	0.235	20.0	0.752	0.269
	ELICIT	**21.9**	**0.872**	**0.123**	**21.5**	**0.824**	**0.143**
S_{test}	NB[39]	19.0	0.813	0.229	20.4	0.752	0.269
	NHP[22]	21.0	0.869	0.175	21.7	0.825	0.175
	ELICIT	**21.4**	**0.886**	**0.118**	**21.8**	**0.829**	**0.146**

Table 2: Quantitative comparison of novel view synthesis on ZJU-MoCap and Human3.6M in PSNR, SSIM (higher is better) and LPIPS (lower is better). ELICIT outperforms NB and NHP on all metrics.

Comparison on novel pose synthesis. For both datasets, we select one front-view image as input for each subject and evaluate the entire video clip synthesized with motion annotations. For Ani-NeRF, we use the pose-dependent displacement field model proposed in [38], which reports their best results. As shown in Table 3, our method also produces high-quality synthesis when generalized to novel poses.

Method	ZJU-MoCAP			Human 3.6m		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS ↓
NeuralBody[39]	20.2	0.784	0.282	21.5	0.799	0.264
Ani-NeRF[37]	20.3	0.791	0.277	22.2	0.807	0.229
ELICIT	**21.8**	**0.853**	**0.143**	**22.6**	**0.859**	**0.123**

Table 3: Quantitative comparison of **novel pose synthesis** on ZJU-MoCap and Human 3.6M in PSNR, SSIM (higher is better) and LPIPS (lower is better). ELICIT outperforms both NB and Ani-NeRF in all the metrics.

We show sampled novel view and pose synthesis results in Figure 5. Compared to the latest NeRF-based methods, ELICIT performs better in rendering realistic visual details and inferring occluded contents of clothed human bodies.

4.3. Qualitative Analysis

Our single-image-based method aims to enable users to create animatable 3D characters from simply available photos of real people. Therefore, in addition to the quantitative evaluation on multi-view human video datasets, we evaluate

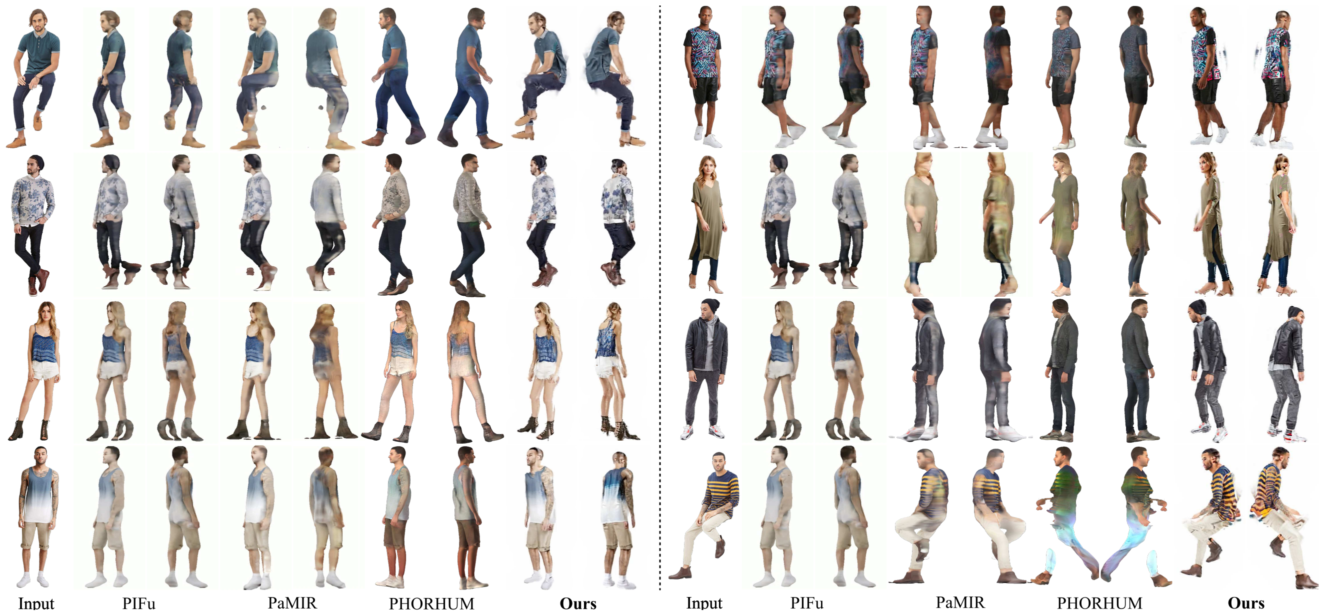

Figure 6: **Qualitative results** of PIFu [47], PaMIR [72], PHORHUM [1] and ELICIT on DeepFashion [28]. ELICIT generates more realistic details in occluded views and generalizes well on challenging body poses.

our approach on 2D human images from DeepFashion [28] dataset, with SMPL parameters estimated by off-the-shelf pose estimation models [25, 67]. Among previous data-driven non-NeRF methods, PIFu [47] and PaMIR [72] and PHOHRUM [1] support both reconstruction of geometry and texture from single image input, which have also shown impressive results on DeepFashion dataset. Here we choose these three methods for qualitative comparison. Figure 6 illustrates that our training-data-free one-shot method generalizes well on real-world human images and creates rich details for body textures, such as patterns on clothes and shoes, tattoos on the skin, and details of face and hair. While PIFu and PaMIR produce blurry results, limited by the distribution gap between training data and in-the-wild data.

4.4. Ablation Studies

We conduct our ablation studies on introduced model-based priors and select representative subjects from ZJU-MoCap and DeepFashion for comparison.

Implicit representation. We compare our method with a simple baseline of modeling the animatable character explicitly by SMPL meshes, which only optimizes its per-triangle texture parameters during training. Such an explicit model produces noisy textures, and its SMPL-based geometry is also inaccurate compared to the actual human shape. As shown in Table 4 and Figure 7(a), an implicit representation such as HumanNeRF, which models the character appearance with a spatially continuous function, is necessary for the one-shot learning stage.

SMPL mesh initialization. Initializing our implicit representation with the rendered views of SMPL mesh imparts an approximate human shape and body part semantics at the beginning of the optimization. The significant performance drop in Table 4 and Figure 7(a) illustrates that this step is necessary for our approach. Only on this basis can semantic loss and geometric constraints guide the completion of detailed geometry and textures.

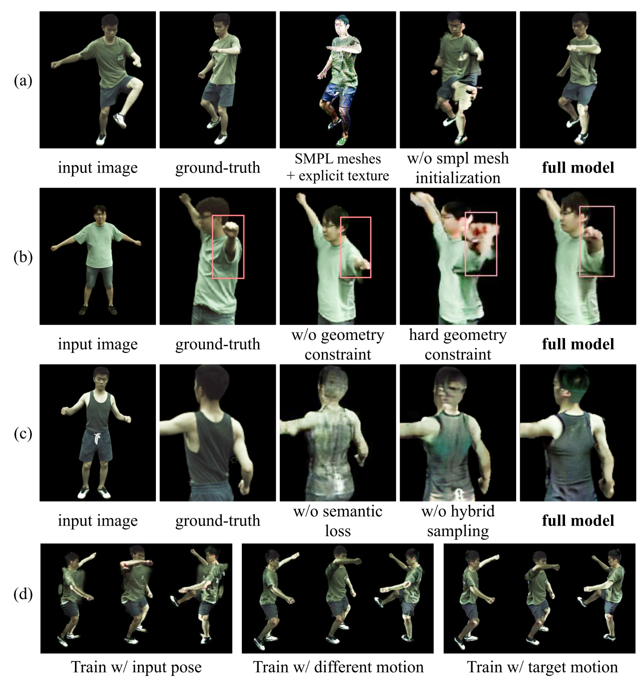

Figure 7: **Qualitative results** for the ablation studies of priors used in our method, selected from ZJU-MoCap [39] dataset.

Soft geometric constraint. As shown in Figure 7(b), optimizing the model without geometric constraints may lead to error poses Moreover, in contrast to matching the SMPL geometry directly by a hard constraint of silhouette loss, we only penalize the internal misalignment. This soft constraint allows the implicit model to learn human geometry with clothes and affiliate objects, while the hard one brings in artifacts due to the misalignment of the SMPL shape and the clothed body shape.

CLIP-based semantic loss. As shown in Figure 7(c), our semantic loss plays a vital role in generating plausible content for the occluded areas. We also compare the performance of different pre-trained vision models in Sup. Mat.. The results indicate that vision models pre-trained with large multi-modal data and large-capacity models are particularly effective in the semantic loss.

Sampling strategy. Figure 7(c) illustrates that certain artifacts from some small areas can significantly affect the overall visual quality, such as texture artifacts on cloth textures and missing hair on the back of the head. Our hybrid sampling strategy helps to generate vivid details and avoid multi-faced artifacts for avatar creation.

Training poses. In Figure 7(d), we show a comparison of animation results of different training poses. The failure of body shape in the input-pose-only training result illustrated the necessity of diverse training poses. While the results of training with different motion sequences show that ELICIT generalizes well to novel poses in test-time animation.

Setting	PSNR↑	SSIM↑	LPIPS↓
SMPL mesh w/ explicit texture	17.20	0.7779	0.2116
w/o SMPL mesh initialization	19.01	0.7911	0.2122
w/o semantic loss	21.46	0.8592	0.1344
w/o geometric constraint	21.68	0.8633	0.1282
hard geometry constraint	20.58	0.8288	0.1664
w/o hybrid sampling strategy	21.46	0.8592	0.1344
training only w/ input pose	21.47	0.8562	0.1516
full model	**22.61**	**0.8908**	**0.1115**

Table 4: Ablation study on subjects {313, 377, 392} of ZJU-MoCap.

5. Discussion on Limitations

While our reconstruction results are generally promising, there remain certain instances of failure. This section provides a comprehensive analysis regarding the limitations of ELICIT and discusses some potential future directions.

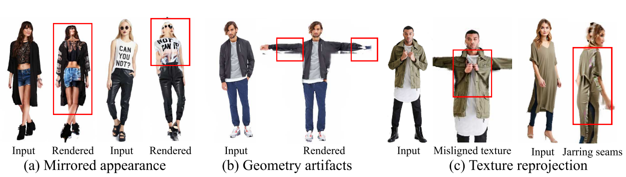

Figure 8: We show failure cases of (a) mirrored appearance, (b) geometry artifacts, and (c) errors in texture re-projection for limitation analysis.

Mirrored appearance: Although ELICIT can successfully recover the back-side appearance in many cases with the help of $\mathcal{L}_{\mathrm{CLIP}}$ and the hybrid sampling strategy, the problem of mirrored appearance still happens sometimes. As shown in Figure 8(a), since ELICIT cannot separate the semantic information of different attributes in $\mathcal{L}_{\mathrm{CLIP}}$, it fails to recover complex garment layering and pattern. The presence of intricate facial attributes can also result in mirrored faces. In Figure 4, we have shown a possible improvement to $\mathcal{L}_{\mathrm{CLIP}}$ with text guidance. We believe that enhancements with richer semantic information (e.g., human parsing segmentation [26] and view-aware text guidance [15]), and integration of text-to-image generative models [46, 40, 41] can further improve the quality of back-side appearance.

Limited geometry quality: Apart from an SMPL-based initialization and a soft geometry constraint, ELICIT has no direct supervision of the clothed body geometry and relies on $\mathcal{L}_{\mathrm{CLIP}}$ to create geometry details indirectly. As shown in Figure 8(b), the artifacts in the self-contact body parts and hands show $\mathcal{L}_{\mathrm{CLIP}}$ has limited ability in modeling geometric details. In future work, to alleviate this problem, we can introduce additional supervision of the geometry, including surface regularization, estimated surface normal and depth, and accurate estimation of face geometry and hand geometry, thus enabling expressive animation from some motion generation methods, e.g., TalkSHOW [64].

Texture re-projection: We use a strong constraint of $\mathcal{L}_{\mathrm{recon}}$ to re-project the input view texture. However, as shown in Fig. 8(d) some texture could be reprojected onto the wrong body parts due to the misalignment between the recovered body shape and the actual geometry, and a strong loss weight of $\mathcal{L}_{\mathrm{recon}}$ at the edge of the input could lead to jarring seams. Improving geometry alignment, disentangling and balancing $\mathcal{L}_{\mathrm{recon}}$ of different body parts could be potential solutions for this problem.

6. Concluding Remarks

We introduce ELICIT, a novel method to construct an animatable implicit representation from a single image input and generate a free-view video of the character in the target motion. Two model-based priors drive the one-shot optimization of ELICIT: the visual-model-based visual semantic prior and the SMPL-based human body prior, which enables the reconstruction of body geometry and the inference of full body clothing. We evaluate our methods both qualitatively and quantitatively. We demonstrate our superior performance in single-image settings compared to prior work on novel view and novel pose synthesis, and strong generalizability on real-world human images.

Acknowledgements. This work was supported in part by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B, in part by The National Nature Science Foundation of China (Nos: 62273301, 62273302, 62273303, 62036009, 61936006), in part by the Key R&D Program of Zhejiang Province, China (2023C01135), and in part by Yongjiang Talent Introduction Programme (No: 2022A-240-G).

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1506–1515, 2022. 2, 8

[2] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8340–8348, 2018. 2

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody Dance Now. In *International Conference on Computer Vision (ICCV)*, pages 5933–5942, 2019. 2

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 3

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast Generalizable Radiance Field Reconstruction From Multi-View Stereo. In *International Conference on Computer Vision (ICCV)*, pages 14124–14133, 2021. 2, 3

[7] Hongsuk Choi, Gyeongsik Moon, Matthieu Armando, Vincent Leroy, Kyoung Mu Lee, and Grégory Rogez. Mononhr: Monocular neural human renderer. *International Conference on 3D Vision (3DV)*, pages 242–251, 2022. 2, 3, 4

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11875–11885, 2021. 3

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2022. 3

[10] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:5207–5218, 2022. 5

[11] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2, 4

[12] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 5

[13] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. In *International Conference on Computer Vision (ICCV)*, pages 11046–11056, 2021. 2, 4

[14] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[15] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 3, 5, 6, 9

[16] 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

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 6

[18] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 867–876, 2022. 3, 4, 6

[19] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *International Conference on Computer Vision (ICCV)*, pages 5865–5874, Oct. 2021. 2, 3, 4

[20] Kyungmin Jo, Gyumin Shim, Sanghun Jung, Soyoung Yang, and Jaegul Choo. Cg-nerf: Conditional generative neural radiance fields. *arXiv preprint arXiv:2112.03517*, 2021. 3

[21] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. GeoNeRF: Generalizing NeRF With Geometry Priors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18365–18375, 2022. 3

[22] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:24741–24752, 2021. 2, 4, 6, 7

[23] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. MINE: Towards Continuous Depth MPI With NeRF for Novel View Synthesis. In *International Conference on Computer Vision (ICCV)*, pages 12578–12588, 2021. 3

[24] Zhong Li, Lele Chen, Celong Liu, Yu Gao, Yuanzhou Ha, Chenliang Xu, Shuxue Quan, and Yi Xu. 3d human avatar digitization from a single image. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–8, 2019. 2

[25] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 590–606, 2022. 5, 8

[26] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep

human parsing with active template regression. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(12):2402–2414, Dec 2015. 9

[27] 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. *Transactions on Graphics (TOG)*, 40(6):219:1–219:16, Dec. 2021. 2, 7

[28] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3, 6, 8

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2, 3

[30] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose Guided Person Image Generation. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 2

[31] Lu Mi, Abhijit Kundu, David Ross, Frank Dellaert, Noah Snavely, and Alireza Fathi. im2nerf: Image to neural radiance field in the wild. *arXiv preprint arXiv:2209.04061*, 2022. 3

[32] 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

[33] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 20–36, 2022. 3

[34] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4480–4490, 2019. 2

[35] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense Pose Transfer. In *European Conference on Computer Vision (ECCV)*, pages 123–138, 2018. 2

[36] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5480, June 2022. 2, 3

[37] 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 *International Conference on Computer Vision (ICCV)*, pages 14314–14323, 2021. 2, 6, 7

[38] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022. 2, 7

[39] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 2, 6, 7, 8

[40] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 9

[41] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*, 2023. 9

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, July 2021. 3, 4

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831, July 2021. 3, 5

[45] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *Transactions on Graphics (TOG)*, 2021. 3

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 3, 9

[47] 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 *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2, 4, 8

[48] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 2, 4

[49] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

[50] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 20154–20166, 2020. 3

[51] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2

[52] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12278–12291, 2021. 2

[53] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *International Conference on Computer Vision (ICCV)*, pages 15182–15192, 2021. 3

[54] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3835–3844, 2022. 3, 5

[55] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2021. 3

[56] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018. 2

[57] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2, 3, 4, 7

[58] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-View Neural Human Rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691, 2020. 2

[59] 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, June 2022. 4

[60] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022. 2, 3, 6

[61] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4479–4489, 2023. 3, 4

[62] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 14955–14966. Curran Associates, Inc., 2021. 2

[63] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose Guided Human Video Generation. In *European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2

[64] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, June 2023. 9

[65] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-Guided Human Animation from a Single Image in the Wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15034–15043, June 2021. 2

[66] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields From One or Few Images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 2, 3

[67] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 8

[68] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *Transactions on Graphics (TOG)*, 40(4):149:1–149:18, July 2021. 2

[69] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild. In *European Conference on Computer Vision (ECCV)*, pages 34–51, 2020. 5

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7

[71] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. HumanNeRF: Efficiently Generated Human Radiance Field From Sparse Inputs. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, 2022. 2, 7

[72] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):3170–3184, 2021. 2, 8