

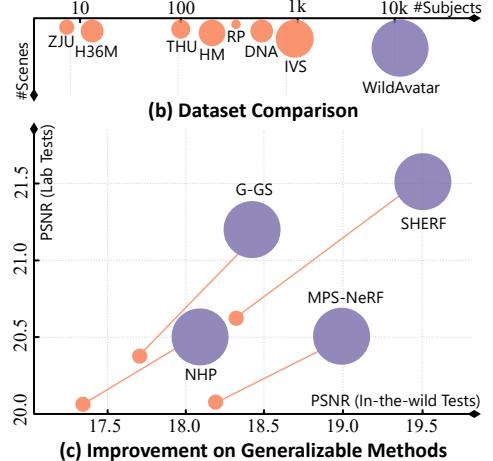
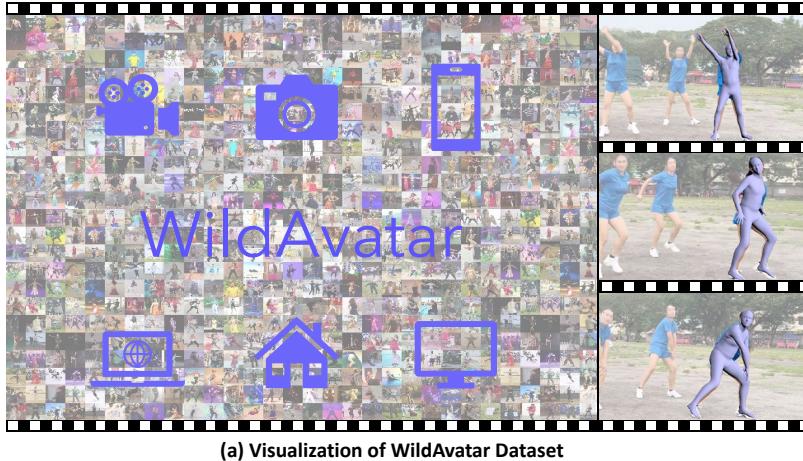
WildAvatar: Learning In-the-wild 3D Avatars from the Web

Zihao Huang^{1,2}, Shoukang Hu², Guangcong Wang³, Tianqi Liu^{1,2},
Yuhang Zang⁴, Zhiguo Cao¹, Wei Li^{2,†}, Ziwei Liu²

¹School of AIA, Huazhong University of Science and Technology,

²S-Lab, Nanyang Technological University, ³Great Bay University, ⁴Shanghai AI Laboratory

<https://wildavatar.github.io/>



(c) Improvement on Generalizable Methods

Figure 1. Overview of WildAvatar. (a) Unlike previous laboratory datasets for 3D avatar creation, WildAvatar curates in-the-wild web videos. (b) With 10k+ human subjects and scenes, WildAvatar is at least 10× richer than the previous datasets. (c) It contains high-quality annotations and demonstrates impressive potential to boost the quality and generalizability of avatar-creation methods.

Abstract

Existing research on avatar creation is typically limited to laboratory datasets, which require high costs against scalability and exhibit insufficient representation of the real world. On the other hand, the web abounds with off-the-shelf real-world human videos, but these videos vary in quality and require accurate annotations for avatar creation. To this end, we propose an automatic annotating pipeline with filtering protocols to curate these humans from the web. Our pipeline surpasses state-of-the-art methods on the EMDB benchmark, and the filtering protocols boost verification metrics on web videos. We then curate **WildAvatar**, a web-scale in-the-wild human avatar creation dataset extracted from YouTube, with 10,000+ different human subjects and scenes. WildAvatar is at least 10× richer than previous datasets for 3D human avatar creation and closer to the real world. To explore its potential, we demonstrate the quality and generalizability of avatar creation methods on WildAvatar. We will publicly release our code, data source links and annotations to push forward 3D human avatar creation and other related fields for real-world applications.

†: Corresponding author.

1. Introduction

3D Human Avatar creation has extensive applications in VR/AR, film-making, metaverse, etc., attracting significant attention recently. With the advent of neural radiance fields (NeRF) [57, 77], recent research aims to recover 3D avatars from 2D observations, enabling the synthesis of novel images from arbitrary viewpoints and body poses [21, 29, 31, 45, 59, 60, 84]. For example, many works focus on reconstructing animatable human models from well-annotated and calibrated multi-view images and videos [59–61, 85], and have achieved photo-realistic results in laboratory datasets [4, 60, 64]. However, their performances are restricted by small-scale indoor human data, especially for generalizable creation. This is down to the fact that current avatar creation datasets are mainly collected via well-designed laboratory systems, which are expensive and time-consuming to collect. In addition, there still exists domain gaps between laboratory and real-world scenarios. As shown in Table 1, existing human data collections mostly rely on annotations from advanced devices, such as well-calibrated multi-view cameras [16, 56, 73, 95], depth sensors [9, 33, 95, 100], IMUs [79], or expensive scanners [64, 93], as well as specialized actors and light-

Dataset	#Sub./Sce.	Type	Cost	Dataset	#Sub./Sce.	Type	Cost
ZJU-Mocap [60]	9/6	Lab	 + 	RenderPeople [64]	482/-	Lab	 + 
HuMMan [9]	339/20	Lab	 +  + 	THuman [100]	100/10	Lab	 + 
DyMVHumans [99]	32/45	Lab	 + 	THuman2.0 [95]	500/-	Lab	 +  + 
Human3.6M [33]	11/15	Lab	 +  + 	THuman3.0 [73]	20/-	Lab	 +  + 
THuman4.0 [101]	3/3	Lab	 + 	THuman5.0 [70]	10/10	Lab	 + 
Hi4D [93]	40/10	Lab	 +  + 	DNA [16]	500/1187	Lab	 +  + 
DynaCap [26]	4/5	Lab	 + 	MultiHuman [98]	50/-	Lab	 + 
UltraStage [102]	100/20	Lab	 + 	Actors-HQ [35]	8/21	Lab	 + 
HUMBI [94]	772/4	Lab	 + 	NHR [85]	4/3	Lab	 + 
AIST++ [78]	30/10	Lab	 + 	MPII [5]	8/1	Lab	 + 
ENeRF [50]	7/4	ITW	 + 	3DPW [79]	7/4	ITW	 +  + 
FreeMan [80]	40/123	ITW	 + 	SynWild [25]	5/5	ITW	--
NeuMan [37]	6/6	ITW	--	TikTok [36]	340/340	ITW	--
IVS-Net [18]	<700/<700	ITW	--	WildAvatar	10k+/10k+	ITW	--

Table 1. Statistics on different human datasets for avatar creation. We only consider human datasets that include human appearances. : multi-camera system. : 3D scanner. : depth sensor. : inertial measurement unit. : professional actor. Our WildAvatar is a *large-scale* and *in-the-wild* avatar dataset collected with our designed *automatic collection* pipeline.

stages [26, 50, 78, 80, 94]. These ideal conditions require high costs against scalability, exhibit limited representations of the real world, and are unavailable in in-the-wild scenarios (*i.e.*, monocular web videos) or consumer applications.

To tackle this problem, recent efforts are attempting to collect in-the-wild monocular human data [18, 20, 36] from the web. However, they still rely heavily on costly manual interventions, making it difficult to scale up. Therefore, their inadequate diversity fails to meet the requirements for dealing with in-the-wild challenges of 3D avatar creation. Given the shortage of real-world human data, scaling up avatar creation from real-world scenarios is worth exploring.

To this end, we propose a novel pipeline with filtering protocols to extract human movements automatically from the web. Specifically, we first streamline the annotation process with off-the-shelf state-of-the-art annotation methods [2, 23, 41, 63], *e.g.*, YOLO [63], Segment Anything [41]. We then propose a suite of assessment protocols for automatic filtering to retain only qualified video clips (*e.g.*, without occlusions and annotated in high confidence). We report our performance on the EMDB benchmark to investigate our pipeline’s performance in real-world scenarios. With the help of the additional stage for aligning and smoothing, our pipeline surpasses state-of-the-art methods (up to 4%). We also demonstrate that the proposed filtering protocols can boost verification metrics among web video sources.

We implement the proposed annotating and filtering pipeline to collect real-world human samples from the web. Without costly sensors or lightstages, we curate WildAvatar, a large-scale in-the-wild human avatar creation dataset extracted from YouTube, with 10,000+ different real-world human subjects and scenes. WildAvatar fills the gap in large-

scale in-the-wild human data collection, offering at least 10× richer human subjects than the existing human datasets.

To explore the potential of WildAvatar, we investigate the scalability of existing generalizable human avatar creation methods [21, 31, 45] with WildAvatar, which suggests a significant improvement (up to 7%) in real-world scenarios. We will release the WildAvatar dataset, providing the video IDs, frame IDs, extracting scripts, and annotations obtained from our pipeline. We hope that our dataset will push forward the development of 3D human avatar creation and other related topics, *e.g.*, human mesh estimation (HPS) [32, 39, 42–44, 48, 71, 96], 3D human avatar generation [11, 14, 28, 30, 54, 81], 3D human interactions [72, 82], and 3D relightable avatar [13, 34, 88]. Overall, our main contributions can be summarized as follows:

- We propose a novel annotating pipeline with filtering protocols to curate human movements from the web. Our pipeline surpasses state-of-the-art methods on the in-the-wild benchmark, and our filtering protocols can boost verification metrics on annotating web video sources.
- We curate WildAvatar, a large-scale avatar dataset collected from in-the-wild videos with 10,000+ human subjects, at least 10× larger than previous datasets. The scale-up facilitates the creation of per-subject avatars and paves a new avenue for generalizable avatar reconstruction.
- We illustrate the great potential of our pipeline and WildAvatar, with exploratory experiments on supporting popular downstream 3D avatar creation applications. Further data scale-up for large-scale model training will be unlocked. The open code and data could provide important insights for future avatar creation and relevant tasks.

2. Related Work

SMPL Annotation in the Wild. Parametric models (*e.g.*, SMPL [55], SMPLX [58]) encode coarse human surfaces with pose and shape parameters for 3D avatar creation. For example, end-to-end methods [32, 39, 43, 48, 96] estimate the SMPL parameters efficiently from single in-the-wild images, but they are not robust with complex scenes. Subsequent works [38, 44] refine end-to-end estimation results via fitting SMPL to 2D annotations in the loop. Recent works further consider imposing temporal consistency [2, 42, 75] to smooth SMPL estimations across in-the-wild videos. Yet, it is far from achieving a comprehensive annotating and filtering streamline for real-world video sources.

3D Avatar Creation Datasets. Table 1 presents a system overview of previous avatar creation datasets. Previous datasets mainly obtain high-quality human mask and SMPL annotations with ideal laboratory systems, *e.g.*, well-calibrated multi-view cameras [9, 16, 26, 33, 35, 50, 56, 60, 70, 73, 78, 79, 85, 93–95, 98, 99, 101, 102], depth sensors [9, 16, 33, 73, 95, 100], IMUs [79], or expensive scanners [64, 93], as well as specialized actors and light-stages [9, 16, 26, 33, 35, 50, 56, 60, 64, 70, 73, 78, 79, 85, 93–95, 98–102], depth sensors [9, 16, 33, 73, 95, 100]. Unfortunately, these optimal conditions are typically high-cost against scaling up and limited in scenarios compared to the real world. Towards in-the-wild avatar creation, previous efforts also attempt to collect avatar data from web human movement videos [18, 36]. However, they rely heavily on costly manual interventions (*e.g.*, pre-filtering or mask extraction based on [52]) and still fail to scale up. And they either show little viewpoint change [36] or not public released [18]. Therefore, there are still significant demands for a public large-scale in-the-wild dataset for avatar creation.

3D Avatar Creation Methods. Existing works on avatar creation learn coarse human surfaces with parametric mesh models [6, 55, 58, 65, 87], but these explicit meshes cannot express detailed geometry or appearance. Pifu-based methods [19, 27, 49, 66, 67] represent the human body with pixel-aligned functions and require high-quality synthetic data, but fail to generalize to real-world scenarios due to the domain gap. NeRF-based approaches [8, 22, 47, 59–61, 84, 91] learn implicit human representation from high-quality videos, and achieve photo-realistic rendering in the laboratory benchmarks. In addition, generalizable models [21, 31, 45] further simplify the data demand to a single image. Nonetheless, these methods still rely on accurate annotations, which only exist in ideal laboratory environments. Recent attempts [25, 37] decompose avatars and scenes for construction at once, demanding accurate global alignments between human bodies and backgrounds that are not available in real-world videos. Web-scale in-the-wild video data is crucial for the next phase of 3D avatar development.

3. Methodology

Our goal is to build a large-scale in-the-wild dataset for human avatar creation. To this end, we collect abundant real-world human videos from YouTube and annotate corresponding labels. Without time-consuming and high-cost manual filtering and annotation [18], we design an efficient pipeline to filter video candidates and obtain high-quality annotations. Specifically, we download human movement videos from the web with our automatic tools, as described in Section 3.1. Then, we collect high-quality annotated data with several processing steps on downloaded web videos (*e.g.*, annotating in Section 3.2 and filtering in Section 3.3). Finally, we evaluate our designs in detail in Section 3.4.

3.1. Data Collection

We follow two steps to collect human movement videos: 1) We first search for a large and diverse set of video candidates on YouTube. Note that video candidates could contain some noisy videos without human subjects; 2) We then download the candidate videos with automatic tools [3], cut them into manageable clips, and filter out short video clips. Details can be found in Appendix B.

3.2. Data Annotating

To obtain high-quality annotations (*i.e.*, SMPL, camera parameters, and human segmentation mask) for the collected human videos, we design the following four stages to annotate the in-the-wild human video clips automatically.

Stage I: Human Bounding Box Detection and Tracking. Precise human bounding boxes are important for monocular SMPL estimation methods [39, 43, 48, 96] and human segmentation (using *Segment Anything* [41]). Hence, we first obtain the bounding box of human subjects with off-the-shelf state-of-the-art detection methods [63, 86].

Stage II: Human Segmentation Mask Extraction. Foreground segmentation masks that distinguish human subjects from the background are important for 3D avatar creation. Previous methods [18, 84] adopt the *background matting* [52] methods for segmentation. These methods, however, require manual intervention to select the collected videos with at least one still background image, which is time-consuming to prepare. To efficiently segment human video clips, we adopt state-of-the-art *Segment Anything* (SAM) [41], which only requires bounding boxes and keypoints for each frame obtained in previous human detection, tracking, and 2D pose estimation steps (See Fig. 2).

Stage III & IV: SMPL and Camera Estimation. We first estimate SMPL and camera parameters frame by frame, using state-of-the-art single-image-based human pose and shape estimation methods [23, 43] in Stage III. However, these coarse parameters may ignore the temporal consistency of human motions. Inspired by previous SMPL annotation pipelines [2, 38, 44], we further refine SMPL and camera

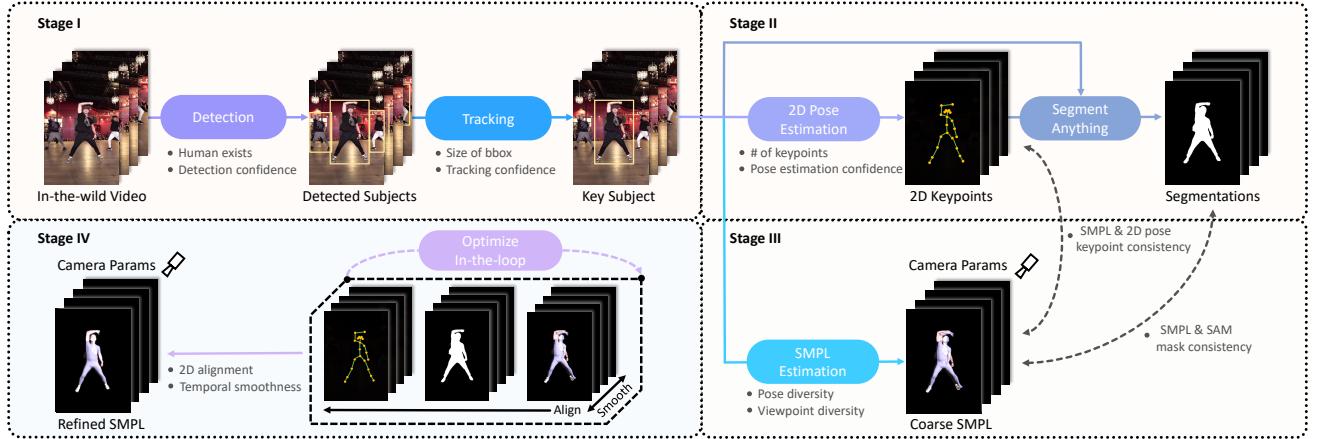


Figure 2. The four-stage data processing pipeline. We first obtain the bounding box of key subjects in videos in Stage I and extract human segmentation masks in Stage II. Then, the SMPL and camera parameters are coarsely estimated in Stage III and later refined in Stage IV.

parameters in Stage IV, smoothing these parameters across the whole video clip sequence via gradient descent. We also incorporate the estimated 2D keypoints and SAM masks into a refinement loop, providing additional supervision for precise SMPL and camera parameter estimation. Qualitative comparisons between coarse and refined SMPL and camera parameters can be found in Fig. E.

3.3. Data Filtering

In contrast to high-quality human videos collected in well-designed laboratory systems, real-world videos are not always qualified for annotating or curating into datasets. To automatically filter out these unqualified videos (*e.g.*, in severe occlusions or low resolution), we propose a suite of assessment protocols within the four-stage pipeline.

Protocol I: Clear Body with Significant Movement. The primary focus is to ensure the presence of a clear human body with no occlusion. To this end, we only retain in-the-wild videos with high confidence in both detection and 2D pose estimation, as illustrated in Fig. 3 (a) and (b). Then, we further reduce occlusion by only focusing on the key subject within each video clip. This is achieved by calculating the average size of the bounding box that encloses the tracked subject throughout the entire video. We also eliminate those trivial human movement videos, specifically those too brief or with insignificant viewpoint shifts and human movements.

Protocol II: Ensemble of Annotating Experts. To ensure the quality of in-the-wild annotation, we adopt the ensemble of different state-of-the-art annotating model experts. Specifically, we denote the model libraries for detection [63, 86], 2D pose estimation [10, 74, 92], and SMPL Estimation [23, 43] as $\{\mathcal{D}, \mathcal{J}, \mathcal{S}\}$, and their predictions as $\{D_i, J_i, S_i\}$. We calculate the average $\{\mu(D_i), \mu(J_i), \mu(S_i)\}$ as the fused predictions, and the standard deviation $\{\sigma(D_i), \sigma(J_i), \sigma(S_i)\}$ as references for filtering. To reduce annotating errors, we only retain the videos that are annotated consistently among

different state-of-the-art annotating experts (see Fig. 3 (c)).

Protocol III: Consistency on 2D Keypoints. To augment the reliability of SMPL estimations, we implement an additional double-check of the consistency in 2D keypoints, between the monocular SMPL estimation and 2D pose estimation results. Specially, we project the 3D SMPL keypoints $J_{3D}(\mu(S_i))$ to 2D. Then, we compute the Percentage of Correct Keypoints (PCK) [90] between the projected $\Pi(J_{3D}(\mu(S_i)))$ and those predicted from 2D pose estimation $\mu(J_i)$. Subsequently, we discard video clips that exhibit low average PCK values to ensure that only videos with high annotation confidence are retained for further curation.

Protocol IV: Consistency in SMPL and SAM Masks. With inaccurate predictions, monocular SMPL estimation and SAM might have low overlaps in estimated human masks. We double-check their consistency by comparing SMPL projection masks with the SAM masks. Intuitively, the SMPL mask denotes the naked body, while the SAM mask contains the clothed body. Therefore, the SMPL mask should be mostly covered by the SAM mask. To guarantee high-quality annotations, we discard video clips whose SAM-masked region has small overlaps with the SMPL-masked region, as illustrated in Fig. 3 (e) (more examples in Fig. F).

3.4. Analysis

Evaluation of Annotating Pipeline. To demonstrate the accuracy of our four-stage pipeline among in-the-wild videos, we first provide quantitative analysis on real-world EMDB [40] benchmark in Table 2. Due to the additional 2D aligning and temporal smoothing in Stage IV, our annotations have more accurate alignment on extremities (*e.g.*, feet, arms, and heads). Our pipeline surpasses existing methods, especially on the MPJPE [33] (+3.27%) and PVE [89] (+4.16%), where global alignments are not taken into account. Evaluation details and qualitative comparisons can be found in Appendix C.



Figure 3. Visualizations of filtering protocols. We only retain video clips that (a) show high confidence in human detections; (b) obtain high average confidence in 2D pose estimations; (c) consistency annotated by different expert models; (d) consistency on keypoints between projected SMPL keypoints and 2D pose estimations; and (e) consistency on segmentation masks between Segment-Anything and SMPL.

Method	PA-MPJPE \downarrow	MPJPE \downarrow	PVE \downarrow
SPIN [44]	87.1	140.3	174.9
VIBE [42]	81.4	125.9	146.8
PARE [43]	72.2	113.9	133.2
TRACE [76]	70.9	109.9	127.4
CLIFF [48]	68.1	103.3	128.0
TCMR [17]	65.6	119.1	137.7
HybrIK [46]	65.6	103.0	122.2
HMR2.0 [23]	60.6	98.0	120.3
Ours	59.9	94.9	115.5

Table 2. Performance comparisons on the public EMDB [40] benchmark. Our proposed four-stage pipeline surpasses existing methods on the quality of SMPL annotations.

Verification of Filtering Strategies. We further demonstrate the efficacy of our filtering protocols and validate the quality of our final dataset. Given that there is no ground truth (GT) for evaluation, we adopt double-checks for verification to assess the quality of annotations for in-the-wild data. Specifically, we evaluate the consistency between predicted annotations after applying our protocols. The detailed results are presented in Table 3, where PCK represents the consistency between the 2D pose estimation models and the SMPL models, while the SOIOU measures the proportion of the SMPL masks that lie outside the SAM mask. Comparing adjacent columns, we find our pipeline designs compelling: 1) Protocol I is a vital pre-processing, which brings a marginal improvement on PCK and SOIOU. It eliminates a substantial number of clips in poor quality (*e.g.*, no human subjects or movements); 2) Protocol II also yields a significant enhancement on PCK (10.1%), demonstrating the importance of adopting various annotating experts. 3) Protocol III & IV step further towards consistency, raise PCK by 7.51% and drop SOIOU by 0.094 in total; and 4) Stage IV finally refines the SMPL annotations, resulting in PCK over 0.92 and SOIOU lower than 0.03. Notably, this final PCK deviates by only 1.7% from the 3DPW dataset.

P.I	P.II	P.III	P.IV	S.IV	PCK \uparrow	SOIOU \downarrow	#Sub.
--	--	--	--	--	0.282	0.760	465801
✓	--	--	--	--	0.762	0.214	43824
✓	✓	--	--	--	0.839	0.146	25392
✓	✓	✓	--	--	0.882	0.129	12482
✓	✓	✓	✓	--	0.902	0.052	10647
✓	✓	✓	✓	✓	0.921	0.028	10647
Ground Truth of 3DPW				0.937	--	--	7

Table 3. Evaluation of the proposed pipeline on web videos. PCK (threshold 0.1) represents the consistency between the 2D pose and SMPL estimation, while SOIOU measures the proportion of the SMPL mask outside the SAM mask. #Sub. denotes the number of qualified video clips. P.X/S.X denotes the Protocol/Stagex.

4. WildAvatar

Our collected WildAvatar contains 10,647 real-world human video clips. We randomly split WildAvatar into training, validation, and test splits with 7k, 1.5k, and 1.5k subjects, respectively. Notably, although selected by our protocols, the remaining curated data is still much more diverse than in the laboratory systems (See Section 4 for statistics and Fig. A for unusual camera viewpoints and body poses). We show data statistics of our WildAvatar in Fig. 4 and describe the details of pose, viewpoint, and clothing distribution as follows.

Poses: Various vs. Trivial. In Fig. 4 (c), we compare the pose distribution differences among WildAvatar and previous human datasets. Specifically, we visualize the top-2 components of body poses with t-SNE [62] dimension reduction. The pose distribution comparison shows that our WildAvatar contains more diverse poses than previous laboratory datasets, as laboratory datasets only contain several designed motion sequences. This statistic illustrates the importance of our large-scale in-the-wild WildAvatar when applying the existing models to real-world scenarios.

Viewpoints: Free vs. Fixed. We compare viewpoint distribution among previous human datasets [9, 60, 100] and

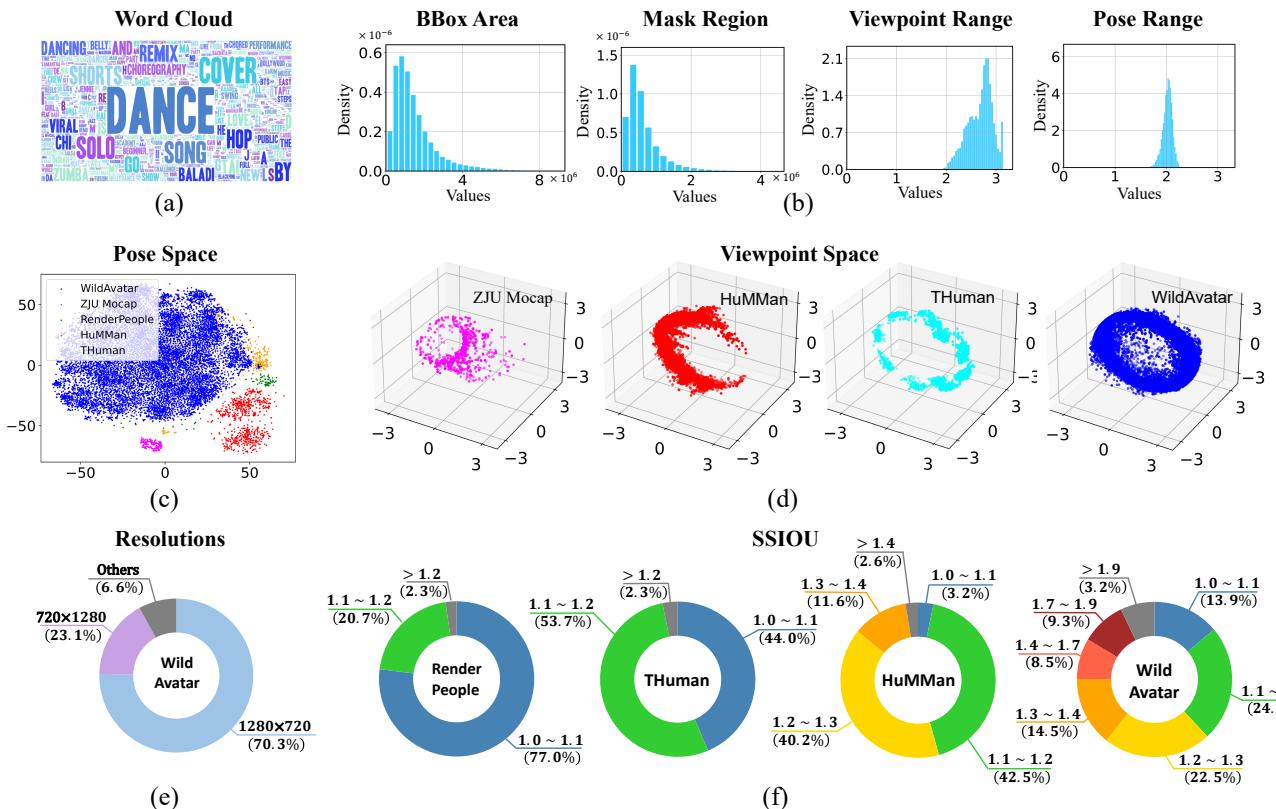


Figure 4. Data Analysis: (a) word cloud of the video titles in WildAvatar, (b) histograms of annotations across video clips, here we count the bounding box and human mask region in pixels, and “Range” denotes the difference between the maximum and minimum values. (c) comparison of the body pose spaces with popular laboratory human datasets, (d) comparison of the viewpoints spaces with popular laboratory human datasets, (e) resolutions of videos in WildAvatar, and (f) comparison with the previous dataset on the abundance of clothing. We introduce the SSIOU, the inverse IOU between SMPL [55, 58] masks and segmentation masks.

WildAvatar. Specifically, we visualize the 3D rotations of observed cameras relative to human subjects. As suggested in Table 1, previous laboratory datasets are mainly collected in indoor lightstages, where the RGB(D) cameras’ positions are fixed for whole datasets. In real-world scenarios, however, the viewpoints of humans are arbitrary. As shown in Fig. 4 (d), the viewpoint distributions in previous datasets are sparse or unbalanced, while our WildAvatar covers a more dense viewpoint distribution.

Cloth: Diversity vs. Unitary. Previous laboratory human datasets mainly contain unitary tight clothes, while complex real-world appearances (*e.g.*, different hairstyles) and diverse loose clothes are seldom involved. Our WildAvatar involves diverse tight and loose clothes in the real world. To quantitatively measure the clothing diversity between real-world and laboratory data, we define a metric (SSIOU) that calculates the inverse IOU between SMPL [55, 58] projected masks and human segmentation masks among different datasets in Fig. 4 (f). For humans with tight clothes, the measured SSIOU is close to 1.0 (blue in Fig. 4 (f)), indicating their SMPL masks are similar to human segmentation masks. Correspondingly, a larger SSIOU (red in Fig. 4 (f)) reveals that human clothes are looser, with segmentation masks being

larger than SMPL masks. Examples of various SSIOU can be found in Fig. D. As shown in Fig. 4 (f), more than 30% subjects in WildAvatar have SSIOU values over 1.4, suggesting that WildAvatar covers various types of clothes. This analysis validates that WildAvatar more satisfies data distributions in real-world scenarios.

5. Experiments

We conduct exploratory experiments on our WildAvatar with the following commonly used metrics (details in Section 5.1) on 3D avatar creation. Section 5.2 illustrates the real-world avatars from our WildAvatar, created by per-subject avatar creation methods. It also illustrates the improvement of our annotating pipeline, by comparing annotations from our pipeline or the state-of-the-art HMR2.0 [23]. Section 5.3 further demonstrates the potential of generalizable avatar creation methods, when provided our large-scale WildAvatar.

5.1. Evaluation Metrics

To quantitatively evaluate the quality of created avatars, we apply three commonly used metrics: peak signal-to-noise ratio (PSNR) [68], structural similarity index (SSIM) [83],

Method	HMR2.0			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NHR [85]	18.23	89.7	10.3	18.93	91.6	9.9
NB [60]	16.17	84.2	12.6	16.74	86.2	12.1
AN [59]	19.02	89.1	11.4	19.72	90.5	10.9
AS [61]	18.72	90.9	11.1	19.20	92.3	10.7
HN [84]	22.52	86.3	15.2	23.12	88.0	14.6
IN [22]	24.39	92.7	8.2	25.28	94.3	7.7
GH [29]	24.73	93.8	6.3	25.89	95.7	5.7

Table 4. Quantitative comparisons on WildAvatar. We report the quality of novel pose synthesis of popular methods on the state-of-the-art HMR2.0 annotations and our annotations, revealing the advantages of our pipeline towards in-the-wild avatar creation.

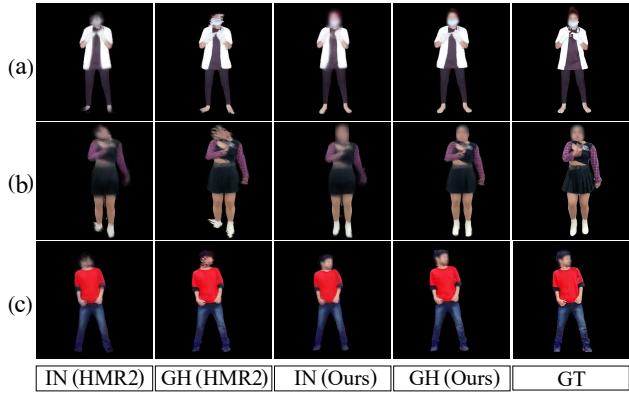


Figure 5. Qualitative comparisons of avatars created with our annotations and the state-of-the-art HMR2.0. IN and GH denote InstantNVR and GauHuman, respectively. More accurate avatars can be created with our annotations. Human faces are blurred.

and Learned Perceptual Image Patch Similarity (LPIPS) [97]. Following previous literature [31, 45], we compute whole-image metrics for per-subject reconstruction methods, while reporting the metrics based on projected 3D human bounding box areas for generalizable methods.

5.2. Per-Subject Avatar Creation

Setup. To explore the quality of WildAvatar and investigate the reconstruction performance on monocular in-the-wild videos with our annotations. We evaluate popular per-subject avatar creation methods on our WildAvatar dataset. We consider a total of 7 baselines: Neural Human Rendering (NHR) [85], NeuralBody (NB) [60], Animatable NeRF (AN) [59], Animatable SDF (AS) [61], HumanNeRF (HN) [84], InstantNVR (IN) [22], and GauHuman (GH) [29]. Considering the calculating and time cost, it is impractical and inefficient to train all subjects and scenes in WildAvatar. Instead, we manually select 133 representative human subjects from WildAvatar for exploring. For each subject, we randomly choose 100 frames for training and 100 for testing. We follow the default model settings (learning rate, batch size, etc.) in their official implementations and report

average metrics of these baselines in Table 4.

Quantitative and Qualitative Analysis. As shown in Table 4, more accurate avatars can be created with our annotations. Compared to the annotations from the state-of-the-art HMR2.0 [23], our annotations improve PSNR (+3.51%), SSIM (+1.91%), and LPIPS (−5.06%) on average. We further present qualitative comparisons in Fig. 5. Avatars from our annotations have more accurate extremities (*e.g.*, feet, arms, and heads), which are challenging to predict and align from single images. This improvement may prove the benefit of Stage IV (*e.g.*, 2D alignment and temporal smoothness).

5.3. Generalizable Avatar Creation

Setup. We evaluate four state-of-the-art generalizable baselines on our WildAvatar, including 1) Neural Human Performer (NHP) [21], 2) MPS-NeRF [45], 3) SHERF [31], and 4) Generalizable 3D Gaussian Splatting (Generalizable GS). Specifically, the Generalizable GS is adapted from GauHuman [29], with the Gaussians’ attributes decoded from the input features, following the approaches in [12, 53]. To evaluate the overall generalizability among various scenarios, we also report results on four laboratory datasets, including 1) ZJU-Mocap [60], 2) RenderPeople [64], 3) HuMMan [9], and 4) THuman [100]. To avoid overfitting to a small indoor dataset, we adopt the cross-domain testing strategy, *e.g.*, mask out the ZJU training set when testing on the ZJU.

Quantitative and Qualitative Analysis. There are three main observations in Tab. 5. First, large-scale WildAvatar is essential for improving generalizability towards in-the-wild scenarios. Comparing odd and even rows, the WildAvatar training set brings considerable improvements in real-world scenarios. On the WildAvatar test set, it improves PSNR, SSIM, and LPIPS by 4.52%, 3.53%, and 5.06% on average, respectively. Notably, this is fairly evident as over 60% of pixels are background. This improvement suggests that large-scale in-the-wild data is essential for the generalizable avatar creation methods. As shown in Fig. 6, models without (w/o) training on WildAvatar, tend to predict dark-colored artifacts on the lower body. Models with (w/) training on WildAvatar can provide more realistic appearances and geometries. Second, large-scale in-the-wild data is also beneficial for laboratory benchmarks. However, there are significant domain gaps, and the annotation of in-the-wild data is not as accurate as in the laboratory. By introducing WildAvatar dataset, we obtain an improvement of 4.17% and 4.54% PSNR on novel view and pose synthesis, respectively. Third, although 3D Gaussian Splatting (3DGS) performs well in per-subject human avatar creation, it shows slightly worse than NeRF in this generalizable setting. Our insight is that 3DGS highly relies on the point cloud initialization [15, 29, 53]. Unfortunately, due to the depth ambiguity, it is challenging to infer accurate point cloud initialization from a single image.

Method	WA	Training Set				WildAvatar Test			Novel Lab Test	
		RP	THU	HM	ZJU	PSNR(NP)	SSIM(NP)	LPIPS(NP)	PSNR(NV)	PSNR(NP)
NHP [21]	✓		✓	✓	✓	17.39	62.6	33.2	20.05	19.83
			✓	✓	✓	18.12	66.3	31.4	20.49	20.24
		✓		✓	✓	17.35	62.2	33.4	20.84	20.81
	✓	✓		✓	✓	18.07	66.9	31.2	21.20	21.27
MPS-NeRF [45]	✓	✓	✓		✓	18.21	74.2	24.0	16.99	17.03
	✓	✓	✓		✓	18.79	76.8	22.2	18.48	18.77
		✓	✓	✓	✓	18.20	74.3	23.8	20.17	20.03
	✓	✓	✓	✓		19.11	77.2	22.1	21.19	21.48
SHERF [31]	✓		✓	✓	✓	18.39	78.1	20.2	20.67	20.10
			✓	✓	✓	19.43	80.5	19.3	20.96	20.49
		✓	✓		✓	18.31	77.9	20.4	19.07	18.99
	✓	✓	✓		✓	19.50	80.7	19.2	19.57	19.64
Generalizable GS	✓	✓		✓	✓	17.76	81.1	22.8	20.31	20.03
	✓	✓		✓	✓	18.43	81.5	22.3	21.73	21.65
		✓	✓	✓	✓	17.73	81.2	22.7	21.47	21.78
	✓	✓	✓	✓		18.39	81.4	22.4	22.46	22.02

Table 5. Generalizability comparisons on challenging WildAvatar and laboratory benchmarks. We report the results of previous generalizable avatar creation methods on the cross-domain setting, offering a clear perspective across different domains. NP/NV denotes Novel Pose/View. RP/THU/HM/ZJU/WA are short for RenderPeople [64] / THuman [100] / HuMMan [9] / ZJU Mocap [60] / **WildAvatar (Ours)** dataset. Novel Lab Test refers to the test split of laboratory datasets not included in the training.

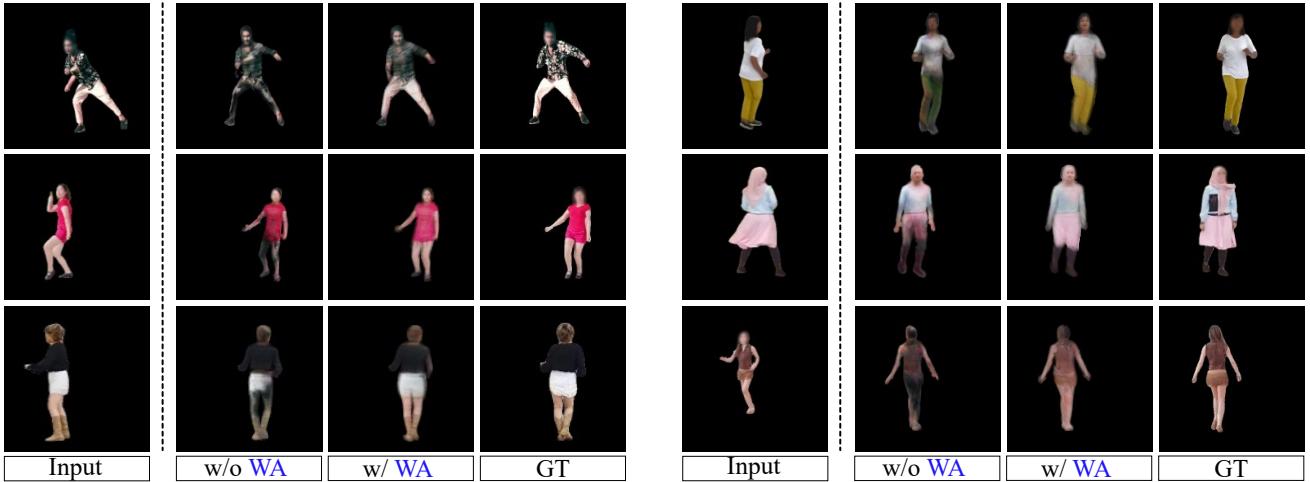


Figure 6. Qualitative comparisons on the state-of-the-art generalizable avatar creation method [31]. “w/ WA” or “w/o WA” denotes training with or without WildAvatar, respectively. Human faces are blurred to protect privacy.

6. Conclusion

This work introduces an automatic pipeline with filtering protocols for collecting in-the-wild human annotations from the web. From this pipeline, we curate WildAvatar, a large-scale in-the-wild human avatar creation dataset from YouTube, with 10,000+ various human subjects and scenes. Compared with traditional avatar creation datasets, our WildAvatar consists of at least 10× more subjects or scenes. We demonstrate the effectiveness of the proposed pipeline and

illuminate the potential of WildAvatar on avatar creation tasks with data. We hope our pipeline and dataset will shed insights on following in-the-wild human avatar creation and benefit 3D/4D human content generation works.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

- [1] Video scene cut detection and analysis tool. Github, 2014. [14](#)
- [2] Easymocap - make human motion capture easier. Github, 2021. [2](#), [3](#)
- [3] A feature-rich command-line audio/video downloader. Github, 2021. [3](#)
- [4] Thiem Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8387–8397. Computer Vision Foundation / IEEE Computer Society, 2018. [1](#)
- [5] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 3686–3693. Computer Vision Foundation / IEEE, 2014. [2](#)
- [6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005. [3](#)
- [7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *Proc. Eur. Conf. Comput. Vis.*, pages 561–578. Springer-Verlag, 2016. [14](#)
- [8] Aljaz Bozic, Pablo R. Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1450–1459. Computer Vision Foundation / IEEE, 2021. [3](#)
- [9] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. [1](#), [2](#), [3](#), [5](#), [7](#), [8](#)
- [10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186, 2018. [4](#)
- [11] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. [2](#)
- [12] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *arXiv*, 2023. [7](#)
- [13] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, pages 606–623. Springer, 2022. [2](#)
- [14] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdifusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems*, 36:13664–13677, 2023. [2](#)
- [15] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. *arXiv preprint arXiv:2402.14650*, 2024. [7](#)
- [16] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19925–19936. IEEE, 2023. [1](#), [2](#), [3](#)
- [17] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. [5](#)
- [18] Junting Dong, Qi Fang, Tianshuo Yang, Qing Shuai, Chengyu Qiao, and Sida Peng. ivs-net: Learning human view synthesis from internet videos. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 22885–22894. IEEE, 2023. [2](#), [3](#)
- [19] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. PINA: learning a personalized implicit neural avatar from a single RGB-D video sequence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20438–20448. IEEE, 2022. [3](#)
- [20] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021. [2](#)
- [21] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *CoRR*, abs/2203.16875, 2022. [1](#), [2](#), [3](#), [7](#), [8](#)
- [22] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8759–8770. IEEE, 2023. [3](#), [7](#)
- [23] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14737–14748. IEEE, 2023. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [14](#), [15](#), [16](#)

- [24] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. [16](#)
- [25] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12858–12868. IEEE, 2023. [2, 3](#)
- [26] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph.*, 40(4):94:1–94:16, 2021. [2, 3](#)
- [27] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [3](#)
- [28] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. [2](#)
- [29] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *CoRR*, abs/2312.02973, 2023. [1, 7](#)
- [30] Shoukang Hu, Fangzhou Hong, Tao Hu, Liang Pan, Haiyi Mei, Weiye Xiao, Lei Yang, and Ziwei Liu. Humanlif: Layer-wise 3d human generation with diffusion model. *arXiv preprint arXiv:2308.09712*, 2023. [2](#)
- [31] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. SHERF: generalizable human nerf from a single image. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9318–9330. IEEE, 2023. [1, 2, 3, 7, 8](#)
- [32] Zihao Huang, Min Shi, Chengxin Liu, Ke Xian, and Zhiguo Cao. Simhmhr: A simple query-based framework for parameterized human mesh reconstruction. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 6918–6927. ACM, 2023. [2, 3](#)
- [33] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. [1, 2, 3, 4, 15](#)
- [34] Umar Iqbal, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, and Jan Kautz. RANA: relightable articulated neural avatars. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23085–23096. IEEE, 2023. [2](#)
- [35] Mustafa Isik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Trans. Graph.*, 42(4):160:1–160:12, 2023. [2, 3](#)
- [36] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12753–12762. Computer Vision Foundation / IEEE, 2021. [2, 3](#)
- [37] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022. [2, 3](#)
- [38] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *Proc. Int. Conf. 3D Vis.*, pages 42–52. Computer Vision Foundation / IEEE, 2021. [3](#)
- [39] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 7122–7131. Computer Vision Foundation / IEEE, 2018. [2, 3](#)
- [40] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zarate, and Otmar Hilliges. EMDB: the electromagnetic database of global 3d human pose and shape in the wild. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14586–14597. IEEE, 2023. [4, 5, 15](#)
- [41] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. [2, 3](#)
- [42] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 5252–5262. Computer Vision Foundation / IEEE, 2020. [2, 3, 5](#)
- [43] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: part attention regressor for 3d human body estimation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 11107–11117. Computer Vision Foundation / IEEE, 2021. [3, 4, 5, 14](#)
- [44] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2252–2261. Computer Vision Foundation / IEEE, 2019. [2, 3, 5](#)
- [45] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24741–24752, 2021. [1, 2, 3, 7, 8](#)
- [46] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *CoRR*, abs/2304.05690, 2023. [5](#)
- [47] Rui long Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric

- performance capture. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, pages 49–67. Springer, 2020. 3
- [48] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: carrying location information in full frames into human pose and shape estimation. In *Proc. Eur. Conf. Comput. Vis.*, pages 590–606. Springer-Verlag, 2022. 2, 3, 5
- [49] Zhe Li, Tao Yu, Zerong Zheng, and Yebin Liu. Robust and accurate 3d self-portraits in seconds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7854–7870, 2022. 3
- [50] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022*, pages 39:1–39:9. ACM, 2022. 2, 3
- [51] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 14
- [52] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8762–8771. Computer Vision Foundation / IEEE, 2021. 3
- [53] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Fast generalizable gaussian splatting reconstruction from multi-view stereo. *arXiv preprint arXiv:2405.12218*, 2024. 7
- [54] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061*, 2023. 2
- [55] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 3, 6
- [56] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *Proc. Int. Conf. 3D Vis.*, pages 506–516. Computer Vision Foundation / IEEE, 2017. 1, 3
- [57] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [58] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands,
- face, and body from a single image. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 3, 6, 15
- [59] 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. 1, 3, 7
- [60] 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 *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 9054–9063. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5, 7, 8
- [61] Sida Peng, Zhen Xu, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable implicit neural representations for creating realistic avatars from videos. *TPAMI*, 2024. 1, 3, 7
- [62] Pavlin G. Policar and Blaz Zupan. Visualizing high-dimensional temporal data using direction-aware t-sne. *CoRR*, abs/2403.19040, 2024. 5
- [63] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016. 2, 3, 4, 14
- [64] Renderpeople. Renderpeople, 2018. <https://renderpeople.com/3d-people>. 1, 2, 3, 7, 8
- [65] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *CoRR*, abs/2201.02610, 2022. 3
- [66] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2304–2314. Computer Vision Foundation / IEEE, 2019. 3
- [67] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 81–90. Computer Vision Foundation / IEEE, 2020. 3
- [68] Umme Sara, Morium Akter, and Mohammad Sharif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, page 8–18, 2019. 6
- [69] M. Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab. 2005. 15
- [70] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, 2022. 2, 3
- [71] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 2

- [72] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 57:1–57:10. ACM, 2022. 2
- [73] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deep-cloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. 1, 2, 3
- [74] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 5693–5703. Computer Vision Foundation / IEEE, 2019. 4, 14
- [75] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 11159–11168. Computer Vision Foundation / IEEE, 2021. 3
- [76] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8856–8866. IEEE, 2023. 5
- [77] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul P. Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Nießner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. Advances in neural rendering. *Comput. Graph. Forum*, 41(2):703–735, 2022. 1
- [78] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 501–510, 2019. 2, 3
- [79] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proc. Eur. Conf. Comput. Vis.*, pages 614–631. Springer-Verlag, 2018. 1, 2, 3, 14
- [80] Jiong Wang, Fengyu Yang, Wenbo Gou, Bingliang Li, Danqi Yan, Ailing Zeng, Yijun Gao, Junle Wang, and Ruimao Zhang. Freeman: Towards benchmarking 3d human pose estimation in the wild. *CoRR*, abs/2309.05073, 2023. 2
- [81] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 2
- [82] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20213–20224. IEEE, 2023. 2
- [83] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6
- [84] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humanerf: Free-viewpoint rendering of moving people from monocular video. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 16189–16199. Computer Vision Foundation / IEEE, 2022. 1, 3, 7
- [85] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1679–1688. Computer Vision Foundation / IEEE, 2020. 1, 2, 3, 7
- [86] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 3, 4, 14
- [87] 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 *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 6183–6192. Computer Vision Foundation / IEEE, 2020. 3
- [88] Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Relightable and animatable neural avatar from sparse-view video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 16-22, 2024*, pages 990–1000. IEEE, 2024. 2
- [89] Youze Xue, Jiansheng Chen, Yudong Zhang, Cheng Yu, Huimin Ma, and Hongbing Ma. 3d human mesh reconstruction by learning to sample joint adaptive tokens for transformers. In *Proc. ACM Int. Conf. Multimedia*, pages 6765–6773. ACM Press, 2022. 4, 15
- [90] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2878–2890, 2013. 4
- [91] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13284–13293. Computer Vision Foundation / IEEE, 2021. 3
- [92] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 4212–4222. IEEE, 2023. 4, 14

- [93] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, June 17-24, 2023*, pages 17016–17027. IEEE, 2023. [1](#), [2](#), [3](#)
- [94] Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions and benchmark challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):623–640, 2023. [2](#)
- [95] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. [1](#), [2](#), [3](#)
- [96] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 11426–11436. Computer Vision Foundation / IEEE, 2021. [2](#), [3](#)
- [97] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. [7](#)
- [98] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *IEEE International Conference on Computer Vision*, 2021. [2](#), [3](#)
- [99] Xiaoyun Zheng, Liwei Liao, Xufeng Li, Jianbo Jiao, Rongjie Wang, Feng Gao, Shiqi Wang, and Ronggang Wang. Pkudymvhumans: A multi-view video benchmark for high-fidelity dynamic human modeling. *CoRR*, abs/2403.16080, 2024. [2](#), [3](#)
- [100] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7738–7748. IEEE, 2019. [1](#), [2](#), [3](#), [5](#), [7](#), [8](#)
- [101] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [2](#), [3](#)
- [102] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, June 17-24, 2023*, pages 4315–4327. IEEE, 2023. [2](#), [3](#)

A. Details in Data Collection

Downloading Video Candidates. Our first goal is to collect videos from the web with human motions. To cover a wide range of in-the-wild human-central activities, we start from a label pool of human motion datasets [51]. Based on these human motion labels (*e.g.*, fishing and playing tennis), we download over 100k+ video candidates from YouTube API. **Per-filtering Video Clips.** Some collected video candidates could not meet the high-quality human avatar creation requirement. For example, human bodies may not exist at all (*e.g.*, blank preview, severe occlusion) or frequently change across scenes (*e.g.*, montage) in some subsections of these videos. To exclude such unqualified subsections, we utilize SceneDetect [1] to cut video candidates into clips and eliminate those with insufficient length (less than 2 seconds). Subsequently, we apply human detection models with low FPS to these clips to efficiently filter out those without human subjects at minimal cost. After filtering video candidates, we obtain 460k+ video clip candidates for further processing.

B. Details in Data Pipeline

In addition to the main paper, we provide more details on the data processing pipeline and filtering protocols.

Stage I: Human Bounding Box Detection and Tracking. We first obtain the bounding box of human subjects with off-the-shelf state-of-the-art detection methods (*e.g.*, Yolo [63] and Detectron2 [86]). We only keep the video clip with at least one “person” instance with its detection threshold over 0.8 on all detection models. The tracking step is finding the largest IOU overlay of bounding boxes among frames. We discard low-resolution human subjects whose bounding box areas are lower than 64×64 . To ensure the richness of the dataset, we only keep one “key subject” for each video, as clips from the same video may probably share the same key subject.

Stage II: Human Segmentation Mask Extraction. We first obtain the 2D keypoints J_{2D} for human subjects using the popular HRNet [74] and DWPose [92]. Given the 2D keypoint annotations, we can also discard over part-occluded subjects. In particular, we only keep the subject with the average confidence of 2D keypoints over 0.65. For segmentation, we feed the 2D bounding box and the 2D keypoints into the *sam_vit_h* sub-model to extract the foreground mask.

Stage III: Coarse SMPL and Camera Estimation. We first estimate SMPL and camera parameters frame by frame, using state-of-the-art single-image-based human pose and shape (HPS) estimation methods [23, 43]. To perform better in complex scenes in the wild, we adapt the model pre-trained on the in-the-wild 3DPW dataset [79]. The HPS models infer human body pose/shape parameters (θ/β) and the global camera parameters (rotation matrix R and the 3D offset T). To retain the remaining video clips with consid-

erable viewpoint shifts and human movements, we discard the clips with viewpoint angle changes lower than $\frac{\pi}{4} rad$. We also automatically select the most non-trivial $N = 20$ frames, which keeps the pose and viewpoint diversity to the greatest extent possible. As mentioned in the main paper, we double-check the consistency of the SAM and SMPL masks. Intuitively, the SMPL mask denotes the naked body, while the SAM mask contains the clothed body. Therefore, the SMPL mask should be mostly covered by the SAM mask (See Fig. F (a) ~ (d)). We discard the subjects whose SAM masks are over $3\times$ larger than their SMPL masks (See Fig. F (e) ~ (h)). We also discard the subjects whose 10% SMPL mask pixels from main bodies are not covered by the SAM mask (See Fig. F (i) ~ (l)). Similarly, we double-check the consistency of the 2D keypoints from 2D pose and SMPL estimations and discard the clips with the averaged PCK less than 0.85.

Stage IV: Refining SMPL and Camera In-the-loop. We refine the coarse SMPL parameters (θ, β) and camera parameters (R, T) obtained in Stage I for high-quality annotations. To achieve temporally smooth results, we regularize the differences in parameters between adjacent frames, which is given by

$$\begin{aligned}\mathcal{L}_\theta^s &= \sum_{i=1}^{N-1} \|\theta^i - \theta^{i+1}\|_2, \\ \mathcal{L}_R^s &= \sum_{i=1}^{N-1} \|R^i - R^{i+1}\|_2, \\ \mathcal{L}_T^s &= \sum_{i=1}^{N-1} \|T^i - T^{i+1}\|_2, \\ \mathcal{L}_{2D}^s &= \sum_{i=1}^{N-1} \|\Pi(J_{3D}(\theta^i, \beta^i); R^i, T^i) \\ &\quad - \Pi(J_{3D}(\theta^{i+1}, \beta^{i+1}); R^{i+1}, T^{i+1})\|_2,\end{aligned}\tag{1}$$

where $J_{3D}(\theta, \beta)$ infers the 3D keypoints of the human body, and the Π denotes the 2D projection, and i denotes the i_{th} frame of the input video. Notice that the body shapes (β) are treated as constants across the input video.

In addition, we align the human body parameters to the 2D keypoints and SAM masks, which are given by

$$\begin{aligned}\mathcal{L}_{2D} &= \sum_{i=1}^N \|\Pi(J_{3D}(\theta^i, \beta^i)) - J_{2D}^i\|_2, \\ \mathcal{L}_{\text{mask}} &= \sum_{i=1}^N (M_{rd}(\theta^i, \beta^i), M^i),\end{aligned}\tag{2}$$

where $M_{rd}(\theta, \beta)$ denotes the rendered mask related to the SMPL parameters. The M^i denotes the foreground of the i frame. We also regularize θ to avoid out-of-domain poses [7],

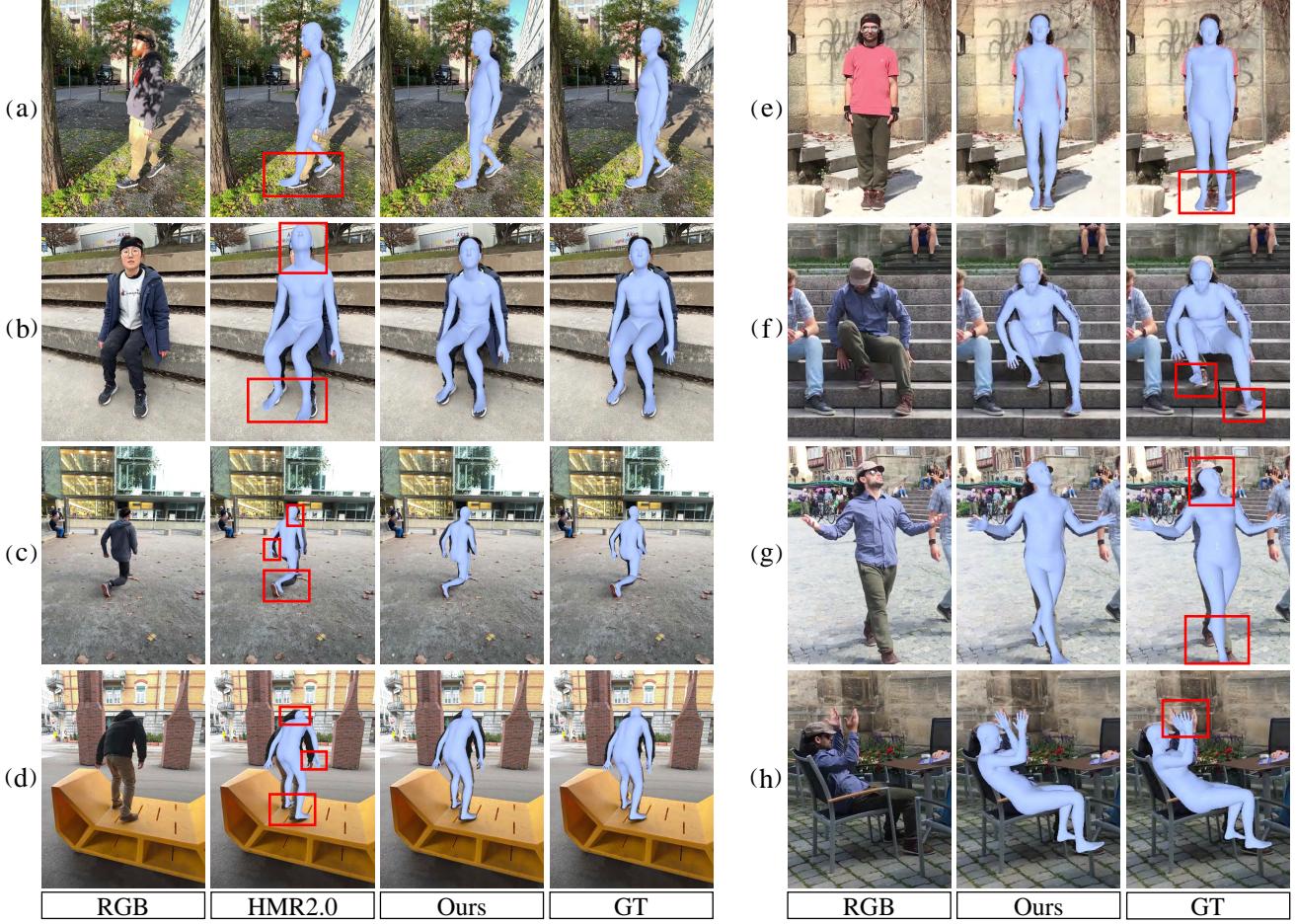


Figure A. Qualitative comparison of our four-stage pipeline and the state-of-the-art HMR2.0 [23]. Our pipeline can adapt to complex environmental scenes and output reasonable results in uncommon scenarios.

using the Gaussian Mixture Model (GMM) prior [58]

$$\mathcal{L}_{\text{prior}} = \sum_{i=1}^N \|GMM(\theta^i)\|_2. \quad (3)$$

We adopt the loss functions as mentioned above for supervision:

$$\begin{aligned} \mathcal{L} = & \lambda_\theta^s \mathcal{L}_\theta^s + \lambda_R^s \mathcal{L}_R^s + \lambda_T^s \mathcal{L}_T^s + \lambda_{2D}^s \mathcal{L}_{2D}^s \\ & + \lambda_{2D} \mathcal{L}_{2D} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}. \end{aligned} \quad (4)$$

We empirically set the loss weights as $\lambda_\theta^s = 100$, $\lambda_R^s = 1000$, $\lambda_T^s = 50$, $\lambda_{2D}^s = 100$, $\lambda_{2D} = 100$, $\lambda_{\text{mask}} = 100$ and $\lambda_{\text{prior}} = 0.1$. We adopt the LBFGS [69] optimizer with the learning rate $lr = 1.0$. Qualitative comparisons between coarse and refined SMPL and camera parameters can be found in Fig. E. Stage IV boosts local alignments on extremities.

Efficiency. Data collection and annotation efficiency are also crucial for data scale-up and application. Despite the

complex multiple-stage design, our data-collecting pipeline only takes less than 200 seconds to generate full annotations for a 20 seconds in-the-wild video clip.

C. Details in Pipeline Evaluation

Dataset. We evaluate our four-stage pipeline on the in-the-wild EMDB [40], which is widely recognized for its challenging and diverse real-world scenarios. We use the EMDB-1 split to evaluate the camera-coordinate performance. EMDB-1 contains 17 sequences totaling 13.5 minutes.

Metrics. For joints, we compute error on the 24 main joints of the human body under the SMPL convention. As for vertex, we calculate the point-to-point corresponding error on the SMPL vertices. We report quantitative results on MPJPE, PA-MPJPE [33], and PVE [89]. **MPJPE (Mean Per Joint Position Error)** calculates the mean distances between the predicted and ground-truth 3D joints after the translation alignment at the pelvis joint. The predicted or



Figure B. Visualization of SMPL overlay on unusual camera viewpoints. Our pipeline can show robust annotations on unusual camera viewpoints and body poses.

ground-truth 3D joints are regressed from corresponding pose and shape parameters. **PA-MPJPE (Procrustes analysis MPJPE)** calculates the mean distances between the predicted and ground-truth 3D joints after Procrustes Analysis [24], including alignment in scale, translation and rotation. PA-MPJPE mainly focuses on the quality of pose and shape estimation, regardless of global rotation. **PVE (Per Vertex Error)** calculates the mean distances between the vertices on the human mesh without any alignment, which evaluates the reconstruction accuracy of the human surface. **Qualitative Comparisons.** Qualitative comparisons on the EMDB dataset are also shown in Fig. A (a) ~ (d). Compared to previous state-of-the-art HMR2.0 [23], our pipeline can adapt to complex environmental scenes and output reasonable results in uncommon scenarios. For example, our pipeline can accurately predict 1) the foot and ankle pose of Fig. A (a) & (c); 2) the head and neck pose of Fig. A (b) & (d); 3) the global alignment of Fig. A (a), (b) & (d). Qualitative comparisons on the 3DPW dataset are also shown in Fig. A (e) ~ (h). Compared to the official ground truth

(GT) from expensive IMUs, our annotations also exhibit better alignment on 1) the foot and ankle pose of Fig. A (e) ~ (g); 2) the head and neck pose of Fig. A (g); 3) the hand pose of Fig. A (h). These comparisons validate the annotation quality of our pipeline for in-the-wild videos.

D. License, Statistics and Visualizations

The authors bear all responsibility in case of violation of rights and confirm that this dataset is open-sourced under the **S-Lab License 1.0 license**. We shall enforce strict regulations when applying our code and data to mitigate potential negative social impacts. 69.1% / 30.9% scenes in WildAvatar are indoor/outdoor, respectively. 45.3% / 54.7% of the scenes in WildAvatar have single/multiple human(s), respectively. And 34.6% / 65.4% subjects in WildAvatar are male/female, respectively. More visualization of SMPL overlay on unusual camera viewpoints can be found in Fig. B. More RGB examples from the proposed WildAvatar dataset can be found in Fig. C, and examples of different SSIOU ranges can be found in Fig. D.

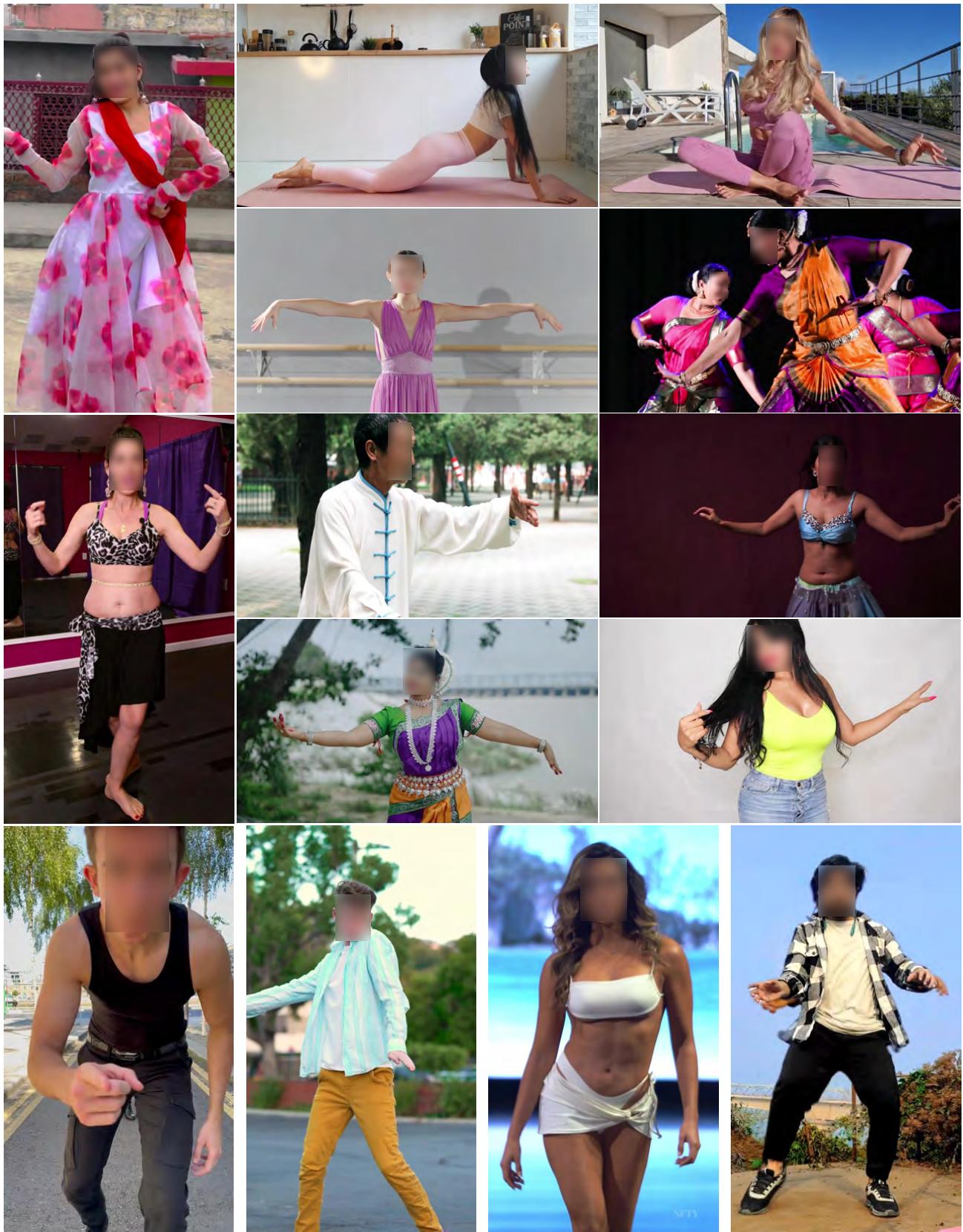


Figure C. More RGB examples from the proposed WildAvatar dataset. The best view zoomed in on-screen for details.



Figure D. Examples of different SSIOU ranges. SSIOU rises from up to down. The last row shows the SSIOU larger than 1.9. The samples with SSIOU larger than 1.9 are mostly caused by loose dresses rather than erroneous SMPL fitting.

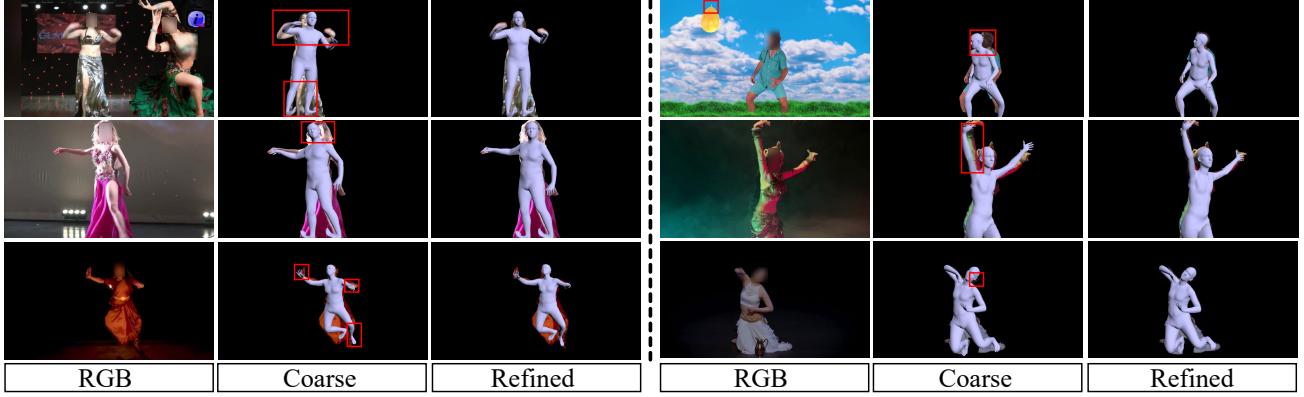


Figure E. Comparison of the coarse and refined SMPL parameters. The coarse SMPL annotations are from Stage III, and are later refined in Stage IV. The refined SMPL parameters achieve better alignments to the raw RGB images.

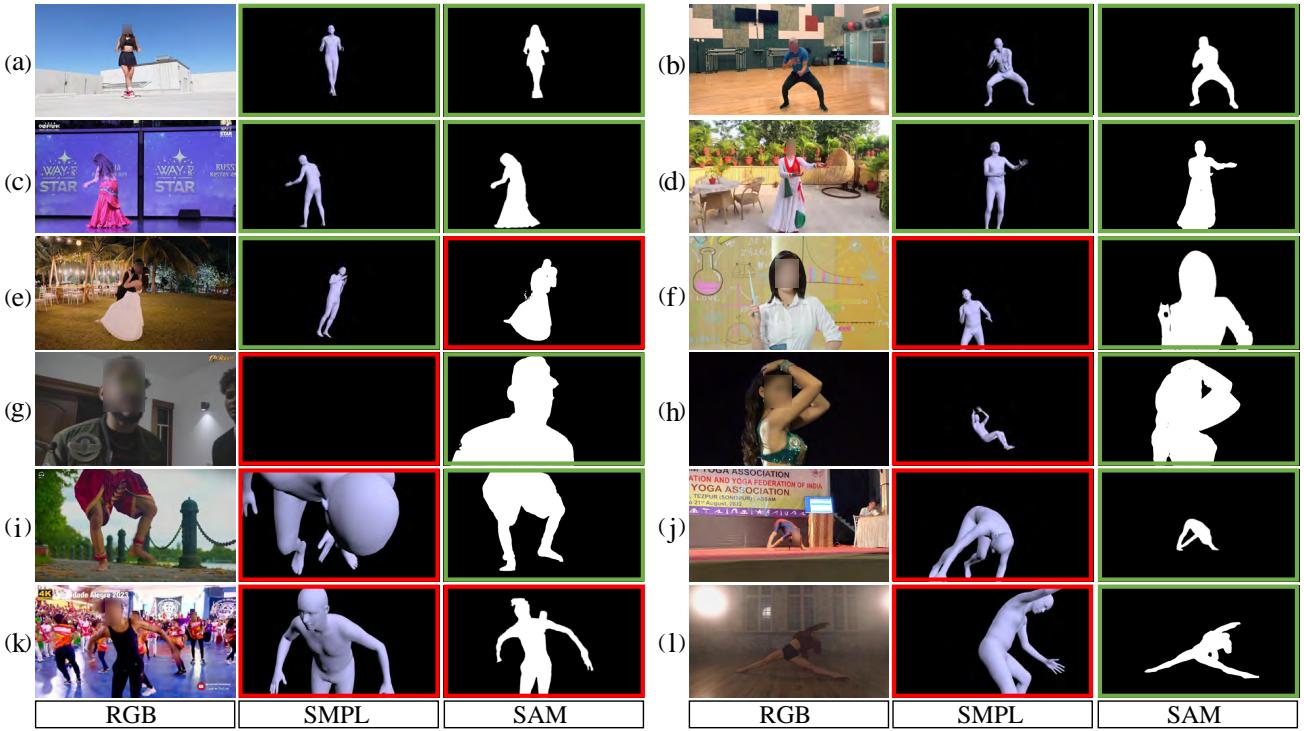


Figure F. SMPL and SAM consistency. The green/red borders denote good/bad outputs, respectively. Note that the SMPL annotations are coarse results from Stage III, which are later refined in Stage IV.