One-shot Implicit Animatable Avatars with Model-based Priors

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

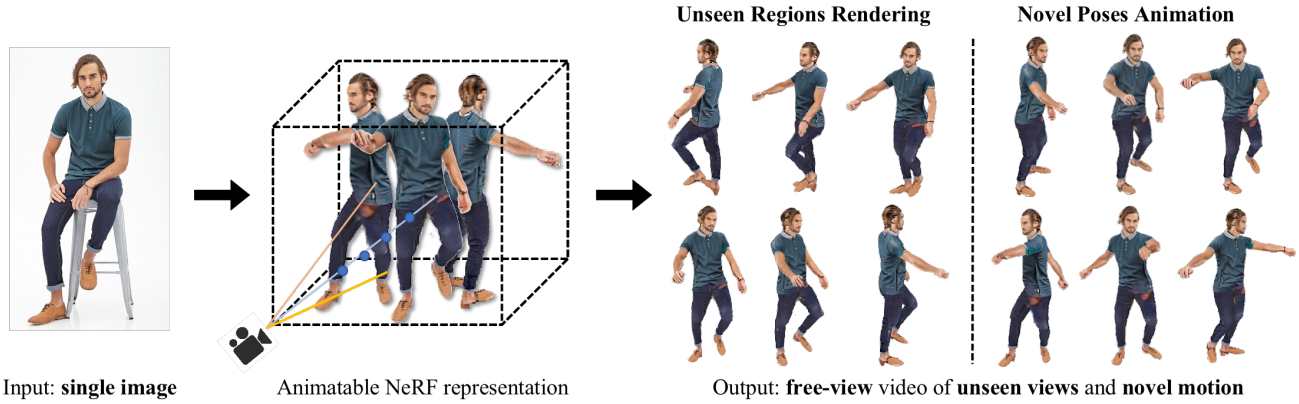

Figure 1. Our method creates free-viewpoint motion videos from a single image by constructing an animatable NeRF representation in one-shot learning.

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 easily reconstruct the body geometry and infer the full-body clothing from a single image, we leverage two priors in ELICIT: 3D geometry prior and visual semantic prior. Specifically, ELICIT introduces 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 pre-trained models. Both priors are used to jointly guide the optimization for creating plausible content in the invisible areas. 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 has outperformed current state-of-the-art avatar creation methods when only a single image is available. Code will be public for research purpose at https://elicit3d.github.io.

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 dancing motion synthesis and pose-specific view rendering. 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-controlled inputs are available, e.g., multi-view videos captured by well-calibrated multi-camera systems [41, 43, 62, 65, 69], or long monocular videos [61] where almost all parts of the human body are visible. Despite their excel-

*These authors contributed equally to this work.

lent performance, it is inconvenient (sometimes impossible) for ordinary users to obtain such high-quality dense inputs. Various methods have been proposed to tackle this data inefficiency. For example, ARCH [20] and ARCH++ [17] train reconstruction models with a single image input on large 3D scans datasets, but they do not generalize well to in-the-wild data. Human-specific neural radiance fields (NeRF) [14, 26, 29] trains conditioned 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 image inputs.

Instead of learning conditioned models from large-scale datasets [6, 68], recent works novelly introduce various regularizations on geometry [39] and appearance [23, 64] to avoid degeneration, which makes it possible to synthesize visually plausible views in a semi-supervised framework without extra training data. Nevertheless, due to the lack of 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, in this work, we propose a novel method, ELICIT, to learn human-specific neural radiance fields from a single image. We explicitly utilize body shape geometry and visual clothing semantic priors to guide the optimization and achieve free-view rendering from single image inputs.

In summary, our contributions are summarized below:

- We present a new approach, called ELICIT, for training an animatable neural radiance field from a single image. Our method enables the creation of animatable avatars that can be rendered from arbitrary views and poses.
- We introduce a CLIP-based semantic prior and a SMPL-based geometry prior to jointly guide the optimization of a NeRF model in order to produce plausible visual content for the invisible body areas. Additionally, we propose a body-part refinement sampling strategy that can further improve the visual details.
- We conduct quantitative and qualitative comparisons with previous state-of-the-art human-specific neural rendering methods in the single image input setting. ELICIT consistently outperforms existing methods in both novel view synthesis and avatar animation, and shows promising performance on in-the-wild images.

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 human pose transfer methods [1, 33, 37, 53], leveraging explicit temporal constraints [3, 66], optical flow estimation [60], and warping field [55, 67] to create temporally consistent posed-guided video from input videos or images. Most single-image-based 3D methods [50, 51] learn encoder-decoder models from large-scale human 3D scans data, among which ARCH [20] and ARCH++ [17] reconstruct animation-ready 3D representation. However, such data-driven methods depend on extra human-specific training data and are likely to encounter generalization issues on in-the-wild data.

Method	Subject data	Extra training data	Invisible area completion	Animatable
NeuralBody [43] Ani-NeRF [41] HumanNeRF [61]	multi-view images, monocular videos	**data-free**	✗	✓
PiFU [50] PaMIR [72] ARCH [20] ARCH++ [17]	**monocular images**	3D scans	✓	✓
MPS-NeRF [14] NHP [26]	sparse videos, multi-view images	multi-view videos	✗	✓
MonoNHR [8]	**monocular images**	multi-view images	✓	✗
EVA3D [18]	**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 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) can generate animatable videos of the character.

Recent works about human-specific neural radiance fields reconstruct animatable 3D human NeRF representation from multi-view or single-view video (For NeRF, see [35]). Most of them do per-subject optimization on an implicit model, using the whole video sequence as training data. Among which [43] learns structured latent codes on SMPL [31] mesh vertices, other methods construct the representation in a canonical space by modeling pose-driven deformation [41, 42, 56, 61, 65, 71]. Although such methods can synthesize impressive results, they are only applicable with dense inputs that cover most areas of the human body. Our approach allows the creation of an animatable realistic character from a single image input or a short video clip, which is more user-friendly and flexible in 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. Also, it is difficult for the model to synthesize regions in the new view which is not visible for the input due to occlusion. Some existing methods utilized learned prior about scene geometry and appearance in a data-driven manner, e.g., generative adversarial models [54, 57], supervised learning [6, 25, 59, 68] and unsupervised learning [34] for conditioned NeRF. Most of them only work on simple 3D shapes [5]. Eg3D [4] and CG-Nerf [24] work on specific kinds of objects such as human faces using conditional generative NeRF.

There are also non-data-driven methods introducing pri-

ors from off-the-shelf models, including depth cues [11,27] and other knowledge such as object geometry [39]. SinNeRF [64] and DietNeRF [23] 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 [45] is one of the latest cross-modality representation learning methods. CLIP-pretrained models have been widely applied to text-driven image generation [46,47,49]. Recently, some work started incorporating CLIP and radiance fields for 3D-aware synthesis tasks. DietCLIP [23] synthesizes view-consistent novel views from sparse view input, with a CLIP-based loss as a regularization on NeRF. CLIP-NeRF [58] directly apply joint image-text latent space for conditioned nerf for NeRF manipulation with multi-model inputs. LaTeRF [36] uses CLIP loss to extract objects of interest from the scene, similar to the texture cue. CLIP-based image similarity. AvatarCLIP [19] and dream fields [22] apply CLIP on the optimization process for text-driven 3D avatar generation. In our work, we further explore the potential of CLIP-driven NeRF by using it for human-specific rendering from a single image.

Concurrent works. Recently, MonoNHR [8] propose a data-driven approach that renders free-viewpoint images of a character from only a single image input by a conditioned NeRF. Besides, EVA3D [18] learns an unconditional 3D human generative model on DeepFasion [30], a large-scale 2D human image dataset. EVA3D can reconstruct 3D humans from a single image by GAN inversion [9,48] but has limited generalizability because of the biased distribution of the training datasets. We summarize the difference between our method and related methods in Tab. 1.

3. Method

3.1. Problem Specification

We formulate our 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 [32], 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 [61] is a human-specific variant of neural radiance field(NeRF), which supports free-view rendering of a moving character from monocular video inputs. It can handle complex body motions by a motion field mapping from observation to a canonical space. 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:

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

, where $F_c :\rightarrow (\boldsymbol{c}, \sigma)$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 $T : (\boldsymbol{x}_o, \boldsymbol{p}) \rightarrow \boldsymbol{x}_c$ 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 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.

CLIP [45] is a visual-language model pre-trained on 400 million diverse image-text pairs data collected from a massive web scrape. The model consists of an image encoder E_I and a text encoder E_T, which are both transformer-based models. CLIP learns a joint embedding space of text and image by pulling the embeddings of paired images and texts and pushing unpaired ones apart. The CLIP-pretrained image encoder can capture view-consistent high-level visual attributes of images like the art style, colors, and high-level semantic attributes, including object tags and categories [15,58]. It also can provide 3D-aware prior knowledge for novel view synthesis [23].

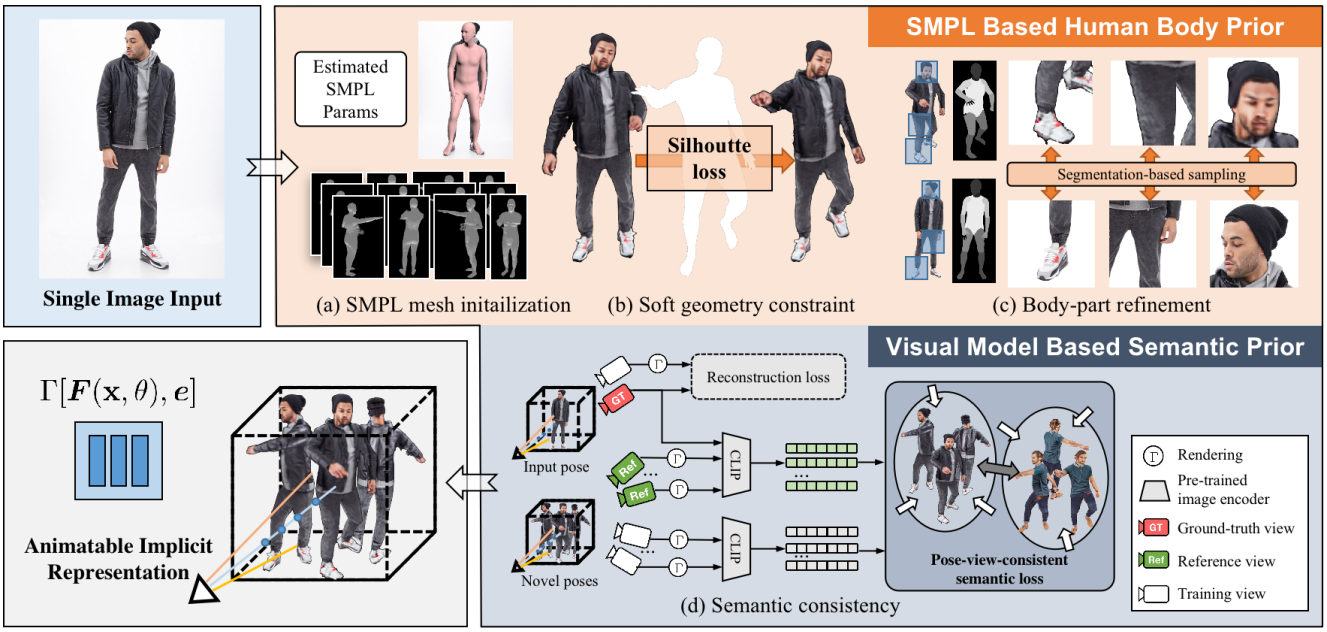

Figure 2. **Method overview.** Our method takes a single source image of a person to generate an animatable avatar, which can be applied to generate pose-guided free-view renderings of the person animated by any target motion in SMPL format. Specifically, ELICIT trains an animatable implicit human representation of HumanNeRF by one-shot prior-based learning. We introduce two model-based prior to guide the optimization, SMPL-based Human Body Prior and Visual-Model-based Semantic Prior. For the human body prior, we first use (a) multi-view video frames rendered by SMPL meshes to initialize the pose-conditioned geometry of the implicit representation. (b) During training, a silhouette loss is used to constrain synthesized geometry and body poses, and (c) a segmentation sampling strategy is used to refine body-part details. Moreover, the semantic prior plays an essential role in one-shot learning, which provides (d) pose-view-consistent semantic supervision for novel views of novel poses with a powerful pre-trained visual model.

3.3. Prior-driven One-shot Learning for Single-image-based 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 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 neighboring view sampled by specific rules.

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 [61].

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

, 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, (3) and 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 [8, 14, 26] or 3D scans training data [17, 50, 52, 63], 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

Generating plausible 3D-aware contents for a clothed human, especially the body areas invisible in the input view, is a crucial challenge for single-image 3D human digitalization. One way to achieve that is by minimizing an embedding distance between synthesized novel views and the input view, which requires a visual model that captures appearance identity from different views of the subject. In other words, we need a visual model to embed the images from different views of 3D humans in a semantically meaningful space, where different views of the same person should be significantly closer than those of a different person. Moreover, in contrast to person-ReID [7] models, such models should also be able to capture detailed visual

attributes such as low-level textures to provide rich supervision signals for a vivid generation.

Recent works show that visual models pre-trained on large-scale image data could guide novel view synthesis from sparse input with semantic loss. DietNeRF [23] utilizes a CLIP pre-trained ViT to synthesize semantically consistent novel views, and SinNeRF [64] uses a DINO pre-trained ViT to introduce global structure prior for semi-supervision with pseudo semantic labels. Additionally, CLIP-NeRF [58] has evaluated that the CLIP pre-trained models embed images of 3D objects into a view-consistent semantic space. We conducted a similar evaluation and found that such embedding is also view-pose-consistent for human images, as shown in Fig. 3.

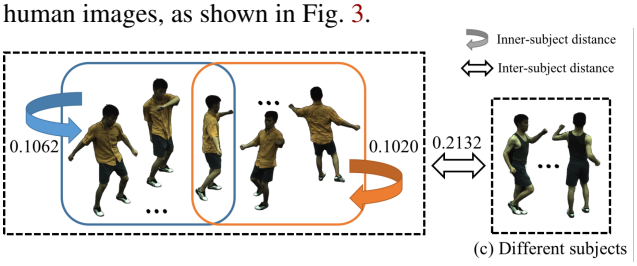

Figure 3. **View-pose-consistency of the CLIP embeddings.** The embedding distance of the same character under different views and poses is significantly smaller than the distance between two different characters.

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 [13, 19, 47, 58]. It enables the generation of vivid visual details with its ability to capture important high-level semantic properties and visual attributes of images like the art style and colors [15].

3.3.2 SMPL-based human body prior

By incorporating off-the-shelf pose estimation models [28], we can obtain knowledge about approximate human shapes and body-part segmentation from SMPL. Our method utilizes this human-specific prior as geometric and semantic clues for 3D human reconstruction and animation by introducing an SMPL-based NeRF initialization, a soft geometry constraint in training, and a fine-grained sampling strategy for body-part refinement.

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 [19], 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 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 $\boldsymbol{E}_s = \{\mathbf{e}_s^i\}_{i=1}^m$ 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 mentioned above, we have introduced human body geometry prior to our NeRF-based implicit representation by SMPL meshes so that the model can render approximate shapes for the target character in certain poses. Although the CLIP embedding is similar among different poses and views, optimizing the model only with semantic loss may lead to degenerated results with inconsistent rendered poses and missing body parts.

For this issue, we introduce a soft geometry constraint loss based on the assumption that the estimated SMPL meshes are close to the geometry of the naked body of the target character, thus, should be contained by the actual shape of the clothed character. This loss function is a masked version of silhouette loss [70], consisting of an MSE loss and a one-way chamfer loss for the silhouette boundary, only computed for the rendered alpha pixels contained by the SMPL silhouette. Given SMPL silhouette mask S and rendered alpha map A, we compute the loss as follows:

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

, where $\circ$ is the element-wise product, $\text{Edge}(S)$ computes the edge of mask S, $A(p)$ is the pixel value of A at p, $S(p)$ is the pixel-wise mask of S at p. This constraint maintains the character's pose in target motion during training. Also, it ensures that the semantic loss supervision guides the optimization in the direction of plausible completions of unseen content rather than making the rendered character pose closer to the reference view.

Body-part refinement. CLIP loss between unseen views or unseen poses and the input view enforces global semantic consistency. To get rid of the resolution constraint of the pre-trained CLIP and improve the quality of the synthesized image, SinNeRF [64] calculates semantic feature loss between the extracted features of random sampled local patches instead of complete global views. However, in

Dataset	Subjects	Novel View Synthesis									Novel Pose Synthesis								
		PSNR↑			SSIM↑			LPIPS↓			PSNR↑			SSIM↑			LPIPS↓		
		NB	NHP	ELICIT	NB	NHP	ELICIT	NB	NHP	ELICIT	NB	AniNeRF	ELICIT	NB	AniNeRF	ELICIT	NB	AniNeRF	ELICIT
ZJU-MoCap	313*	21.1	24.0	**24.7**	0.828	0.896	**0.931**	0.237	0.168	**0.068**	20.9	20.8	**24.5**	0.800	0.816	**0.933**	0.278	0.237	**0.061**
	315*	17.8	19.0	**22.1**	0.779	0.828	**0.925**	0.255	0.222	**0.073**	18.2	18.5	**21.8**	0.747	0.782	**0.921**	0.309	0.297	**0.075**
	377*	19.4	23.1	**23.6**	0.784	0.879	**0.931**	0.273	0.193	**0.074**	18.2	19.5	**24.6**	0.686	0.788	**0.939**	0.362	0.293	**0.063**
	386*	21.2	25.1	**26.9**	0.886	0.882	**0.933**	0.312	0.225	**0.081**	22.0	24.3	**24.5**	0.745	0.824	**0.910**	0.357	0.289	**0.095**
	387	17.7	21.0	**22.9**	0.760	0.865	**0.920**	0.283	0.190	**0.088**	18.5	19.8	**22.7**	0.743	0.813	**0.922**	0.316	0.266	**0.083**
	390*	19.2	23.6	**25.0**	0.706	0.871	**0.924**	0.357	0.208	**0.086**	20.2	20.8	**24.0**	0.675	0.773	**0.921**	0.377	0.298	**0.076**
	392*	22.8	24.7	**25.6**	0.840	0.887	**0.936**	0.229	0.193	**0.072**	20.9	19.3	**24.3**	0.742	0.786	**0.931**	0.339	0.293	**0.070**
	393	19.6	24.0	**24.5**	0.779	0.892	**0.930**	0.264	0.168	**0.074**	18.8	19.6	**23.7**	0.752	0.784	**0.926**	0.255	0.274	**0.072**
	394	18.1	24.1	**24.5**	0.680	0.869	**0.918**	0.346	0.196	**0.082**	19.4	22.8	**24.8**	0.715	0.858	**0.920**	0.337	0.181	**0.073**
	Average	19.9	23.1	**24.4**	0.795	0.875	**0.929**	0.276	0.196	**0.077**	19.7	20.3	**23.8**	0.736	0.796	**0.925**	0.324	0.276	**0.074**
Human 3.6M	S1*	18.3	**25.1**	24.2	0.710	0.896	**0.920**	0.328	0.121	**0.065**	21.6	21.5	**25.7**	0.843	0.842	**0.933**	0.224	0.212	**0.061**
	S5*	19.7	**25.1**	24.0	0.785	0.895	**0.918**	0.266	0.126	**0.076**	21.3	22.0	**23.5**	0.828	0.821	**0.924**	0.236	0.244	**0.070**
	S6*	19.6	23.5	**23.7**	0.777	0.839	**0.912**	0.257	0.160	**0.074**	21.7	21.2	**24.8**	0.820	0.812	**0.924**	0.230	0.237	**0.071**
	S7*	18.6	23.9	**23.9**	0.751	0.856	**0.919**	0.299	0.153	**0.076**	18.3	20.2	**25.3**	0.771	0.783	**0.928**	0.297	0.280	**0.070**
	S8	18.2	20.5	**22.1**	0.778	0.838	**0.912**	0.254	0.166	**0.081**	21.2	22.0	**25.2**	0.847	0.863	**0.932**	0.220	0.191	**0.063**
	S9	20.8	23.0	**25.0**	0.780	0.818	**0.910**	0.261	0.192	**0.082**	22.1	22.3	**25.9**	0.811	0.815	**0.925**	0.250	0.236	**0.069**
	S11	21.2	22.5	**25.5**	0.783	0.822	**0.921**	0.280	0.201	**0.082**	20.9	23.7	**25.7**	0.809	0.858	**0.934**	0.272	0.224	**0.064**
	Average	19.5	23.4	**24.1**	0.766	0.852	**0.916**	0.278	0.160	**0.077**	21.0	21.8	**25.2**	0.818	0.828	**0.928**	0.247	0.232	**0.067**

Table 2. Detailed quantitative results of **novel pose synthesis** and **novel pose synthesis** on **ZJU-MoCap** [43] and **Human 3.6M** [21] dataset in PSNR↑, SSIM↑(higher is better) and LPIPS↓(lower is better), our method outperforms in both datasets compared to the per-subject optimization baseline NeuralBody [43](NB) and Animatable NeRF [41, 42](AniNeRF), also the generalizable baseline Neural Human Performer(NHP) which is pretrained on part of the subjects. * Subjects included in training set of NHP.

human-specific rendering tasks, the appearance and semantic features vary a lot among different body parts. Directly applying such patch-based similarity loss may lead to the misalignment of synthetic contents.

Here we use a body-part-aware patch sampling strategy for each sampled view $V_{\text{train}} = (\theta_{\text{train}}, \mathbf{e}_{\text{train}})$, our strategy randomly samples a body part k (including the whole body) to refine. The rendered segmentation of SMPL can determine the corresponding region, $S^k_{\text{SMPL}}(\theta_{\text{train}}, \mathbf{e}_{\text{train}})$, explicitly defined by groupings of SMPL meshes. Accordingly, we adjust the training camera to render a local patch $V^k_{\text{train}} = (\theta_{\text{train}}, \mathbf{e}^k_{\text{train}})$ for this body part. We can also crop a corresponding reference patch V^k_s from the input image by the SMPL segmentation $S^k_{\text{SMPL}}(\theta_s, \mathbf{e}_s)$.

To improve the refinement quality, we further utilize the prior knowledge about body orientation. Existing NeRF optimization-based 3D generative approaches have encountered the Janus problem [22, 44], which means that the learned 3D model has multiple faces. For this issue, instead of sampling reference patches from the input image V_s for every training V_{train}, here we sample neighboring rendered views as reference $V^k_{\text{ref}} = (\theta_{\text{train}}, \mathbf{e}^k_{\text{ref}})$ for some body parts k viewed from a specific direction. For example, when the rear view of the character's head is sampled for training, we select a reference view from the left-rear or the right-rear direction, which could be partially visible in the input image. We use this rendered image as the reference to avoid generating another face on the back side and robustly solve most bad cases of the Janus problem in our task.

4. Experiments

4.1. Datasets

We evaluate on two multi-view human video datasets (ZJU-MoCap [43] and Human3.6M [21]) and a 2D human image dataset (DeepFashion [30]). We use all of the 9 subjects in ZJU-MoCap and the "Posing" video of all 7 subjects in Human3.6M to evaluate free-view animation. The single-image inputs are sampled from the first camera of ZJU-MoCap and the third camera of Human3.6M, along with annotated SMPL parameters, camera matrices, and segmentation mask. We directly apply the annotated motion sequence of the video clip for animation. In addition, we use high-resolution full-body photos from DeepFashion [30] for the evaluation on humans with various cloth styles.

4.2. Comparisons

As far as we know, NeRF-based human-specific novel view synthesis can be categorized into per-subject optimization methods and generalizable methods. We select three state-of-the-art methods as baselines: Neural Body [43](NB) and Ani-NeRF [41] from per-subject optimization methods, and Neural Human Performer [26](NHP) from generalizable methods. All three methods utilize SMPL-based human body priors, among which NB and Ani-NeRF support novel pose synthesis for animation. We adapt these baselines to take single-image inputs for a fair comparison. We perform overall comparison in two different task settings: novel view synthesis for free-view rendering and novel pose synthesis for character animation.

Metrics. Following previous works [26, 29, 43], we report the results on two standard metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). We also report a perceptual similarity metric, LPIPS, used in [61, 71], to measure the visual quality.

Comparison on novel view synthesis. For ZJU-MoCap and Human3.6M, we uniformly sample 10 frames from each subject video and evaluate the results on all available camera views except the input one. For per-subject optimization methods NB and ELICIT, we optimize one model for each frame. For NHP, we train one model for each dataset and report both results on training subjects and testing subjects. As shown in Tab. 2, ELICIT outperforms under the metrics for all subjects, except for subjects S1 and

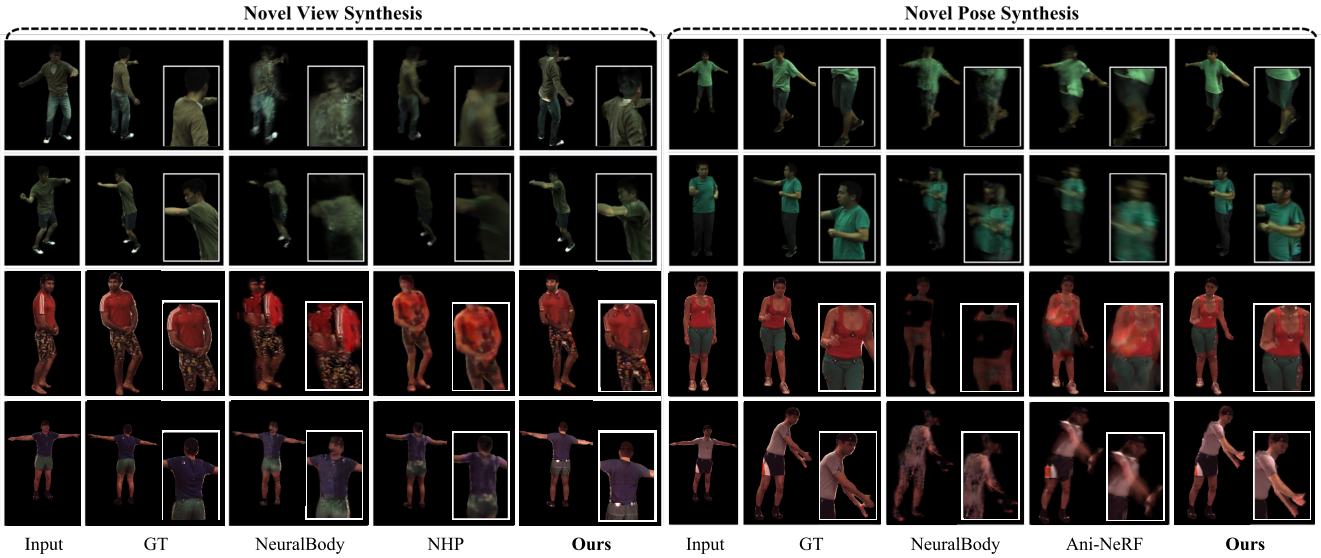

Figure 4. **Overall comparison.** Compared with state-of-the-art NeRF based methods [26,41,43] on novel view synthesis and novel pose synthesis, ELICIT can generate human 3D renderings with more consistent apperance and more realistic visual details from only a single image.

S5 on PSNR, which are included in the training set of NHP. The superior performance on the SSIM and LPIPS metrics reveals the advantage of our ELICIT in the perceptual quality of the rendering results.

Concurrently, MonoNHR [8] reports state-of-the-art single image-based novel view synthesis results on ZJU-MoCap. We follow their evaluation setting and present our results in Tab. 3 for comparison. We perform better on SSIM which means our results are more perceptually similar to the ground-truth image. Notably, we are slightly lower on PSNR, which is known to favor smooth results compared with MonoNHR, but MonoNHR is pre-trained on seven subjects from ZJU-MoCap and does not support animation.

Method	PSNR↑	SSIM↑
pixelNeRF [68]	22.13	0.8604
NHP [26]	24.01	0.8953
MonoNHR [8]	**25.36**	0.9093
ELICIT	24.63	**0.9326**

Table 3. Comparison of state-of-the-art NeRF based neural human rendering methods [8, 26, 68] on ZJU-MoCap subjects {313, 377, 392}. All the baseline results are reported by MonoNHR [8]. Except for a slightly lower PSNR than MonoNHR, ELICIT has a better performance than other methods under quantitative comparison.

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 [42], which reports their best results. As shown in Tab. 2, our method also produces high-quality synthesis when generalized to novel poses.

We show sampled novel view and pose synthesis results in Fig. 4. Compared to the latest NeRF-based methods, ELICIT performs better in reconstructing visual details and inference of occluded contents of clothed human bodies.

Figure 5. **Qualitative results** of PIFu [50], PaMIR [72] and ELICIT on DeepFashion. ELICIT generates more realistic details in occluded views.

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 our approach on 2D human images from DeepFashion [30] dataset, with SMPL parameters estimated by an off-the-shelf pose estimation model [28]. Among previous data-

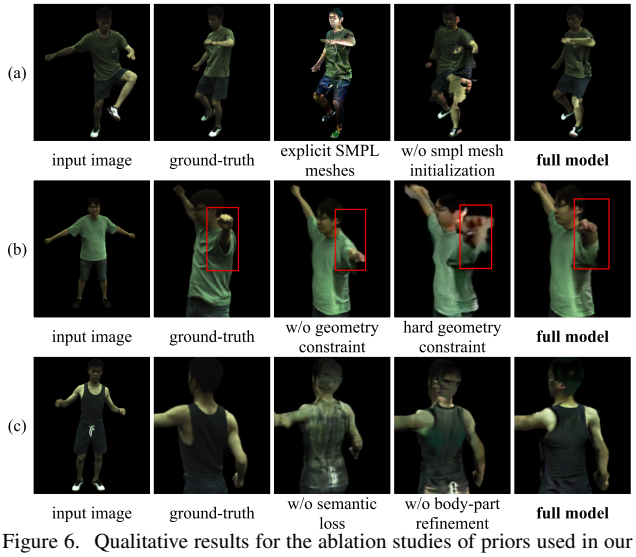

Figure 6. Qualitative results for the ablation studies of priors used in our method, selected from ZJU-MoCap [43] dataset.

driven non-NeRF methods, PIFu [50] and PaMIR [72] support both reconstruction of geometry and texture from single image input, which have also shown impressive results on DeepFashion dataset. Here we choose these two methods for qualitative comparison. Figure 5 illustrates that our training-data-free one-shot method generalize 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 texture parameters during training. Such an explicit model produces noisy textures, and its SMPL-based geometry also has a gap with the actual human shape. As shown in Tab. 4 and Fig. 6(a), an implicit representation such as HumanNeRF, which models character appearance with a spatially continuous function, is necessary for our one-shot learning.

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 Tab. 4 and Fig. 6(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.

Soft geometric constraint. Optimizing the model without geometric constraints may lead to error poses, as shown in Fig. 6(b), which can affect the character motion in video results. Moreover, in contrast to matching the SMPL geometry directly by a hard constraint of silhouette loss, we penalize the silhouette misalignment only internally. 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 actual body shape of the input view.

CLIP-based semantic loss. As shown in Fig. 6(c), semantic loss plays a vital role in generating plausible content in invisible areas of the input view. We also compare the performance of different pre-trained visual models, including an DINO [2] ViT used by SinNeRF [64], an ViT/L-14 [10, 12] supervised pre-trained on ImageNet, an unsupervised pre-trained ViT/L-14 by MAE [16], also a lighter version of CLIP ViT/B-32. As shown in Fig. 7, among all the ViT models we evaluated, CLIP ViT/L-14 shows best performance in capturing 3D-aware human body structure and generating vivid visual details, and the two CLIP pre-trained models have a better performance on head structure than Image pre-trained models. It indicates that rich pre-training data and a large model capacity of the visual model are key factors for the effectiveness of the semantic loss.

Figure 7. Qualitative results for the ablation studies of vision models used for the semantic loss, selected from DeepFashion [30] dataset. The CLIP ViT/L-14 model we use produce best detailed geometry and textures.

Body part refinement. Figure 6(c) illustrates that certain small areas' error content can significantly affect the overall visual quality, such as artifacts on clothes textures and missing hair on the back of the head. Sampling local patch of specific body parts for refinement enables the model to generate high-resolution results by volume rendering, and synthesize vivid geometric and textual details.

Setting	PSNR↑	SSIM↑	LPIPS↓
explicit SMPL mesh	19.68	0.8889	0.1129
w/o SMPL mesh initialization	21.41	0.8961	0.1138
w/o semantic loss	24.04	0.9275	0.0775
w/o geometric constraint	23.98	0.9313	0.0679
hard geometry constraint	22.81	0.9104	0.0915
w/o body part refinement	23.99	0.9328	**0.0661**
full model	**24.46**	**0.9343**	0.0674

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

5. Conclusion

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 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.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 8

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody Dance Now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[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. 2

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[7] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 4

[8] Hongsuk Choi, Gyeongsik Moon, Matthieu Armando, Vincent Leroy, Kyoung Mu Lee, and Gregory Rogez. Mononhr: Monocular neural human renderer. *arXiv preprint arXiv:2210.00627*, 2022. 2, 3, 4, 7, 13

[9] 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. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[11] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[13] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 5

[14] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *arXiv preprint arXiv:2203.16875*, 2022. 2, 4

[15] 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. 3, 5

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 8

[17] 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, 4

[18] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 2, 3

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

[20] 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

[21] 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

[22] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3, 6

[23] 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, Montreal, QC, Canada, Oct. 2021. IEEE. 2, 3, 5

[24] Kyungmin Jo, Gyumin Shim, Sanghun Jung, Soyoung Yang, and Jaegul Choo. Cg-nerf: Conditional generative neural radiance fields. *arXiv preprint arXiv:2112.03517*, 2021. 2

[25] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. GeoNeRF: Generalizing NeRF With Geometry Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 2

[26] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. volume 34, pages 24741–24752, 2021. 2, 4, 6, 7

[27] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 3

[28] 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*, pages 590–606. Springer, 2022. 5, 7

[29] 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*, 40(6):219:1–219:16, Dec. 2021. 2, 6

[30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3, 6, 7, 8

[31] 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

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

[33] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose Guided Person Image Generation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[34] 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. 2

[35] 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

[36] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. *arXiv preprint arXiv:2207.01583*, 2022. 3

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

[38] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 13

[39] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5480, New Orleans, LA, USA, June 2022. IEEE. 2, 3

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 13

[41] 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. 1, 2, 6, 7, 13

[42] 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, 6, 7, 13

[43] 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, 6, 7, 8, 13

[44] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 6

[45] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. 3

[46] 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

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, July 2021. 3, 5

[48] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 3

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3, 13

[50] 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, 4, 7, 8

[51] 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

[52] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 4

[53] 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

[54] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 20154–20166. Curran Associates, Inc., 2020. 2

[55] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[56] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose. In *Advances in Neural Information Processing Systems*, volume 34, pages 12278–12291. Curran Associates, Inc., 2021. 2

[57] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2

[58] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3, 5

[59] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2

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

[61] 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, 4, 6, 13

[62] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-View Neural Human Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 1

[63] 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

[64] 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, 5, 8

[65] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion. In *Advances in Neural Information Processing Systems*, volume 34, pages 14955–14966. Curran Associates, Inc., 2021. 1, 2

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

[67] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15034–15043, Nashville, TN, USA, June 2021. IEEE. 2

[68] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields From One or Few Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 7

[69] 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. *ACM Transactions on Graphics*, 40(4):149:1–149:18, July 2021. 1

[70] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 34–51, Cham, 2020. Springer International Publishing. 5

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2, 6

[72] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 2, 7, 8

Appendices

A. Implementation Details

A.1. Optimization

ELICIT constructs animatable avatars by a two-stage optimization. For the SMPL-based initialization stage, we optimize the model with only the reconstruction loss in Eq. 4. For the stage of one-shot training on the input image and target motion, we optimize the model with L_{recon}, L_{CLIP} and L_{sil} as follows:

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

where the loss weight $\lambda_{\text{CLIP}} = 0.1$, $\lambda_{\text{sil}} = 0.01$. These two stages takes $T_{\text{init}} = 15K$ and $T_{\text{train}} = 20K$ iterations respectively. And the entire optimization costs about 5 hours on 4 NVIDIA Tesla V100 GPUs. Besides, We follow the same settings of the optimizer, learning rate, and ray sampling in HumanNeRF [61].

A.2. Sampling Strategies for One-shot Training

Here we describe our detailed sampling strategy in the optimization stage of one-shot training.

For each iteration, we randomly decide whether to sample a novel view from $\{(\theta_i, \mathbf{e}_j)\}_{i=1,j=1}^{L,M}$ or the input view $V_s = (\theta_s, \mathbf{e}_s)$ with a probability of $p_{\text{novel}} = 0.5$. If $V_{\text{train}} = V_s$, we follow HumanNeRF to sample a pair of patches for reconstruction. Otherwise, we randomly select a body part k (including the whole body) with weighted probability$\{p_{\text{part}}^k\}_{k=1}^K$, and sample a training patch V_{train}^k which is decided by the bounding box of SMPL rendered body-part segmentation $S_{\text{SMPL}}^k(V_{\text{train}})$.

After sampling the training patch, we sample the reference patch from the V_s or other views of the same pose $\{(\theta_{\text{train}}, \mathbf{e}_j)\}_{j=1,j\neq i}^M$. The camera views of the current pose are divided into front views, rear views, left views, and right views according to the body rotation angle. We assume that the input image is close to the front view of the character. If a rear view of specific body parts (e.g. head, upper body, or whole body) is sampled as the training view, we randomly sampled nearest views from left views and right views as V_{ref}. Then we render body-part patch V_{ref}^k by our NeRF model as reference. Otherwise, the reference patch will be constructed by the resized patch V_s^k cropped from the input image.

We set the size of patches in training to 224×224 for all experiments, the same as the input resolution of the CLIP ViT/L-14 model we use for semantic prior.

B. Additional Results

For a more comprehensive comparison with MonoNHR [8], which reports state-of-the-art results on human-specific novel view synthesis from single monocular input, we show qualitative results of MonoNHR and ELICIT on ZJU-MoCAP dataset. In particular, we take the novel view synthesis results from MonoNHR's official qualitative video*, and use the same input view on ELICIT for comparison. Figure 8 shows that ELICIT generates more realistic details on human faces, bodies, and clothing, although in Tab. 3 MonoNHR is slightly higher on PSNR.

Figure 8. Qualitative comparison of novel view synthesis between ELICIT and MonoNHR [8].

C. Limitations

The human body geometry prior we use requires well-aligned SMPL annotation of body shape and postures. When body parts such as hands and legs are heavily misaligned, it produces artifacts as it fails to reconstruct image input on the model which is incorrectly initialized, or fails to sample reference patches for body-part refinement. For similar reasons, it's also difficult for our method to model the hand geometry and complex clothing geometry precisely.

Besides, a computational cost of 5 hours on 4 Tesla V100s for each avatar is still too expensive for application. Using more efficient human-specific NeRFs and improving the training pipeline might be helpful for this issue.

D. Future works

We will further explore the model-based priors that can potentially improve ELICIT, such as image diffusion models [38,49] for semantic prior and SMPL-X [40] for human-body prior. In addition, we will also work to enhance the versatility of our one-shot training framework, including complicity with different kinds of inputs (e.g., multiple images, short video, images with a text description) and various human-specific NeRFs(e.g., NeuralBody [43], AniNeRF [41,42]).

*https://www.youtube.com/watch?v=9-hfGf7dRw4