TotalSelfScan: Learning Full-body Avatars from Self-Portrait Videos of Faces, Hands, and Bodies

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances in neural implicit functions make it possible to reconstruct a human model from a monocular self-rotation human video. While they present impressive results of the human body, the quality of reconstructed face and hands are relatively low. The main reason is the image region occupied by these parts is very small compared to the body. To solve this problem, we propose TotalSelfScan, which reconstructs the full-body human from several monocular self-rotation videos that focus on the face, hands, and body, respectively. Compared to recording a single body video, this setting has almost no additional cost but provides abundant details of essential parts. To learn the full-body model, instead of encoding the whole body in a single-part network, we propose a novel multi-part representation to model separate parts and then fuse the part-specific observations into the unified human model. Once learned, the human model enables rendering photorealistic free-viewpoint videos under novel human poses. Experiments show that TotalSelfScan can significantly improve the performance on the face and hands compared to the existing methods.

1 Introduction

Human reconstruction and rendering can be ubiquitously applicable in promising areas such as immersive viewing experiences and telepresence. In most applications, hand motion and high-quality face are essential components for social communications. While previous methods [18, 6, 43] demonstrate the impressive performance of full-body reconstruction, they usually require hundreds of calibrated and synchronized cameras, which makes them impossible to create personalized avatars for general users.

To make human avatar creation a commodity, many recent methods propose to reconstruct the human model from monocular RGB inputs. Based on the pixel-aligned implicit function, some works [36, 37, 51] propose to reconstruct the human geometry and appearance from a single image. However, relying on the paired 3D data and images for training, their methods have difficulty in generalizing to the in-the-wild images. In addition, several works propose to reconstruct a person-specific model from a monocular video that records a self-rotation performer holding a fixed pose. The subjects can scan themselves only using a fixed camera without any assistance, which makes the personalized avatar creation possible. VideoAvatar [3] relies on the statistical human model and further adds displacements to the vertices of the human mesh for modeling clothing. More recently, SelfRecon [17] proposes to combine the explicit mesh and implicit signed distance field to improve the reconstruction. While these self-rotation video based methods achieve great results for the human body, the quality of hands and face are relatively low. The main reason is that limited to the resolution of input images that contain the whole body, the image region occupied by the hands and face is very small.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

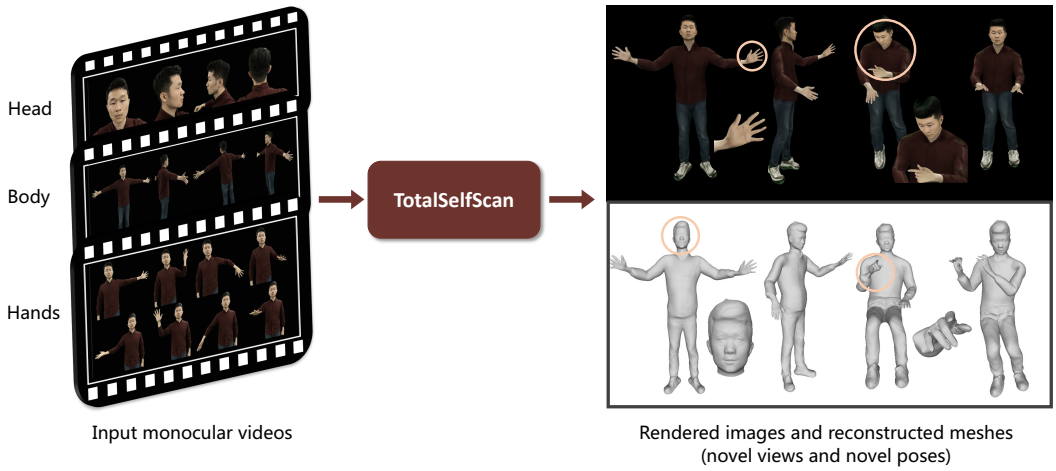

Figure 1: Given several monocular part-specific videos of a performer, our method is able to reconstruct an animatable full-body digital double.

In this paper, following the self-rotation video setting, we propose to use a fixed camera to record several monocular videos that focus on the human body, head, and hands respectively, and the performer rotates the part of concern in each video, as shown in Figure 1. Compared to recording a single body video, this setting has almost no additional cost but provides abundant details of essential parts. Given these monocular videos as input, our target is to recover the detailed full-body geometry and appearance, which enables rendering photorealistic free-viewpoint videos under novel human poses. However, this new problem brings many challenges. First, the human parts own very different scales, such as the hands and the body, which requests the human model to preserve the details at various scales. Second, each video is part-centric and it is unclear how to combine these videos into a unified human model. Third, the appearance of the same part among videos may be not the same due to different lighting. Lastly, the self-rotation videos only cover very limited view directions, which makes the rendering under both novel views and novel poses challenging.

To address these problems, we propose a novel multi-part representation to model the body, head, and hands respectively, and fuse the part-specific observations into a unified human model. Specifically, we utilize multiple signed distance fields (SDFs) to represent the geometry of different human parts in the canonical space. To learn the representation, we propose the part deformation fields to establish the correspondence between each observation space and the canonical space, and the SDF-based volume rendering is utilized to train the model. Then, we propose to composite the geometry and appearance of adjacent parts to obtain a consistent full-body model. Finally, we extend the ray transformation [49] to the non-rigid human to improve the quality of rendered images under novel human poses.

In summary, this work makes the following contributions:

- We introduce a new task of total avatar creation from monocular part-specific videos, which provides much more information on critical human parts almost for free.
- We propose a novel multi-part representation modeling separate parts and the part deformation fields and compositions are proposed to obtain the unified human model. At inference time, the non-rigid ray transformation is utilized to improve the quality of rendered images.
- We show that, compared to using a single body video, the joint analysis of multiple part-specific videos demonstrates significant performance improvement on the face and hands.

2 Related work

Human reconstruction. How to model the underlying geometry and appearance of humans for 3D construction has always been an open question, which can be roughly divided into explicit and implicit representations. The explicit representation, traditionally the mesh, is obtained in advance in

the form of statistical human models [24, 5, 29, 45] or pre-scanned personalized templates [12, 14]. Based on the statistical human model, some works add surface deformation to capture more details [54, 19, 39, 10, 52]. Moreover, a series of methods [47, 15, 46, 16] utilize personalized templates and deform them by dense non-rigid tracking to achieve impressive performance.

The implicit function based methods are prevailing recently. PiFu based methods [36, 37, 51] learn a pixel-aligned implicit function efficiently for both geometry and texture from a single image. However, high-quality 3D models are required for their training, which limits their generalization ability. In addition, optimizing a network to represent a person-specific model shows impressive results [13, 38, 9, 33, 31]. Here, we focus on the methods using images as input. Inspired by NeRF [25], Neural Body [33] optimizes the radiance field conditioned on the structured latent codes with only images as supervision. Neural Actor [22] integrates the texture features to enhance the rendering performance. MVP [23] proposes a mixture of volumetric primitives that support efficient rendering of avatars. Furthermore, in order to achieve better animatable effects, learning blend weights automatically from data [31] and incorporating articulated structures [28] are explored. To improve the geometry, [44, 32] represent the human geometry as the signed distance field and use the volume rendering to learn the representation from images. In addition, some works [3, 17] propose to reconstruct the human model from monocular self-rotation videos.

Total capture. Total capture aims to recover the whole body of the performer including body, face, and hands. Most existing works [24, 34, 20, 8, 7] consider these parts separately. To model the human as a whole, several approaches [18, 30] stitch the part-specific model together and further parameterize the unified model via additional registration and regression. Based on the parametric models, some works propose to optimize the parameters to fit the multi-view [18] or single-view [30, 42] image evidence. To achieve faster inference, some approaches [35, 11, 53, 26] directly regress the parameters of models with neural networks and then integrate or refine them with observations from local regions. Different from the typical mesh representation, imGHUM [4] proposes a generative human model represented by multiple signed distance functions and learns the model from point clouds. A recent work [6] reconstructs full-body avatars based on a disentangled latent space with variational autoencoders conditioned on driving signals and [43] additionally takes clothing modeling into consideration. They reveal remarkable performance but require hundreds of synchronized cameras and high-resolution images. In contrast, we reconstruct high-quality full-body avatars from monocular videos.

3 Method

This paper is to reconstruct the detailed full-body geometry and appearance from several monocular videos, which enables rendering photorealistic free-viewpoint videos under novel human poses. Figure 2 shows the overview of the proposed model. We represent the human as multi-part networks in the canonical space, modeling body, head, and hands, respectively (Section 3.1). To learn the representation from part-specific videos, we first transform the sample points from each observation space to the canonical space via part deformation fields (Section 3.2), and then combine each part into a full-body model (Section 3.3). We train the model with volume rendering (Section 3.4). After training, we propose the non-rigid ray transformation for novel pose rendering (Section 3.5).

Given four monocular part-specific videos (body, head, and two hands) of the performer, we first utilize the EasyMoCap [1] to estimate SMPL+H [34] parameters for the body and hands videos and utilize an adaptation of [40] to estimate the FLAME [20] parameters for the face video. Then, we adopt [21] to generate the human mask for each frame. In the following, we describe the algorithm details of each component.

3.1 Multi-part canonical human model

Similar to [44, 32], the human geometry and appearance are represented as signed distance fields F_s and color fields F_c using MLP networks. In contrast to previous methods that encode the full human body in a single model, we decompose the human into separate parts (i.e., body, head, and hands) and each part is represented as a single model. Specifically, for the part p, the models can be written as follows:

$$s(\mathbf{x}), \mathbf{z}(\mathbf{x}) = F_s^p(\mathbf{x}), \tag{1}$$

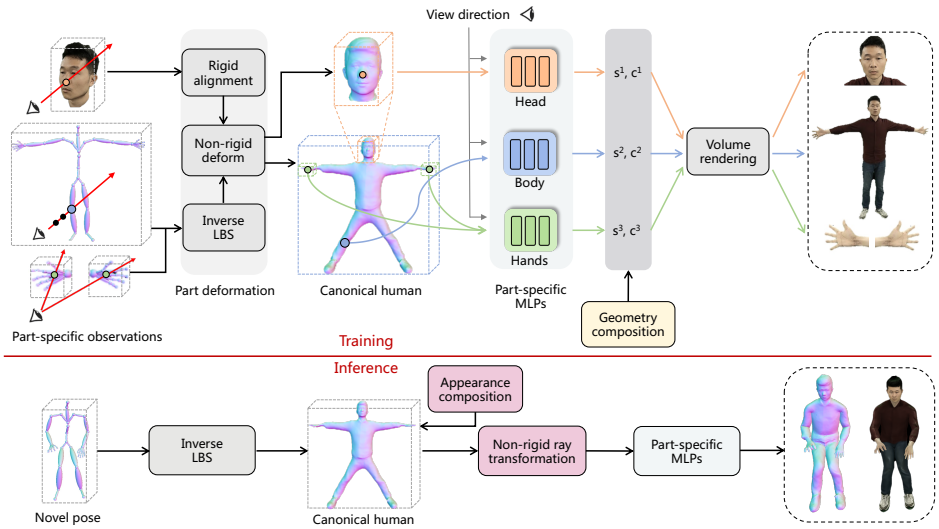

Figure 2: **Overview of the proposed approach.** Given sample points in each part-specific observation space, we transform them to the unified canonical space using part deformations. For each points, the corresponding part network is utilized to predict the signed distance and color and then we use the volume rendering to synthesize images. To obtain the consistent human geometry, we introduce the geometry composition loss in the training. At inference time, we first composite the appearance of adjacent parts and then animate the canonical human model using input novel poses. To render high-quality images, we adopt the non-rigid ray transformation to replace original view directions with global ones.

$$\mathbf{c}(\mathbf{x}) = F_{\mathbf{c}}^{p}(\mathbf{x}, \mathbf{z}(\mathbf{x}), \mathbf{n}(\mathbf{x}), \mathbf{d}, \boldsymbol{\ell}_i^p), \tag{2}$$

where $\mathbf{x}$ and $\mathbf{d}$ denote the point position and view direction in the canonical space, $\mathbf{z}(\mathbf{x})$ and $\mathbf{n}(\mathbf{x})$ denote the corresponding geometry feature and normal, and $\boldsymbol{\ell}_i^p$ denotes the latent code at video frame i. $s(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ denote the signed distance and color to be decoded at the position $\mathbf{x}$.

3.2 Part deformation

To learn the canonical human model, we need to establish the correspondences between each part-specific observation space and the canonical space. Due to the different properties of each part, we adopt part-specific deformation strategies.

Body and Hands. Since the body and hands own a similar articulated structure, we adopt the same deformation strategy consisting of skinning transformation and non-rigid transformation. In particular, for a point $\mathbf{x}$ in the observation space of part p at frame i, the corresponding canonical point $\mathbf{x}_c$ can be written as follows:

$$\mathbf{x}_c = T_{ilbs}(\mathbf{x}, \mathbf{p}_i) + T_{nr}^{p}(T_{ilbs}(\mathbf{x}, \mathbf{p}_i), \boldsymbol{\phi}_i^p), \tag{3}$$

where T_{ilbs} is the standard inverse linear blend skinning algorithm with no learnable parameters and $\mathbf{p}_i$ is the human pose. Note that the blending weights of $\mathbf{x}$ is generated by retrieving the counterpart of closet vertex on the template mesh. T_{nr}^{p} is the part-specific non-rigid displacement field implemented as an MLP network and $\boldsymbol{\phi}_i^p$ is the latent code. Practically, we adopt the SMPL+H model [34] for both body and hands.

Head. Different from the articulated structure of the above parts, the head is closer to a rigid model, whose deformation can be described as the rigid transformation. For more accurate 3D face reconstruction on face-centric images, we select the textured FLAME model rather than a unified human model such as SMPL-X [30] that only contains the geometry model. We adopt [40] for the FLAME-based 3D face reconstruction and then register the FLAME model to the canonical SMPL+H model by solving the following optimization problem:

$$\min_{\mathbf{R},\mathbf{T}} \|\mathbf{R}\mathbf{L}_{\mathbf{flame}} + \mathbf{T} - \mathbf{L}_{\mathbf{smplh}}\|_2, \tag{4}$$

where $\mathbf{R}$ and $\mathbf{T}$ are the rigid transformation, and $\mathbf{L}_{\mathbf{flame}}$ and $\mathbf{L}_{\mathbf{smplh}}$ are the pre-defined corresponding landmark points on FLAME model and SMPL+H model respectively.

However, the above optimization aligns the two parametric models rather than the true face surface. For more accurate head alignment, we propose to further register the detailed face surfaces obtained from separate reconstruction described in following sections. The optimization is as follows:

$$\min_{\mathbf{R},\mathbf{T}} \|(\mathbf{R}\mathbf{X}_{\mathbf{face}} + \mathbf{T} - \mathbf{X}_{\mathbf{fbody}}) \cdot \mathbf{n}\|_2, \tag{5}$$

where $\mathbf{X}_{\mathbf{face}}$ denotes the vertex of reconstructed face mesh from the face video, $\mathbf{X}_{\mathbf{fbody}}$ denotes the corresponding nearest vertex of the reconstructed body mesh from the body video, and $\mathbf{n}$ denotes the normal of the $\mathbf{X}_{\mathbf{fbody}}$. Note that we initialize the $\mathbf{R}$ and $\mathbf{T}$ using the result of Equation (4).

In addition to the rigid transformation, we also introduce the non-rigid transformation similar to the body and hands. Finally, the whole transformation can be written as follows:

$$\mathbf{x}_c = \mathbf{R}^*\mathbf{x} + \mathbf{T}^* + T^p_{nr}(\mathbf{R}^*\mathbf{x} + \mathbf{T}^*, \boldsymbol{\phi}^p_i), \tag{6}$$

where $\mathbf{R}^*$ and $\mathbf{T}^*$ are the solutions of Equation (5).

3.3 Part compositions

After warping points from each part-specific observation space to the canonical space, we need to composite separate part models into a unified human model, which contains the compositions of geometry and appearance. Specifically, we define a bounding box for each part in the canonical space and two adjacent bounding boxes contain some overlap. In the bounding box $\mathbf{B}^p$ of part p, the corresponding human model (i.e., F^p_s and F^p_c) is adopted. To generate a realistic unified model, we utilize the following composition strategies in the intersection region of two bounding boxes.

Geometry composition. To ensure smooth surface transitions between two adjacent part networks, we introduce the following loss function:

$$L_{\mathrm{g}} = \sum_{\mathbf{x}\in\mathcal{X}_{ij}} \|s^{p_i}(\mathbf{x}) - s^{p_j}(\mathbf{x})\|_2 + \sum_{\mathbf{x}\in\mathcal{S}_j} \|s^{p_i}(\mathbf{x})\|_2, \tag{7}$$

where $\mathcal{X}_{ij}$ is the sample points in the intersection region between part i and part j, and $\mathcal{S}_j$ is the sample points on the zero-level-set of part j in the intersection region. The first term encourages the two signed distance fields to be more consistent and the second term further emphasizes the surface consistency.

Appearance composition. Different from the consistent human geometry across part-specific videos, the appearances of the same part in different videos are usually inconsistent due to uneven and varying lighting conditions. As a result, the learned separate part models generate inconsistent appearances in the same position. To solve this problem, we select the body model as the reference and optimize the appearance latent code of other parts $\boldsymbol{\ell}^p$ to obtain a consistent appearance, which can be written as follows:

$$\min_{\boldsymbol{\ell}^p} \sum_{r\in\mathcal{R}} \|\tilde{\mathbf{C}}^{body}(\mathbf{r}, \boldsymbol{\ell}^{body}) - \tilde{\mathbf{C}}^p(\mathbf{r}, \boldsymbol{\ell}^p)\|_2, \tag{8}$$

where $\mathcal{R}$ denotes the set of rays of sampled intersection region, and $\tilde{\mathbf{C}}^p(\mathbf{r}, \boldsymbol{\ell}^p)$ denotes the rendered pixel color of part p. After the latent code optimization, the appearances between two parts become similar but not exactly the same, which leads to an unwanted seam near the boundary. To generate a realistic appearance transition, we further fuse the two adjacent color fields in the overlap region as follows:

$$\mathbf{c}(\mathbf{x}) = \mathbf{c}^{p_i}(\mathbf{x})(1 - \frac{d^{p_i}(\mathbf{x})}{d_s}) + \mathbf{c}^{p_j}(\mathbf{x})\frac{d^{p_i}(\mathbf{x})}{d_s}, \tag{9}$$

where $d^{p_i}(\mathbf{x})$ denotes the distance of $\mathbf{x}$ to part p_i and d_s denotes the length of overlap. Note that the appearance composition is performed after training.

3.4 Training

To learn the signed distance filed $s(\mathbf{x})$ and color filed $\mathbf{c}(\mathbf{x})$ from images, we leverage the SDF-based volume rendering [48, 41] to synthesize images and compare them with the observation images.

For high-quality reconstruction, we introduce a two-stage training strategy. In the first stage, we train each part model using the part-specific video separately and the loss functions are written as follows:

$$L^p = L^p_{\text{rgb}} + L^p_{\text{mask}} + \lambda_1 L^p_{\text{E}} + \lambda^p_2 L^p_{nr}, \tag{10}$$

$$L^p_{\text{rgb}} = \sum_{r\in\mathcal{R}} \|\tilde{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2, \;\; L^p_{\text{mask}} = \sum_{r\in\mathcal{R}} \text{BCE}(\text{sigmoid}(-\rho s^{\mathbf{r}}), M(\mathbf{r})), \tag{11}$$

$$L^p_{\text{E}} = \sum_{\mathbf{x}\in\mathcal{X}^c} (\|\nabla F^p_s(\mathbf{x})\|_2 - 1)^2, \;\; L^p_{nr} = \sum_{\mathbf{x}\in\mathcal{X}^c} \|T^p_{nr}(\mathbf{x}, \boldsymbol{\phi}^p_i)\|_2. \tag{12}$$

The first term L^p_{rgb} denotes the color loss. The second term L^p_{mask} denotes the mask loss, where ρ is a gradually increasing hyper parameter and $M(\mathbf{r})$ is the ground-truth mask value. The third term L^p_{E} denotes the Eikonal loss, where $\mathcal{X}^c$ is the sampled points in canonical space. The last term L^p_{nr} denotes the displacement regularization. The λ_1 and λ^p_2 are two constants and note that λ^p_2 is part-specific.

In the second stage, to obtain better unified human model, we further add geometry composition loss. Instead of training all part models jointly, we only optimize the body model while fixing the other part models. This strategy can fuse adjacent signed distance fields smoothly and preserve the details at the same time. The loss function can be written as follows:

$$L^{union} = L^{body} + L_g. \tag{13}$$

3.5 Non-rigid ray transformation

After training, the canonical human model can be animated and rendered under novel human poses. However, when the rays under novel human poses deviate from the training ray distribution, the color field network F_c produces unexpected artifacts. Inspired by [49] that constructs a ray atlas for a static object, we extend it to the dynamic human. Specifically, we first transform each ray of training frames from the observation space to the canonical space. Then, we extract the human mesh in the canonical space and save all view directions $\mathbf{d}^n_v$ of each vertex v on the human mesh. Based on the saved view directions, we can compute a global view direction $\overline{\mathbf{d}}_v$ as follows:

$$\overline{\mathbf{d}}_v = \frac{1}{N}\sum_n \mathbf{d}^n_v, \tag{14}$$

where N is the number of transformed view directions related to the vertex. Finally, for image synthesis under novel human poses, we replace original view directions related to vertex v with the global one $\overline{\mathbf{d}}_v$ for each ray.

4 Experiments

4.1 Datasets and metrics

Datasets. Since there is no existing dataset for our task, we create two new datasets for evaluation. The first dataset *TotalHuman* consists of four subjects performing specified actions. For each person, we use a fixed camera to record four monocular videos focusing on the body, head, and two hands, respectively. For the body video, the person turns around while holding a T-pose. For the head video, the person also rotates themselves but the camera focuses on the head. For the two hands, the subject holds a fixed hand pose and moves around. We use this dataset for qualitative evaluation.

For quantitative evaluation, we create a synthetic dataset *SynTotalHuman* which contains four animated 3D characters from Mixamo [2]. For each character, similar to the *TotalHuman*, we render four monocular part-specific videos for training and render images under novel human poses for evaluation. In addition, we also utilize this dataset to evaluate the accuracy of 3D reconstruction.

Metrics. For the evaluation of image synthesis, we adopt the following metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) [50]. For 3D reconstruction, we adopt the following two metrics: point-to-surface Euclidean distance (P2S) and Chamfer distance (CD), whose units are both in cm.

Table 1: Results of 3D reconstruction of each part on *SynTotalHuman* dataset.

	Head		Hands		Total	
	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓
NeuralBody [33]	1.55	1.46	1.29	1.19	2.16	1.98
AniNeRF [31]	1.80	1.69	1.12	1.00	2.55	2.21
AniSDF [32]	0.76	0.91	0.79	0.75	1.90	1.97
Ours	**0.59**	**0.82**	**0.55**	**0.52**	**1.84**	**1.89**

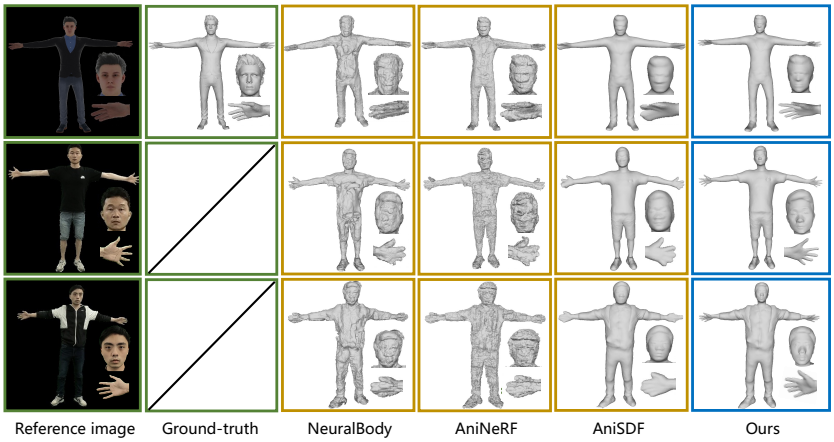

Figure 3: 3D reconstruction on the *SynTotalHuman* and *TotalHuman* datasets.

4.2 Comparison with the baselines

Since most previous methods only focus on body modeling, we extend the state-of-the-art methods AniNeRF [31] and AniSDF [32] with the hands to compare with our method. Since the neural feature field of AniSDF can only converge on small images, we leverage the color field for its image synthesis. We also compare with NeuralBody [33].

3D reconstruction. We first compare the 3D human reconstruction of our method and other baselines. In addition to the full-body evaluation, we also compare the reconstruction of the head and hands individually. The quantitative results on the *SynTotalHuman* dataset are shown in Table 1. Thanks to the multi-part representation and leverage of part-specific videos, our method outperforms the baselines in all parts, especially the head and hands. We present the qualitative results in Figure 3.

Image synthesis. We evaluate the image synthesis on the *TotalHuman* and *SynTotalHuman* datasets. Specifically, we compare the rendering results of novel human poses under novel views. Similar to the 3D reconstruction, we compare the rendered images of each part. The quantitative results on *SynTotalHuman* dataset are shown in Table 2. The results show that our approach achieves the best performance. Figure 4 shows the qualitative results on both datasets. As we can see, the head and hands images rendered by our method significantly outperform the counterparts of baselines.

4.3 Ablation studies

We conduct the ablation studies to justify the algorithm design in the proposed method.

Table 2: Results of image synthesis under novel human pose of each part on *SynTotalHuman* dataset.

	Head			Hands			Total		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody [33]	18.94	0.914	0.184	18.51	0.893	0.240	19.97	0.855	0.239
AniNeRF [31]	20.92	0.928	0.153	20.60	0.907	0.179	23.71	0.889	0.185
AniSDF [32]	21.91	**0.934**	0.104	21.31	0.930	0.111	25.57	0.916	0.129
Ours	**22.23**	**0.934**	**0.084**	**22.49**	**0.941**	**0.076**	**26.15**	**0.921**	**0.114**

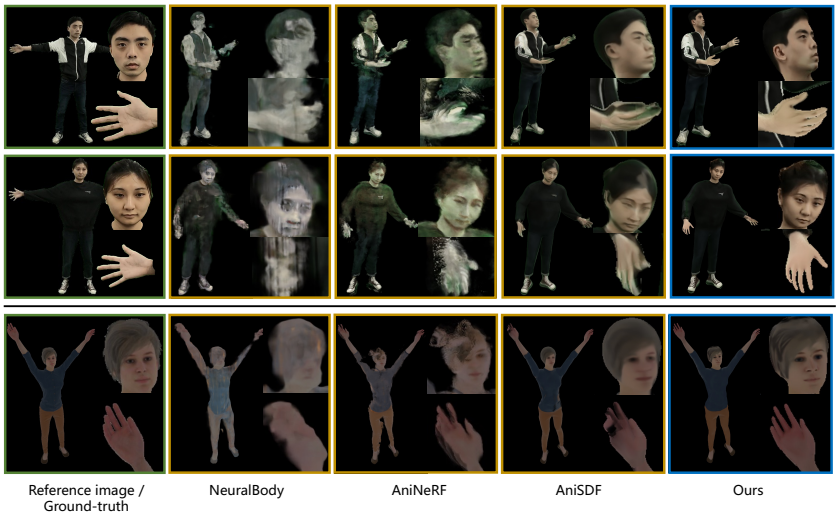

Figure 4: Image synthesis under novel human poses on *TotalHuman* and *SynTotalHuman* dataset. Note that for the *TotalHuman* dataset (first two rows), there is no ground-truth images under novel poses and we put the training images for reference.

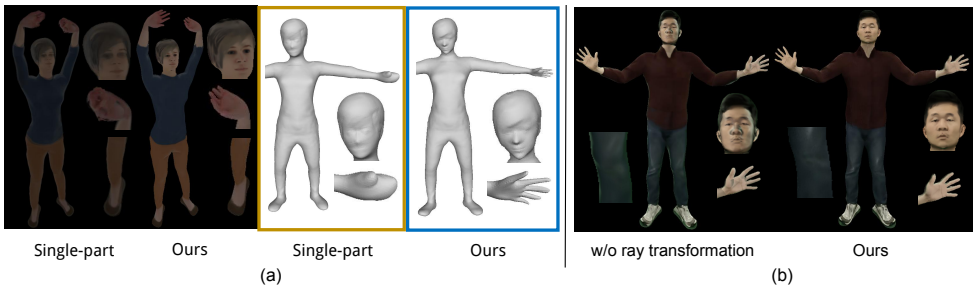

Figure 5: Ablation studies for (a) multi-part networks and (b) non-rigid ray transformation .

Multi-part networks. We use multiple networks to represent the different human parts as described in Section 3.1. An alternative is to represent the whole body using a single network which owns the same number of parameters with the multi-part network. The quantitative results of 3D reconstruction and image synthesis are presented in Table 3 and Table 4, respectively. The results show that our multi-part networks outperform the single-part network by a large margin. We also show the qualitative results in Figure 5 (a).

Head deformation. As described in section 3.2, we first use FLAME-SMPLH registration to initialize the rigid transformation and then refine it with the neural surface alignment. Here, we compare it with the initialized transformation obtained from the FLAME-SMPLH registration. The qualitative results are shown in Figure 6 (a). As we can see, the neural surface alignment significantly improves the alignment between the head part and body part.

Geometry composition. We introduce the L_g loss in Equation (7) to ensure smooth surface transitions between two adjacent sign distance fields. Here, we compare it to the method without this loss. Figure 6 (b) presents the qualitative results on *TotalHuman* dataset. With the L_g loss, the results of geometry transition between two parts are significantly improved.

Appearance composition. For recovering consistent appearance, we first optimize the latent code of head and hands color fields networks and then fuse the color predictions. To evaluate the effect, we compare with: 1) w/o composition: neither latent code optimization nor color fields fusion is used; 2) w/o color fields fusion: no color fields fusion is performed. The qualitative results are shown in Figure 6 (c). As we can see, our method presents much better results of appearance consistency and appearance transition.

Table 3: Ablation study for 3D reconstruction of multi-part networks on *SynTotalHuman* dataset.

	Head		Hands		Total	
	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓
single part model	1.03	1.26	1.25	1.22	1.95	1.97
Ours	**0.59**	**0.82**	**0.55**	**0.52**	**1.84**	**1.89**

Table 4: Ablation study for image synthesis of multi-part networks and dynamic ray transformation on the *SynTotalHuman* dataset.

	Head			Hands			Total		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
single part model	20.38	0.919	0.120	20.75	0.927	0.140	25.19	0.907	0.144
w/o ray transform	21.07	0.933	0.095	22.14	0.939	0.085	25.43	0.916	0.123
Ours	**22.23**	**0.934**	**0.084**	**22.49**	**0.941**	**0.076**	**26.15**	**0.921**	**0.114**

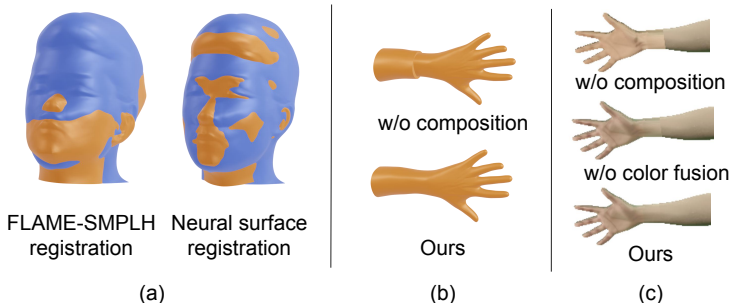

Figure 6: (a) Ablation study for head deformation. The blue mesh is the target mesh reconstructed from the body video. The orange mesh is the source mesh reconstructed from the face video. (b) Ablation study for geometry composition. (c) Ablation study for appearance composition.

Non-rigid ray transformation. After training, we introduce the non-rigid ray transformation to improve the novel view synthesis under novel human poses. To analyze its effect, we compare it with no ray transformation. The quantitative and qualitative results are shown in Table 4 and Figure 5 (b), respectively. The results indicate that our ray transformation greatly improves the rendering quality.

5 Limitations

The proposed method has the following limitations. First, the self-rotation human video provides very limited human motions, which makes our method have difficulty modeling pose-dependent deformations. It would be interesting to learn the prior of pose-dependent deformations from public datasets and finetune on the input videos. Second, our method still needs a relatively long time for training. Recently, based on the hash encoding, [27] significantly reduces the training time of implicit functions, and combining this representation into our model is left as future work.

6 Conclusion

In this paper, we introduced TotalSelfScan, a convenient method to create high-quality full-body avatars from several monocular self-rotation videos that focus on the face, hands, and body, respectively. We propose a multi-part network to represent the whole human in the canonical space and the part deformation is utilized to establish the correspondences between part observation spaces and the canonical space. We also propose the part compositions to obtain a consistent unified human model. After training, the non-rigid ray transformation is adopted to render photorealistic free-viewpoint videos under novel human poses. Both quantitative and qualitative results demonstrate the effectiveness of TotalSelfScan to reconstruct high-fidelity avatars from monocular videos. Lastly, using the reconstructed avatars to synthesize unauthorized personal images may have negative societal impact and we strongly discourage such applications.

References

[1] Easymocap. https://github.com/zju3dv/EasyMocap/.

[2] Mixamo. https://www.mixamo.com/.

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018.

[4] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, pages 5461–5470, 2021.

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *SIGGRAPH*, pages 408–416. 2005.

[6] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM TOG*, 40(4):1–17, 2021.

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[8] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, 2013.

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, pages 11594–11604, 2021.

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787. Springer, 2020.

[11] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, pages 20–40. Springer, 2020.

[12] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *SIGGRAPH*, pages 1–10. 2008.

[13] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *ECCV*, pages 612–628. Springer, 2020.

[14] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. IEEE, 2009.

[15] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM TOG*, 38(2):1–17, 2019.

[16] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, pages 5052–5063, 2020.

[17] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022.

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018.

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019.

[20] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 36(6):194–1, 2017.

[21] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021.

[22] 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 TOG*, 40(6):1–16, 2021.

[23] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM TOG*, 40(4):1–13, 2021.

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

[25] 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 *European conference on computer vision*, pages 405–421. Springer, 2020.

[26] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, 2022.

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

[28] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, pages 5762–5772, 2021.

[29] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A spare trained articulated human body regressor. In *ECCV*, pages 598–613, 2020.

[30] 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 *CVPR*, pages 10975–10985, 2019.

[31] 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*, pages 14314–14323, 2021.

[32] 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.

[33] 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*, pages 9054–9063, 2021.

[34] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM TOG*, 36(6):1–17, 2017.

[35] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV*, pages 1749–1759, 2021.

[36] 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 *ICCV*, pages 2304–2314, 2019.

[37] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 84–93, 2020.

[38] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, pages 2886–2897, 2021.

[39] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *ICCV*, pages 7750–7759, 2019.

[40] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[41] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, volume 34, 2021.

[42] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10965–10974, 2019.

[43] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM TOG*, 40(6):1–15, 2021.

[44] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *NeurIPS*, volume 34, 2021.

[45] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, pages 6184–6193, 2020.

[46] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *CVPR*, pages 4968–4978, 2020.

[47] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM TOG*, 37(2):1–15, 2018.

[48] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34, 2021.

[49] Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, and Xing Tang. Ray priors through reprojection: Improving neural radiance fields for novel view extrapolation. *CVPR*, 2022.

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[51] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, 2021.

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, pages 7739–7749, 2019.

[53] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, pages 4811–4822, 2021.

[54] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, pages 4491–4500, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All data used in the paper and our code will be published.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the supplementary materials due to the space limit.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Unfortunately, the error bars are not provided due to resource constraints. We will add this in the future.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the supplementary materials due to the space limit.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]