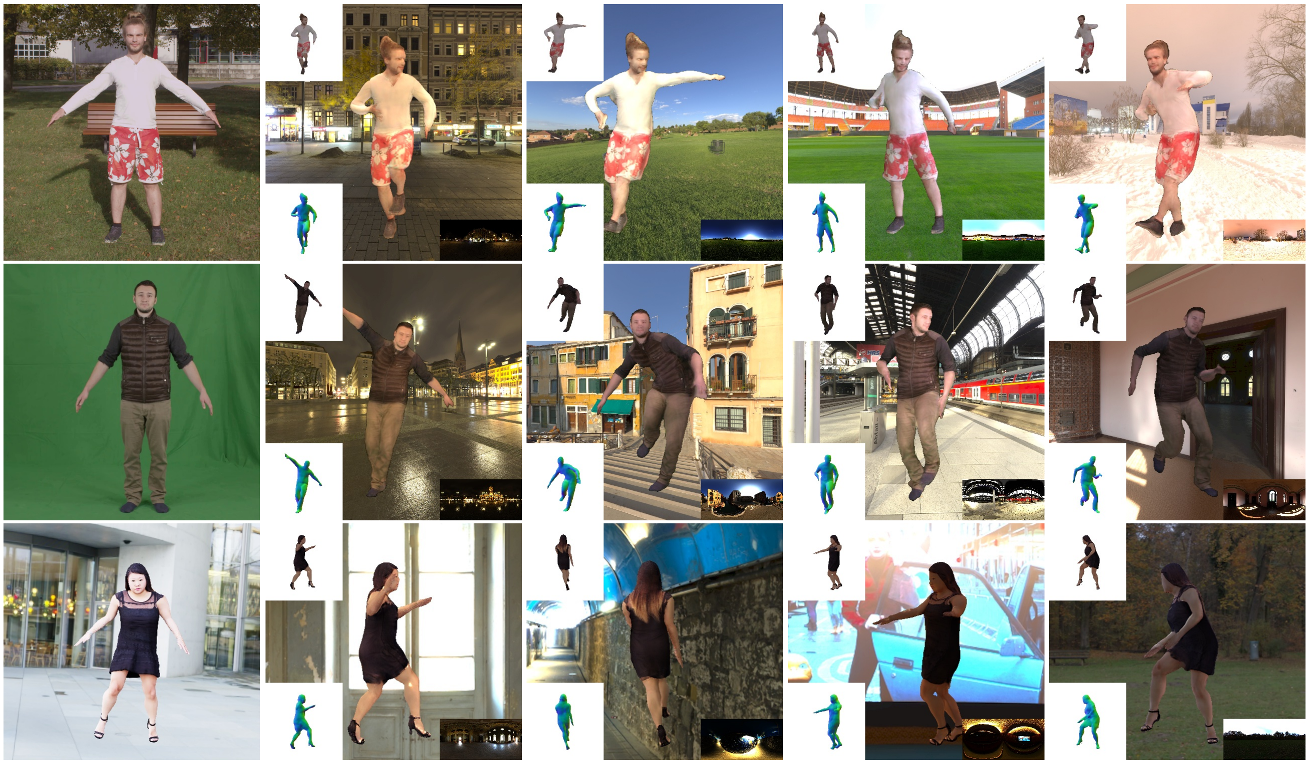
# RANA: Relightable Articulated Neural Avatars

Umar Iqbal[1*] Akin Caliskan[2*] Koki Nagano[1] Sameh Khamis[1] Pavlo Molchanov[1] Jan Kautz[1]
[1]NVIDIA [2]Flawless AI
https://nvlabs.github.io/RANA/



We present, RANA, an approach for learning dynamic and relightable full-body avatars from monocular RGB videos. A training frame of the person is shown in the first column. RANA can synthesize images of the person under novel poses, viewpoints, and lighting environments. In the insets, we show the synthesized albedo image, the normal map, and the target HDRI light map.

## Abstract

*We propose RANA, a relightable and articulated neural avatar for the synthesis of humans under arbitrary viewpoints, body poses, and lighting. We only require a short video clip of the person to create the avatar and assume no knowledge about the lighting environment. We present a novel framework to model humans while disentangling their geometry, texture, and lighting environment from monocular RGB videos. To simplify this otherwise ill-posed task we first estimate the coarse geometry and texture of the person via SMPL+D model fitting and then learn an articulated neural representation for higher quality image synthesis. RANA first generates the normal and albedo maps of the person in any given target body pose and then uses spherical harmonics lighting to generate the shaded image in the target lighting environment. We also propose to pre-train RANA using synthetic images and demonstrate that it leads to better disentanglement between geometry and texture while also improving robustness to novel body poses. Finally, we also present a new photo-realistic synthetic dataset, Relighting Human, to quantitatively evaluate the performance of the proposed approach.*

## 1. Introduction

Articulated neural avatars of humans have numerous applications across telepresence, animation, and visual con-

*equal contribution. The work was partially done during AC's internship at NVIDIA.

tent creation. To enable the widespread adoption of these neural avatars, they should be easy to generate and animate under novel poses and viewpoints, able to render in photo-realistic image quality, and easy to relight in novel environments. Existing methods commonly aim to learn such neural avatars using monocular videos [47, 46, 45, 54, 66, 58, 31]. While this allows photo-realistic image quality and animation, the synthesized images are always limited to the lighting environment of the training video. Other works directly tackle relighting of human avatars but do not provide control over the body pose [23, 41]. Moreover, these approaches often require multiview images recorded in a Light Stage for training, which is limited to controlled settings only. Some recent methods aim to relight RGB videos of a dynamic human but do not provide control over body pose [16].

In this work, we present the Relightable Articulated Neural Avatar (RANA) method, which allows animation of people under any novel body pose, viewpoint, and lighting environment. To create an avatar, we only require a short monocular video clip of the person in unconstrained environment, clothing, and arbitrary body pose. During inference, we only require the target body pose and target lighting information.

Learning relightable neural avatars of dynamics humans from monocular RGB videos recorded in unknown environments is a challenging problem. First, it requires modeling the complex human body articulations and geometry. Second, in order to allow relighting under novel environments, the texture, geometry, and illumination information have to be disentangled, which is an ill-posed problem to solve from RGB videos [8]. To address these challenges, we first extract canonical coarse geometry and texture information from the training frames using a statistical human shape model SMPL+D [40, 5, 36]. We then propose a novel convolutional neural network that is trained on synthetic data to remove the shading information from the coarse texture. We augment the coarse geometry and texture with learnable latent features and pass them to our proposed neural avatar framework, which generates refined normal and albedo maps of the person under the target body pose using two separate convolutional networks. Given the normal map, albedo map, and lighting information, we generate the final shaded image using spherical harmonics (SH) lighting [49]. During training, since the environment lighting is unknown, we jointly optimize it together with the person's appearance and propose novel regularization terms to prevent the leaking of lighting into the albedo texture. We also propose to pre-train the avatar using photo-realistic synthetic data with ground-truth albedo and normal maps. During pretraining, we train a single model for multiple subjects while having separate neural features for each subject. This not only improves the generalizability of the neural avatar to novel body poses but also learns to decouple the texture and geometry information. For a new subject, we only learn a new set of neural features and fine-tune the avatar model to capture fine-grained person-specific details. In our experiments, the avatar for a novel subject can be learned within 15k training iterations.

Since no dataset exists to evaluate the renderings of avatars in terms of both novel light and pose synthesis, we also propose a novel photo-realistic synthetic Relighting Human Dataset (RHD) with ground truth albedo, normals, and lighting information. We also perform a qualitative evaluation of RANA on the People Snapshot dataset [5] to compare with other baselines.

Our contributions can be summarized as follows:

- We present, RANA, a novel framework for learning relightable articulated neural avatars from short unconstrained monocular videos. The proposed approach is very easy to train and does not require any knowledge about the environment lighting of the training video.
- Our proposed approach can synthesize images of articulated humans under any arbitrary body pose, viewpoint, and lighting. It can also be used for relighting videos of dynamic humans.
- We present a new photo-realistic synthetic dataset for quantitative evaluation and to further the research in this direction.

## 2. Related work

**Mesh Based Human Avatars.** These methods represent human avatars using a rigged mesh and an associated texture map. Earlier methods captured human avatars using multi-view cameras [12, 53] or with the help of depth sensors [9, 71]. However, their adoption remained limited due to the constrained hardware requirements. The recent works, therefore, focus on creating the avatars from monocular videos [5, 4, 31, 10], or images [36, 64, 27, 25, 6, 10, 11]. The methods [5, 4, 36] use body model fitting to capture the humans, while more recent methods use data-driven implicit functions combined with pixel-aligned features [50] for human reconstruction [71, 27, 64, 6]. Self-Recon [31] combines both explicit and implicit meshes to reconstruct the detailed geometry of the person from self-rotating videos, however, it requires manual human effort to make the avatar animatable. The main limitation of these methods, however, is that the shading information is baked into the texture. Hence, the avatars cannot be rendered with novel lights. The recent methods PHORUM [6] and S3F [17] are the only exceptions, however, they create avatar from a single image and rely on data-driven priors to hallucinate the occluded regions of the person. Hence, the generated images may not be the true representation of the

| novel view | novel pose | generalizable | relightable | Method |
|---|---|---|---|---|
| ✓ | | | | NeuralBody [46], SelfRecon [31] |
| ✓ | ✓ | | | AnimatableNeRf [45, 14], NeuMan [32] |
| ✓ | ✓ | ✓ | | ANR [47], TNA [52], StylePeople [22] |
| ✓ | | | ✓ | Relighting4D [16] |
| ✓ | ✓ | ✓ | ✓ | **RANA (Ours)** |

Table 1. Comparison of some of the representative methods for neural human avatar creation. Ours (RANA) is the only method that allows novel view, pose and light synthesis.

person. In contrast, our approach uses video data to capture a detailed human representation, while also allowing the rendering of the person in novel lighting.

**Neural Human Avatars.** More recent methods learn a neural representation of the person and use neural renderers [57] to directly generate photo-realistic images in the target body pose and viewpoints [7, 20, 13, 59, 26, 63]. These methods are generally classified into 2D or 3D neural rendering based methods [57]. The 3D neural rendering methods represent the person using neural radiance fields [42] and render the target images using volume rendering [46, 62, 32, 45, 14, 58, 24, 38]. The 2D neural rendering methods, on the other hand, use CNNs to render the images [47, 66, 22, 70]. One limitation of the 3D neural rending methods is that the avatar has to be created from scratch for each person. Some methods learn generalizable avatars but require examplar images of the person during inference [34, 48]. In contrast, the 2D based methods offer some generalizability by sharing the neural renderer across multiple subjects [47]. Our method falls into the 2D neural rendering paradigm as we use CNNs to generate the albedo and normal images of the person. In particular, we take inspiration from ANR [47] for designing our framework. Tab. 1 compares existing neural avatar creation methods. Ours is the only method that allows synthesis under novel poses, viewpoints, and lighting, while also being generalizable.

**Human Relighting.** Relighting of human images has been studied extensively in the literature [18, 51, 33, 55, 72, 61, 56, 30, 37, 35, 69, 17]. However, these methods cannot render relighted images in novel body poses and viewpoints. Some recent methods allow relighting from novel views but provide no control over the body pose [23, 43, 16, 65]. The method TexMesh [71] also provides control over body pose but requires RGB-D videos for training and assumes that the environment map of the training video is known. Our approach, in contrast, provides full control over the body pose, viewpoint, and lighting, and learns the avatar from RGB videos with unknown lighting.

# 3. Method

Our goal is to learn a relightable articulated neural avatar, RANA, that can synthesize high-quality images of the person under any target body pose, viewpoint, and lighting. To create the avatar, we use a small video sequence $\mathbf{I}{=}\{I_{\mathrm{f}}\}_{f=1}^{F}$ with $F$ video frames and assume no knowledge about the lighting environment and body poses present in the video. We parameterize RANA using the SMPL [40] body model to control the animation and use Spherical Harmonics (SH) [49] lighting to model the lighting. During inference, we only need the target body pose and target lighting information in the form of SH coefficients and do not require any exemplar images for ground-truth. Learning RANA from a monocular video requires capturing the geometry and appearance of the dynamic human while also disentangling the shading information. In order to tackle this ill-posed problem, we first capture the coarse geometry using the SMPL+D fits (Sec. 3.1). We use the coarse geometry to extract a coarse texture map from the training images which is converted to an albedo texture map using a convolutional network (Sec. 3.2). We then propose RANA that generates the refined albedo and normal maps. The refined normal maps are used to obtain the shading map using SH lighting which is combined with the refined albedo map to obtain the final image in the target body pose and light (Sec. 3.3). An overview of our method can be seen in Fig. 1. In the following, we describe each of these modules in greater detail.

## 3.1. Coarse Geometry Estimation

Given the training frames, we first estimate the coarse geometry of the person including the clothing and hair details. For this, we employ the SMPL+D [4, 36] variant of the SMPL body model [40]. SMPL is a linear function $M(\theta, \beta)$ that takes the body pose $\theta \in \mathbb{R}^{72}$ and shape parameters $\beta \in \mathbb{R}^{10}$ as input and produces a triangulated mesh $\mathbf{M} \in \mathbb{R}^{V\times 3}$ with $V{=}6890$ vertices. SMPL only captures the undressed shape of the body and ignores the clothing and hair details. For this, SMPL+D adds a set of 3D offsets $\mathbf{D} \in \mathbf{R}^{V\times 3}$ to SMPL to capture the additional geometric details, *i.e.*, $M(\theta, \beta, \mathbf{D}) \in \mathbb{R}^{V\times 3}$ can also model clothed humans. We refer the readers to [36, 4] for a detailed description of SMPL+D.

For fitting SMPL+D to training images, we first estimate the parameters of SMPL using an off-the-shelf method SMPLify3D [29]. Since the person in the video remains the same, we optimize a single $\beta$ for the entire video. We then fix the pose $\{\theta\}_{f=1}^{F}$ and shape $\beta$ parameters and optimize for the offsets $\mathbf{D}$ using the following objective:

$$\mathbf{D} = \underset{\mathbf{D}'}{\operatorname{argmin}} \sum_{f=1}^{F} \mathcal{L}(M(\theta_f, \beta, \mathbf{D}')) = \mathcal{L}_{\text{Sil}} + \mathcal{L}_{\text{smooth}}. \quad (1)$$

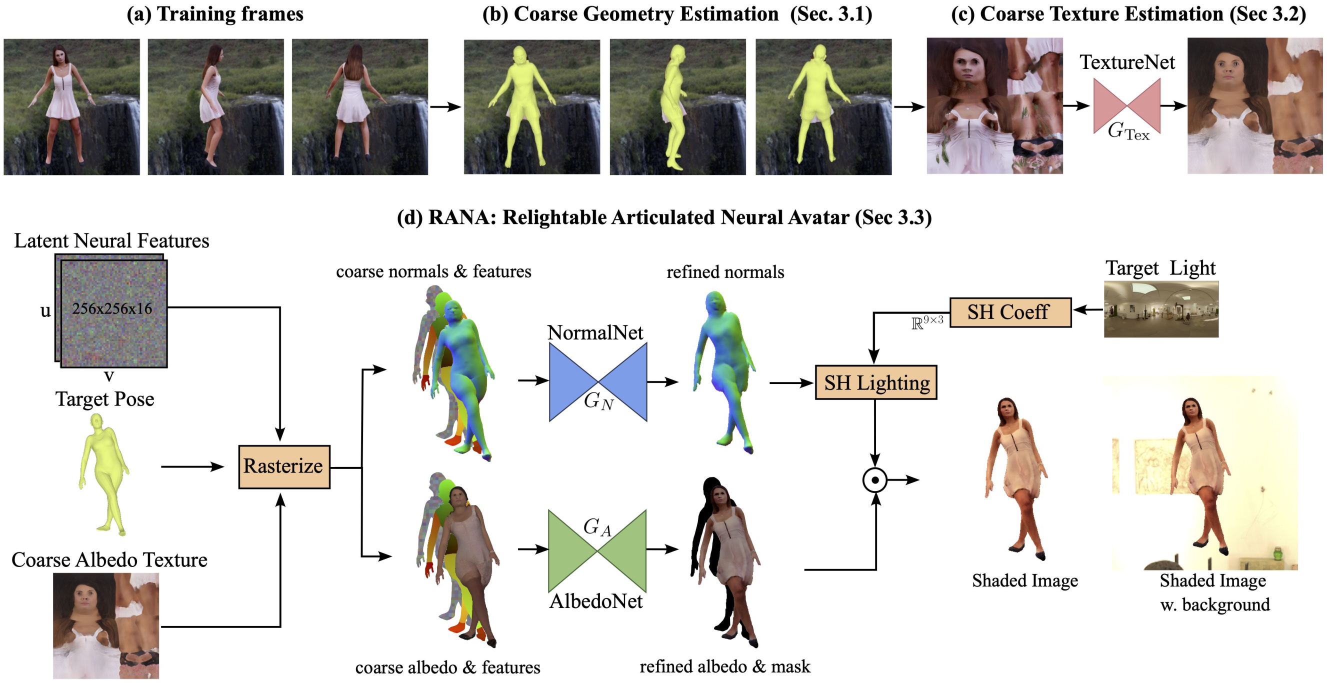


Figure 1. Overview of the proposed approach. (a) shows some training frames. (b) We estimate the coarse geometry of the person using the SMPL+D body model. (c) The SMPL+D fits are used to extract a UV texture map, which we process using TextureNet to obtain a coarse albedo texture map. (d) Given a target body pose, we rasterize person-specific neural features, coarse albedo, and coarse normals from SMPL+D to the target body pose and pass them to NormalNet and AlbedoNet to obtain refined normal and albedo maps, respectively. We then use the normal map and spherical harmonics lighting to obtain the shading image, which is multiplied with refined albedo to produce the shaded image. AlbedoNet also generates a binary mask, which we use to overlay the shaded image onto the background.

Here the term $\mathcal{L}_{\text{Sil}}$ is the silhouette loss. We obtain the silhouette of SMPL+D vertices $S_f$ for frame $f$ using a differentiable renderer [39] while the target silhouette $\hat{S}_t$ is obtained using a person segmentation model [15]. We define the $\mathcal{L}_{\text{Sil}}$ loss as

$$\mathcal{L}_{\text{Sil}} = \frac{1}{F}\sum_{f=1}^{F}|S_f - \hat{S}_f|. \tag{2}$$

The term $\mathcal{L}_{\text{smooth}}$ is a laplacian smoothing term to encourage the smooth surface of the mesh:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{F}\sum_{f=1}^{F}||L\mathbf{M}_f||, \tag{3}$$

where $L$ is the mesh Laplacian operator Note that we optimize a single set of $\mathbf{D}$ for the entire video, hence it does not model any pose-dependent geometric deformations. Some examples of SMPL+D can be seen in Fig. 1b.

### 3.2. Coarse Albedo Estimation

Given the SMPL+D fits for the training frames, we estimate an albedo texture map $T_A$ of the person in the UV space of SMPL. We follow [5] and first extract a partial texture map for each frame by back-projecting the image colors of all visible vertices to the UV space. The final texture map $T_I$ is then generated by calculating the median color value of most orthogonal texels from all frames. Depending on the available body poses in the training video, the obtained texture map can be noisy, and still have missing regions, *e.g.*, hand regions are often very noisy as no hand tracking is performed during SMPL fitting. Also, to ensure plausible relighting, the unknown shading from the texture map has to be removed, which is a challenging problem since decomposing shading and albedo is an ill-posed problem.

To address these problems, we propose TextureNet, $G_{\text{Tex}}$ (Fig. 1c), which takes a noisy texture map $T_I$ with unknown lighting as input and produces a clean albedo texture map as output, *i.e.*, $T_A = G_{\text{Tex}}(T_I)$ [36, 28]. One main challenge for training such a model is the availability of training pairs of noisy/shaded and albedo texture maps. We generate these pairs using 400 rigged characters from the RenderPeople dataset [2]. Since each character in RenderPeople has different UV coordinates, we follow [36] and register the characters with SMPL to obtain ground-truth UV maps in consistent SMPL UV coordinates. For noisy pairs, we generate images with random poses and lighting and extract texture maps like any other video mentioned above. We train $G_{\text{Tex}}$ using the following losses:

$$\mathcal{L}_{\text{Tex}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{VGG}} + \mathcal{L}_{\text{GAN}}. \tag{4}$$

Here $\mathcal{L}_{\text{pixel}}$ is the $L_1$ loss between the predicted and ground-truth albedo texture maps, $\mathcal{L}_{\text{VGG}}$ is $L_1$ difference between their VGG features, and $\mathcal{L}_{\text{GAN}}$ is the GAN loss [21, 44]. We use a vanilla U-Net architecture for TextureNet. It takes a noisy/shaded UV texture map with a resolution of 512×512 as input and produces the albedo texture with the same resolution as output. We also concatenate a 2D tensor of UV coordinates with the input texture map to provide part-specific information to the model. We train the network using Adam optimizer with a batch size of 8 and a learning rate of $1e^{-4}$ with cosine annealing and a minimum learning rate of $1e^{-5}$. To avoid overfitting during training, we perform random noise augmentation to the input texture maps including coarse dropout, gaussian noise, random brightness, and MixUp ($\beta = 0.4$) [67]. Some examples of estimated albedo maps can be seen in Fig. 2.

### 3.3. Relightable Articulated Neural Avatar

The coarse albedo texture and geometry obtained so far lack photo-realism and fine-grained details of the person. First, the topology of SMPL+D is fixed and cannot fully capture the fine geometric details, for example, loose clothing or long hairs. Second, the TextureNet can confuse the texture of the person with shading and may remove some texture details while estimating the albedo texture map. In this section, we present RANA which utilizes the coarse geometry and albedo map and generates higher quality images of the person. We parametrize RANA using the SMPL body model and SH lighting [5]. Specifically, RANA takes the target body pose $\theta$ and the target lighting in the form of second-order spherical harmonics coefficients $E \in \mathbb{R}^{9\times3}$ as input and synthesizes the target image $I^{\theta,E}$ as:

$$I^{\theta,E} = \texttt{RANA}(\theta, E, K), \tag{5}$$

where $K$ corresponds to the intrinsic parameters of the target camera viewpoint. One main challenge in learning such a neural avatar from a short RGB video is to maintain the disentanglement of geometry, albedo, and lighting, as any learnable parameters can overfit the training frames disregarding plausible disentanglement. Hence, we design RANA such that a plausible disentanglement is encouraged during training. Specifically, RANA consists of two convolutional neural networks NormalNet, $G_{\text{N}}$, and AlbedoNet, $G_{\text{A}}$, each responsible for generating the normal map $I_N^\theta \in \mathbb{R}^{h\times w\times3}$, and albedo map, $I_A^\theta \in \mathbb{R}^{h\times w\times3}$, of the person in the body pose $\theta$, respectively. It also consists of a set of subject-specific latent neural features $Z \in \mathbb{R}^{256\times256\times16}$ in UV coordinates to augment the details available in the coarse albedo map and geometry.

More specifically, given the target body pose $\theta$, we first generate the SMPL+D mesh $\mathbf{M}^\theta = M(\theta, \beta, \mathbf{D})$, where the shape parameters $\beta$ and clothing offsets $\mathbf{D}$ are the ones obtained in Sec. 3.1. We then use $\mathbf{M}^\theta$ to differentiably rasterize [39] the latent features $Z$ and coarse albedo texture $T_A$ to obtain a features image $I_Z^\theta$ and coarse albedo image $\bar{I}_A^\theta$ in the target body pose. We also rasterize a coarse normal image $\bar{I}_N^\theta$ and a UV image $I_{\text{uv}}^\theta$ using the normals and UV coordinates of $M^\theta$, respectively. The refined normal image $I_N^\theta$ and refined albedo image $I_A^\theta$ are then obtained as

$$I_N^\theta = G_N(I_Z^\theta, \bar{I}_N^\theta, \gamma(I_{\text{uv}}^\theta)), \tag{6}$$

$$I_A^\theta, S^\theta = G_A(I_Z^\theta, \bar{I}_A^\theta, \gamma(I_{\text{uv}}^\theta)), \tag{7}$$

where $S^\theta$ is the person mask in the target pose and $\gamma$ corresponds to the positional encoding of the UV coordinates [42]. Given the lighting $E$, we obtain the shading image $I_S^{\theta,E}$ using the normal map $I_N^\theta$ and SH lighting [49]. Under the usual assumptions of Lambertian material, distant lighting, and no cast shadows, the final shaded image $I^{\theta,E}$ is then obtained as

$$I^{\theta,E} = I_A^\theta \cdot I_S^{\theta,E}. \tag{8}$$

An overview of RANA can be seen in Fig. 1d. Since the lighting environment of the training video is unknown, we also optimize the second order SH coefficients $E \in \mathbb{R}^{9\times3}$ [49] of the training video during training. Under the assumption of Lambertian surface, none of the learnable parameters in RANA depend on the lighting information. Hence, if the disentanglement of normals, albedo, and lighting during training is correct, we can simply replace $E$ during inference with any other novel lighting environment to obtain relit images. We train RANA with the following objective:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{face}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{VGG}} + \mathcal{L}_{\text{GAN}} \quad &(9) \\ + \mathcal{L}_{\text{reg}}^{\text{albedo}} + \mathcal{L}_{\text{reg}}^{\text{normal}}. \quad &(10) \end{aligned}$$

Here $\mathcal{L}_{\text{pixel}}$ is the $L_1$ difference between the generated image $I^{\theta,E}$ and the ground-truth training frame, $\mathcal{L}_{\text{face}}$ is the $L_1$ difference between their face regions to assign a higher weight to face, and $\mathcal{L}_{\text{mask}}$ is the binary-cross-entropy loss between the estimated mask $S^\theta$ using $G_A$ and the pseudo-ground-truth mask obtained using a person segmentation model [15]. The term $\mathcal{L}_{\text{VGG}}$ is the $L_1$ difference between the VGG features of generated and ground-truth images, and $\mathcal{L}_{\text{GAN}}$ is the commonly used GAN loss [44, 21]. The term $\mathcal{L}_{\text{reg}}^{\text{albedo}}$ is the albedo regularization term that prevents the light information from leaking into the albedo image:

$$\mathcal{L}_{\text{reg}}^{\text{albedo}} = ||\sigma(I_A^\theta, k) - \sigma(\bar{I}_A^\theta, k)||^2. \tag{11}$$

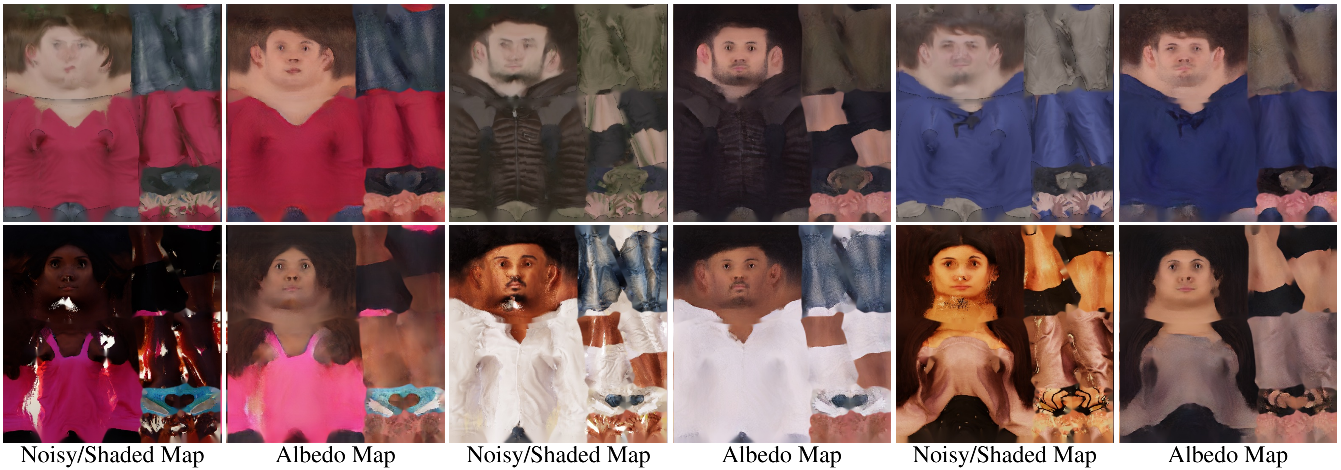


Figure 2. Examples of estimated albedo maps from noisy/shaded maps using our proposed TextureNet (Sec. 3.2). The first row shows some examples from the PeopleSnapshot dataset, while the second row shows examples from our proposed RelightingHuman dataset. s

Here $I_A^\theta$ is the albedo image obtained using $G_A$, $\bar{I}_A^\theta$ is the coarse albedo image, and $\sigma$ is the Gaussian smoothing operator with a kernel size $k$=51. $\mathcal{L}_{\text{reg}}^{\text{albedo}}$ encourages the overall color information in $I_A^\theta$ to be close to $\bar{I}_A^\theta$ while disregarding the texture information. Similarly, $\mathcal{L}_{\text{reg}}^{\text{normal}}$ is the normal regularization loss which prevents the normal image $I_N^\theta$ to move very far from the coarse normal image $\bar{I}_N^\theta$:

$$\mathcal{L}_{\text{reg}}^{\text{normal}} = |S_{\text{smpl}}^\theta I_N^\theta - S_{\text{smpl}}^\theta \bar{I}_N^\theta|, \tag{12}$$

where $S_{\text{smpl}}^\theta$ is the rasterized mask of SMPL+D mesh. It ensures that the regularization is applied only on the pixels where SMPL+D normals are valid. Note that no ground-truth supervision is provided to $G_A$ and $G_N$. They are mostly learned via the image reconstruction losses, while the disentanglement of normals and albedo is ensured via the novel design of RANA and regularization losses. For $\mathcal{L}_{\text{face}}$, we project the nose keypoint of SMPL body model on to the image and crop a $100\times100$ patch around the face to compute the loss. All other losses are calculated on the $512\times512$ generated images.

### 3.3.1 Pre-training using synthetic data

While we design RANA such that it generalizes well to novel body poses, the networks $G_A$ and $G_N$ may still overfit to the body poses available in the training video, in particular, when the coarse geometry and albedo are noisy. A significant advantage of RANA is that it can be trained simultaneously for multiple subjects, *i.e.*, we use different neural features $Z$ for each subject while sharing the networks $G_A$ and $G_N$. This not only allows the model to see diverse body poses during pre-training but also helps in learning to disentangle normals and albedo. Hence, we propose to pretrain $G_A$ and $G_N$ on synthetic data. For this, we use 400 rigged characters from the RenderPeople dataset. We generated 150 albedo and normal images for each subject under random body poses and pretrain both networks using ground-truth albedo and normal images. We use the $L_1$ loss for both terms.

### 3.3.2 Personalization.

After pre-training, given the RGB video of a novel subject, we optimize the latent features $Z$ and lighting environment $E$ of the video from scratch and only fine-tune AlbedoNet ($G_A$). During our experiments, we found that fine-tuning $G_N$ is not required if the model is pretrained (see Sec. 4.3). We keep $G_A$ fixed for the first 1000 iterations and only optimize $Z$ and $E$. This allows optimization of the latent features $Z$ to be compatible with the pretrained $G_A$ and $G_N$. We then optimize $G_A$, $Z$, and $E$ jointly for a total of 15k iterations. Note that, while we do pre-training on synthetic data, RANA does not suffer from domain gap because the neural texture is learned from scratch and AlbedoNet is fine-tuned for every video. Similar to TextureNet, we use vanilla U-Net for AlbedoNet and NormalNet. For training, we use Adam optimizer with a batch size of 16 and learning rate of $1e^{-4}$ with cosine annealing and minimum learning rate of $1e^{-5}$.

## 4. Experiments

In this section, we evaluate the performance of RANA using two different datasets. We perform an ablation study to validate our design choices and also compare our method with state-of-the-art and other baselines.

### 4.1. Datasets

**Relighting Human Dataset.** We propose a new photorealistic synthetic dataset to quantitatively evaluate the performance of our method. We use 49 rigged characters from the

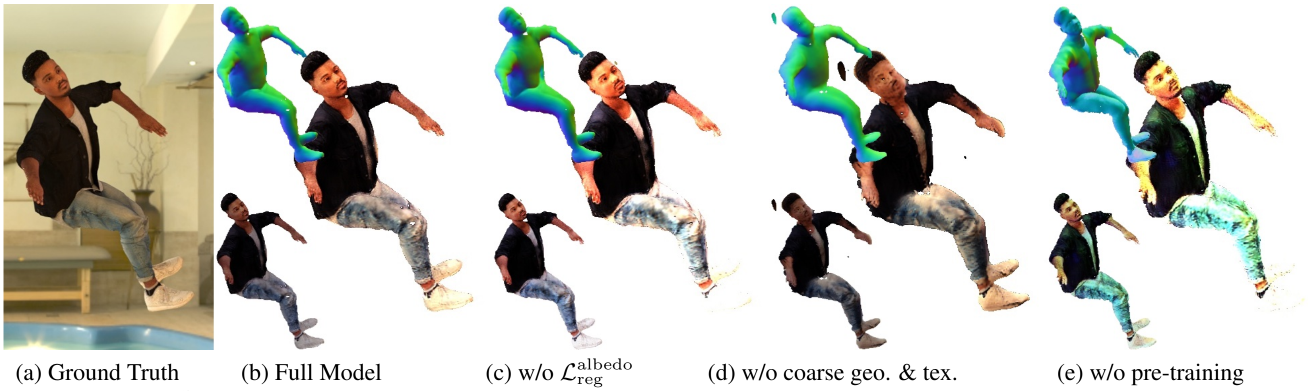

(a) Ground Truth (b) Full Model (c) w/o $\mathcal{L}_{\text{reg}}^{\text{albedo}}$ (d) w/o coarse geo. & tex. (e) w/o pre-training

Figure 3. **Ablation Study**. Impact of the different components of the proposed approach. Our full model yields the best results. Without the $\mathcal{L}_{\text{reg}}^{\text{albedo}}$ loss, the light information leaks into the albedo texture resulting in incorrect illumination. If we do not use coarse geometry and albedo texture, the resulting model does not generalize well to novel body poses. Similarly, training the model from scratch, without any pretraining on synthetic data, can result in an incorrect disentanglement of texture and geometry.

| Method | Image | | | | Normal Map | Albedo Map | | | |
|---|---|---|---|---|---|---|---|---|---|
| | LPIPS ↓ | DISTS ↓ | SSIM ↑ | PSNR ↑ | Degree° ↓ | LPIPS ↓ | DISTS ↓ | SSIM ↑ | PSNR ↑ |
| Full model | **0.217** | **0.204** | **0.751** | **19.498** | 64.350 | **0.219** | **0.207** | **0.779** | **21.832** |
| w/o $\mathcal{L}_{\text{albedo}}$ | 0.249 | 0.241 | 0.697 | 15.199 | 64.064 | 0.264 | 0.257 | 0.713 | 15.688 |
| w/o coarse geo. & tex. | 0.242 | 0.222 | 0.730 | 18.580 | 65.887 | 0.260 | 0.232 | 0.751 | 20.778 |
| w/o pre-training | 0.301 | 0.293 | 0.669 | 13.798 | 74.215 | 0.327 | 0.307 | 0.696 | 15.632 |
| fine-tune $G_N$ | 0.219 | 0.205 | 0.746 | 19.198 | **64.226** | 0.221 | 0.210 | 0.778 | 21.577 |

Table 2. Ablation study. We evaluate the impact of different components of the proposed method. See Fig. 3 for a qualitative comparison.

RenderPeople dataset [2] to generate photo-realistic images for each subject. We use HDRI maps from PolyHaven [1] to illuminate the characters and use the CMU motion capture dataset [3] to pose the characters. In contrast to our proposed method that uses image-based lighting, we use full Path Tracing to generate the dataset. Hence, it is the closest setting to an in-the-wild video, and any future work that uses a more sophisticated lighting model can be evaluated on this dataset. For a fair evaluation, we ensure that none of the characters is used during the training in Sec 3.2 and Sec 3.3. All testing images come with a ground-truth albedo map, a normal map, a segmentation mask, and light information. For our experiments, we evaluate on all 49 characters and train a separate RANA model for each subject. We develop three different protocols for evaluation:

**a) Novel Pose and Light Synthesis** This protocol evaluates the quality in terms of novel pose and light synthesis. We generate 100 training images for each subject rotating 360° with A-pose in front of the camera with fixed lighting. It mimics the self-rotating videos commonly used for training. For testing, we generate 150 frames for each subject with random body pose and light in each frame.

**b) Arbitrary Pose Training** To represent in-the-wild scenarios, we generate 100 training images for each subject with arbitrary poses in fixed lighting. For testing, we generate 50 frames for each subject with random unseen body pose and random light in each frame.

**c) Novel Light Synthesis.** This protocol evaluates the relighting ability of the methods. We generate 150 frames for train and test sets. The train set is generated with fixed lighting and random body poses. The body poses in the test set are exactly the same as the train sets, but each frame is generated using a different light source.

**People Snapshot Dataset. [5].** This dataset consists of real videos of characters rotating in front of the camera. We use this dataset for qualitative evaluation.

### 4.2. Metrics

We report several metrics to evaluate the quality of synthesized images as well as the disentanglement of normal and albedo images. For synthesized images and albedo maps, we use Learned Perceptual Patch Similarity (LPIPS ↓) [68], Deep Image Structure and Texture Similarity (DISTS ↓) [19], Structural Similarity Index (SSIM ↑) [60] and Peak Signal-to-Noise Ratio (PSNR ↑). For normals, we compute the error in degrees (°).

### 4.3. Ablation study

We evaluate different design choices of RANA in Tab. 2 and Fig 3. We use protocol-a of the Relighting Human dataset for all experiments. We first report the results of the final model which includes all loss terms in (10)

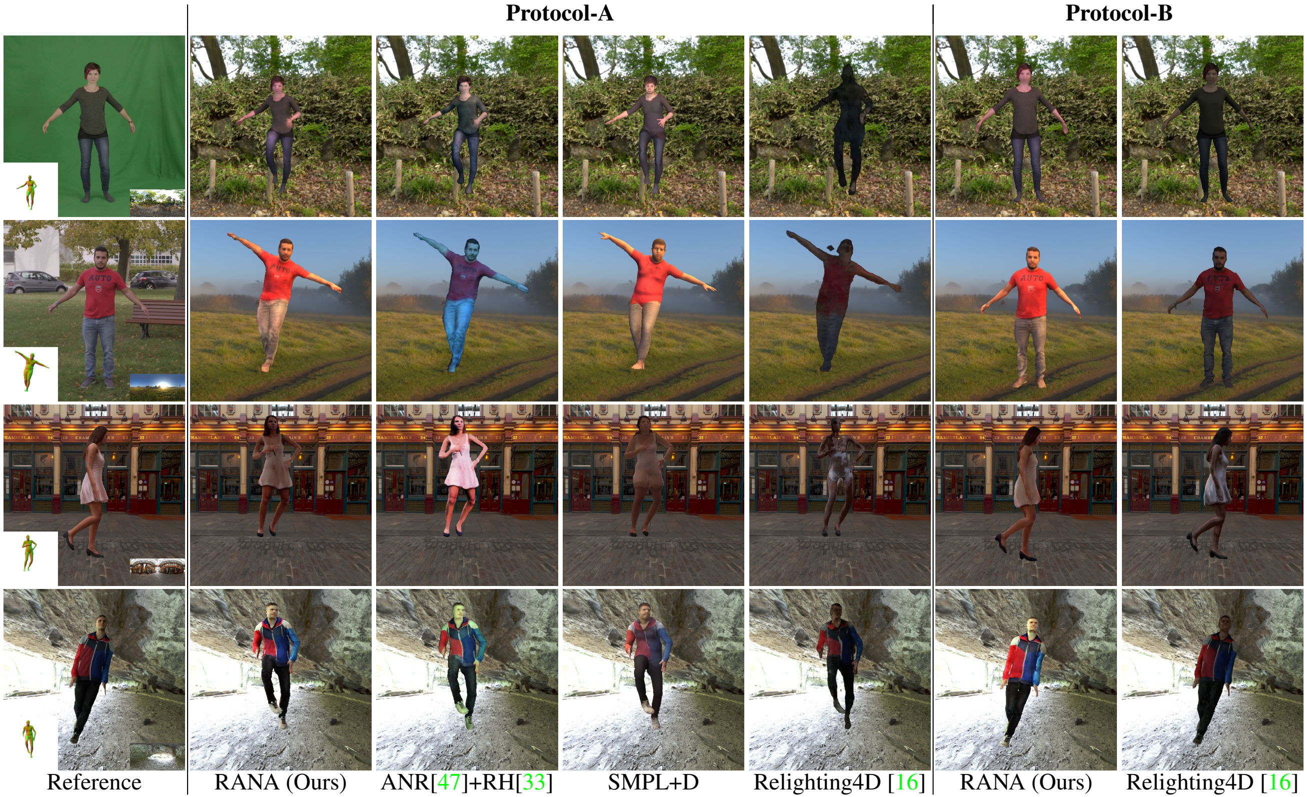


Figure 4. Comparison with the baselines and state-of-the-art methods. Column 1 shows a reference frame with the target body pose and lighting in the insets. In the absence of true reference images, for the Snapshot dataset (rows 1-2), we show training frames for reference. Columns 2-5 compare different methods for protocol-a, while columns 6-7 provide a comparison for protocol-b.

| Method | Image |  |  |  | Normal Map | Albedo Map |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | LPIPS ↓ | DISTS ↓ | SSIM ↑ | PSNR ↑ | Degree° ↓ | LPIPS ↓ | DISTS ↓ | SSIM ↑ | PSNR ↑ |
| **Protocol (a): Novel Pose and Light Synthesis** |  |  |  |  |  |  |  |  |  |
| ANR [47] + RH [33] | 0.275 | 0.416 | 0.664 | 17.495 | N.A. | 0.266 | 0.429 | 0.656 | 14.804 |
| Relightable SMPL+D | 0.265 | 0.225 | 0.751 | **19.678** | **64.121** | **0.216** | **0.182** | **0.811** | **22.623** |
| RANA (Ours) | **0.217** | **0.204** | **0.751** | 19.498 | 64.350 | 0.219 | 0.207 | 0.779 | 21.832 |
| **Protocol (b): Arbitrary Pose Training** |  |  |  |  |  |  |  |  |  |
| Relightable SMPL+D | 0.259 | 0.222 | 0.761 | 19.988 | 63.756 | **0.210** | **0.180** | 0.823 | **23.052** |
| RANA (Ours) | **0.193** | **0.183** | **0.793** | **20.693** | **63.409** | 0.226 | 0.194 | **0.825** | 22.935 |
| **Protocol (c): Novel Light Synthesis** |  |  |  |  |  |  |  |  |  |
| Relighting Humans [33] | 0.226 | 0.353 | 0.668 | 21.205 | - | - | - | - | - |
| Relighting4D [16] | 0.192 | 0.342 | 0.654 | 21.080 | 65.099 | 0.263 | 0.374 | 0.593 | 20.014 |
| RANA (Ours) | **0.173** | **0.171** | **0.842** | **22.338** | **62.823** | **0.200** | **0.179** | **0.865** | **24.721** |

Table 3. Comparison with the baselines and state-of-the-art methods. See Fig. 4 for qualitative comparison.

and pretraining using synthetic data (Sec. 3.3.1). The full model achieves an LPIPS score of 0.217 for image synthesis and 0.219 for albedo map reconstruction. If we remove the loss term $\mathcal{L}_{\text{reg}}^{\text{albedo}}$ from (10), the LPIPS scores for image and albedo map reconstruction increase to 0.249 and 0.264, respectively. Note that the error for albedo maps increases significantly while the error for normal maps remains roughly the same. This indicates that without $\mathcal{L}_{\text{reg}}^{\text{albedo}}$ the light information leaks into the albedo image. An example of this behavior can also be seen in Fig 3c (w/o $\mathcal{L}_{\text{reg}}^{\text{albedo}}$).

Next, we evaluate the impact of coarse geometry and albedo texture on RANA. If we do not use coarse geometry and albedo, LPIPS score increases to 0.242 as compared to 0.217 for the full model. The normal error also increases to 65.9° from 64.3°. This is also evident from the qualitative results shown in Fig 3d, indicating that coarse geometry and albedo help in improved image synthesis quality, in particular when the target body pose is far from the training poses. Next, we evaluate the impact of pretraining on synthetic data. Without the pretraining, all error metrics increase significantly. Specifically, the LPIPS score for image reconstruction increases from 0.217 to 0.301, while the normal error increases from 64.3° to 74.2°. If we look at Fig 3e, we can see that shading information leaks into both the normals and albedo maps. Hence, pretraining the networks also help with the plausible disentanglement of geometry, texture, and light. Thanks to the design of RANA, we can pretrain on as many subjects as available, which is not possible with most of the state-of-the-art methods for human synthesis [45, 46, 58, 31]. Finally, as discussed in Sec 3.3.1, we keep the network $G_N$ fixed during finetuning if RANA is pretrained on synthetic data. In the last row of Tab. 2, we evaluate the case when $G_N$ is also fine-tuned. We can see that it has a negligible impact on the results.

### 4.4. Comparison with other methods

Since ours is the first work that simultaneously evaluates novel light and pose synthesis, we ourselves build some baselines as follows:

**Relightable SMPL+D:** We rasterize the SMPL+D mesh normals and albedo texture estimated using TextureNet in the target body pose and use SH lighting to generate the shaded images.
**ANR [47]+RH[33]:** We train an ANR [47] model which synthesizes images in the lighting of the training video. We then pass the generated images to the single-image human relighting method Relighting Humans (RH) [33] to obtain the relighted images for the target light. We use the publicly available source code and models of [33].
**Relighting4D [16]** is a state-of-the-art human video relighting method. We use the publicly available source code and train it on our dataset.

The results are summarized in Tab. 3 and Fig 4. We do not report results of Relighting4D [16] for protocol-a since it cannot handle novel body poses as can be seen in Fig 4 (column-5). For protocol-a, our method clearly outperforms other baselines for final image synthesis results. Surprisingly, the SMPL+D baseline yields better numbers for albedo reconstruction, even though it provides overly smooth albedo textures. This is because TextureNet is trained on synthetic data, hence albedo maps in SMPL+D baseline achieve the best result in this synthetic data benchmark. However, the SMPL+D baseline does not produce the best image quality. RANA, on the other hand, achieves significantly better image quality while staying close to the unlit texture in SMPL+D.

For protocol-b, we only compare the best two methods from protocol-a as we found it unnecessary to retrain ANR based on the results of protocol-a. As we can see in Tab. 3, the results are consistent with protocol-a.

For protocol-c, we compare with a video-based relighting method Relighting4D [16] and the single-image-based relighting method Relighting Humans [33]. RANA outperforms both methods across all metrics. Some qualitative comparisons with Relighting4D [16] can be seen in Fig. 4 (columns 6-7), where RANA clearly yields better image relighting results. Note that each model for Religthing4D [16] requires 260k iterations for training whereas RANA models are trained only for 15k iterations, thanks to our novel design that allows pre-training on synthetic data, allowing quick fine-tuning for new subjects. In contrast, Relighting4D [16] by design cannot be pretrained easily on multiple subjects. On the other hand, the method [33] is trained for up-right frontal poses only and cannot handle arbitrary body poses, resulting in inferior relighting performance. This is also the main reason behind the inferior performance of the ANR [47]+RH [33] baseline for protocol-a.

Finally, we provide additional qualitative results of avatars animated with complex motions in the supp. video.

## 5. Conclusion and Future Work

We presented RANA which is a novel framework for learning relightable and articulated neural avatars of humans. We demonstrated that RANA can model humans from unconstrained RGB videos while also disentangling their geometry, albedo texture, and environmental lighting. We showed that it can generate high-quality images of people under any novel body pose, viewpoint, and lighting. RANA can be trained simultaneously for multiple people and we showed that pre-training it on multiple (400) synthetic characters significantly improves the image synthesis quality. We also proposed a new photo-realistic synthetic dataset (Relighting Human) to quantitatively evaluate the performance of our proposed method, and believe that it will prove to be very useful to further the research in this direction.

The most pressing limitation of RANA is the assumption of Lambertian surface, no cast shadows, and image-based lighting. In the future, we hope to incorporate more sophisticated physically-based rendering in our framework which will hopefully result in better image quality, self-shadowing and normal maps with more details. Moreover, RANA does not explicitly model motion-dependent clothing deformations. Modeling and neural rendering clothing deformations from a short video clip would be interesting future work.

## References

[1] Poly Haven, 2020. https://hdrihaven.com. 7

[2] Render People, 2020. https://renderpeople.com/3d-people. 4, 7

[3] Carnegie Mellon University Graphics Lab: Motion Capture Database. http://mocap.cs.cmu.edu, 2014. 7

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 2, 3

[5] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, June 2018. 2, 4, 5, 7

[6] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 2022. 2

[7] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 3

[8] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. In *TPAMI*, 2020. 2

[9] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *ICCV*, 2015. 2

[10] Akin Caliskan, Armin Mustafa, and Adrian Hilton. Temporal consistency loss for high resolution textured and clothed 3d human reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1790, 2021. 2

[11] Akin Caliskan, Armin Mustafa, Evren Imre, and Adrian Hilton. Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

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

[13] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 3

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

[15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *CVPR*, 2017. 4, 5

[16] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *ECCV*, 2022. 2, 3, 8, 9

[17] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *CVPR*, 2023. 2, 3

[18] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Annual conference on Computer Graphics and Interactive Techniques*, 2000. 3

[19] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *TPAMI*, 2022. 7

[20] Patrick Esser, Ekaterina Sutter, , and Bjorn Ommer. A variational U-Net for conditional appearance and shape generation. In *CVPR*, 2018. 3

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 5

[22] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *CVPR*, 2021. 3

[23] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: volumetric performance capture of humans with realistic relighting. In *TOG*, 2019. 2, 3

[24] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time Deep Dynamic Characters. *SIGGRAPH*, 2021. 3

[25] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 2

[26] Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *CVPR*, 2021. 3

[27] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 2

[28] Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. Bodymap: Learning full-body dense correspondence map. In *CVPR*, 2022. 4

[29] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *3DV*, 2021. 3

[30] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *ECCV*, 2022. 3

[31] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2, 3, 9

[32] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 3

[33] Yoshihiro Kanamori and Yuki Endo. Relighting humans: Occlusion-aware inverse rendering for full-body human images. In *SIGGRAPH*, 2018. 3, 8, 9

[34] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *NeurIPS*, 2021. 3

[35] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. Single-image full-body human relighting. In *Eurographics Symposium on Rendering (EGSR)*, 2021. 3

[36] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 2, 3, 4

[37] Guannan Li, Chenglei Wu, Carsten Stoll, Yebin Liu, Kiran Varanasi, Qionghai Dai, and Christian Theobalt. Capturing relightable human performances under general uncontrolled illumination. In *Computer Graphics Forum*. Wiley Online Library, 2013. 3

[38] 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. *SIGGRAPH ASIA*, 2021. 3

[39] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019. 4, 5

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH ASIA*, 2015. 2, 3

[41] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. Deep relightable textures - volumetric performance capture with neural rendering. *SIGGRAPH*, 39(6), 2020. 2

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3, 5

[43] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *TOG*, 2021. 3

[44] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019. 5

[45] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2, 3, 9

[46] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 3, 9

[47] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. ANR-articulated neural rendering for virtual avatars. In *CVPR*, 2021. 2, 3, 8, 9

[48] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. In *CVPR*, 2021. 3

[49] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, 2001. 2, 3, 5

[50] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

[51] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 3

[52] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured Neural Avatars. In *CVPR*, 2019. 3

[53] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 2007. 2

[54] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 2

[55] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. In *SIGGRAPH*, 2019. 3

[56] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. *Computer Graphics Forum*, 2021. 3

[57] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. *Computer Graphics Forum*, 41(2):703–735, 2022. 3

[58] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 2, 3, 9

[59] Tuanfeng Y Wang, Duygu Ceylan, Krishna Kumar Singh, and Niloy J Mitra. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis. In *3DV*, 2021. 3

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 2004. 7

[61] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. In *SIGGRAPH*, 2020. 3

[62] 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 *CVPR*, 2022. 3

[63] Lingbo Yang, Pan Wang, Chang Liu, Zhanning Gao, Peiran Ren, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Xiansheng Hua, and Wen Gao. Towards fine-grained human pose transfer with detail replenishing network. In *TIP*, 2021. 3

[64] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 2

[65] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *SIGGRAPH ASIA*, 2022. 3

[66] Jae Shin Yoon, Duygu Ceylan, Tuanfeng Y. Wang, Jingwan Lu, Jimei Yang, Zhixin Shu, and Hyun Soo Park. Learning motion-dependent appearance for high-fidelity rendering of dynamic humans from a single camera. In *CVPR*, 2022. 2, 3

[67] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[69] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *SIGGRAPH*, 2021. 3

[70] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single rgb camera. In *CVPR*, 2022. 3

[71] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *ECCV*, 2020. 2, 3

[72] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *ICCV*, 2019. 3