GAS: Generative Avatar Synthesis from a Single Image

Yixing Lu[1] Junting Dong[2†] Youngjoong Kwon[3] Qin Zhao[2]
Bo Dai[2] Fernando De la Torre[1]
[1]Carnegie Mellon University [2]Shanghai AI Laboratory [3]Stanford University
https://humansensinglab.github.io/GAS/

Figure 1. **In-the-wild avatar synthesis across views and poses.** Starting from a reference image, we generate its novel views and animate the avatar given a pose sequence.

Abstract

We present a unified and generalizable framework for synthesizing view-consistent and temporally coherent avatars from a single image, addressing the challenging task of single-image avatar generation. Existing diffusion-based methods often condition on sparse human templates (e.g., depth or normal maps), which leads to multi-view and temporal inconsistencies due to the mismatch between these signals and the true appearance of the subject. Our approach bridges this gap by combining the reconstruction power of regression-based 3D human reconstruction with the generative capabilities of a diffusion model. In a first step, an initial 3D reconstructed human through a generalized NeRF provides comprehensive conditioning, ensuring high-quality synthesis faithful to the reference appearance and structure. Subsequently, the derived geometry and appearance from the generalized NeRF serve as input to a video-based diffusion model. This strategic integration is pivotal for enforcing both multi-view and temporal consistency throughout the avatar's generation. Empirical results underscore the superior generalization ability of our proposed method, demonstrating its effectiveness across diverse in-domain and out-of-domain in-the-wild datasets.

1. Introduction

Human avatar generation has been a longstanding focus in computer vision and graphics, with wide-ranging applications in gaming, film, sports, fashion, and telepresence. However, despite its transformative potential, existing technologies often rely on expensive capture systems and complex workflows, making them inaccessible to the broader public. Recent advancements have aimed to make avatar generation more accessible by leveraging neural rendering techniques. In particular, studies such as [17, 22–24] explore generalizable 3D human reconstruction, enabling novel view and arbitrary pose synthesis of subjects from highly sparse—sometimes even single—input images. These approaches incorporate 3D human priors to support accurate synthesis of complex body geometry and smooth interpolation between input observations. However, their regression-based nature limits their extrapolation capabilities, often resulting in blurry outputs and restricting them to mostly rigid deformations.

Generative models, such as GANs and diffusion models, have recently demonstrated impressive capabilities in synthesizing photorealistic images and videos with realistic motion. Leveraging these advancements, recent ap-

† Corresponding author

proaches employ diffusion models conditioned on human priors—such as 2D keypoints, depth, or normal maps—to generate high-quality avatars from a single image [6, 16, 43, 64]. However, the sparsity of these conditioning cues often leads to inconsistencies in the generated results, including flickering across views and temporal instability.

To address these challenges, we introduce a method for single-image avatar synthesis that ensures both view and temporal consistency. Rather than depending solely on sparse conditioning signals (e.g., SMPL normal maps), we first generate intermediate novel views and/or poses using a regression-based 3D human reconstruction model. The shape and apperance of these new views/poses are then used as inputs to a video diffusion model. By leveraging the dense structural information from the 3D reconstruction, our approach maintains geometric fidelity and visual richness, delivering high-quality, consistent results across viewpoints and human poses (i.e., time).

Another major challenge lies in improving generalization to real-world data, as the ultimate goal is to develop a system that performs well beyond controlled environments—specifically on casually captured, in-the-wild human imagery. Most existing human digitization models are trained on multi-view datasets collected in studio settings [17, 22, 23], which lack the diversity found in real-world scenes, including variations in lighting, clothing, and movement. To overcome this limitation, we propose leveraging in-the-wild internet videos, which offer a rich and diverse distribution of real-world appearances. However, training with such data introduces new challenges, particularly for novel view synthesis, which traditionally relies on multi-view supervision—an element typically absent in internet-sourced footage.

Our method, Generative Avatar Synthesis (GAS) is a unified framework that jointly learns both novel view and pose synthesis by sharing parameters across tasks. While studio-captured multi-view datasets are used for view synthesis, we augment training with both multi-view and in-the-wild videos for pose synthesis. This parameter sharing allows improvements from pose generalization to transfer naturally to view synthesis, significantly enhancing performance on real-world data (see Figure 1). To enable this joint learning effectively, we incorporate a mode switcher that distinguishes between the two tasks. This design allows the network to prioritize view consistency for novel view synthesis and realistic deformation for novel pose generation.

In summary, our main contributions are:

- A unified framework for novel view and pose synthesis of avatars, which enables shared model parameters across both tasks with real human data (e.g., internet videos) at scale for training, leading to broad generalizability.
- A dense appearance cue derived from generalizable NeRF renderings as diffusion guidance, ensuring consistent appearance preservation over novel views and poses.

2. Related Work

2.1. Generalizable Human Radiance Fields

Radiance field methods like NeRF [34] have demonstrated impressive performance in generating high-fidelity novel views. To extend these models to generalize across scenes and work with sparse-view inputs, several approaches [42, 47, 56] incorporate pixel-aligned features. However, applying them directly to human modeling remains challenging due to complex body geometry and self-occlusions, often resulting in over-smoothed outputs. To address this, recent works [8, 9, 22, 23, 36, 37, 60] leverage 3D human priors such as the SMPL model [33] to better anchor features on the human form, enabling robust synthesis from sparse or even single-view inputs [17]. However, these methods typically suffer from slow rendering speeds.

To overcome this, 3D Gaussian Splatting [21], accelerated by GPU rasterization, has emerged as an efficient radiance field representation. Recent works [24, 63, 65] leverage this framework for photorealistic human rendering from sparse inputs. Yet, they still face difficulties in generating fine details in unseen regions due to the mean-seeking nature of one-to-many mappings [25, 31]. Moreover, animating avatars typically requires costly post-processing—e.g., using linear blend skinning [26]—which often fails to capture non-rigid effects such as garment dynamics [17, 18, 22, 38, 65].

In this work, we propose a novel framework that learns a distribution shift over generalizable human radiance fields, enabling consistent and artifact-free synthesis of novel views and poses. By incorporating a strong generative prior, our method naturally produces realistic non-rigid deformations from pose sequences without relying on complex post-processing pipelines.

2.2. Generative Human Animation

Generative human animation aims to employ generative models to produce coherent videos from static human images, utilizing guidance such as text and motion sequences. This body of work focuses on leveraging generative priors to sample complex dynamic motions, including pose-dependent clothing deformations. Early approaches harnessed the generative capabilities of Generative Adversarial Networks (GANs) [11] for synthesizing novel human poses [5, 30, 49]. In recent years, latent diffusion models [41] have gained traction in the realm of human animation due to their robust controllability and superior generation quality. Various methods [6, 7, 16, 27, 28, 44, 48, 64] implement distinct motion guidance and conditioning techniques. Notably, Animate Anyone [16] introduces a UNet-based ReferenceNet to extract features from reference images, utilizing

DWPose [54] for pose guidance. Subsequent works [64] also incorporate guidance from 3D human parametric models, such as SMPL [33, 61, 62], leveraging the advantages of multiple forms of guidance. Following this trajectory, recent studies [14, 20, 29, 32, 38, 43, 53] explore human view controllability within the diffusion frameworks. Human4DiT [43] develops a hierarchical 4D diffusion transformer that disentangles the learning of 2D images, viewpoints, and time. However, these methods struggle to synthesize view-consistent and temporally coherent results due to the gap between the sparse driving signal and the actual subject. In this paper, we address this challenge by densely conditioning on generalizable geometry and appearance cues, leading to improved appearance preservation and consistency.

2.3. Diffusion Models for Video Generation

Diffusion models have shown impressive results in image synthesis and are now being adapted to the more complex domain of video generation [58]. A common strategy involves extending UNet-based image diffusion models by adding temporal modules. Some approaches [3, 12] freeze the pre-trained image backbone and train only the temporal components on video data. Stable Video Diffusion (SVD) [2], for instance, enhances the UNet architecture by inserting temporal layers after each spatial convolution and attention block, and fine-tunes the full model on large-scale curated video datasets. This approach has proven to be a powerful foundation, enabling downstream applications such as video generation and 3D/4D synthesis [46, 51]. In parallel, diffusion transformers [4, 35, 55] have emerged as an alternative architecture, applying full spatio-temporal attention over latent codes for both images and videos. In this work, we build on the strong capabilities of SVD to model multi-view consistency and temporally coherent, realistic deformations of human subjects.

3. Method

Figure 2 illustrates the main idea of our method. GAS synthesizes view- and pose-consistent avatar renderings from a single image by combining the rich appearance information from a generalizable 3D human reconstruction model with the generative power of a video diffusion model.

3.1. Notation

Functions (e.g., neural network mapping) are denoted with uppercase calligraphic letters (e.g., $\mathcal{U}$). Vectors are denoted with bold lowercase letters (e.g., $\boldsymbol{x}$). Matrices are denoted with uppercase letters (e.g., C). Sets are denoted with bold uppercase letters (e.g., $\boldsymbol{I}_{nerf}$).

3.2. Single-view Generalizable Human Synthesis

Given a single reference image I_{ref}, the camera parameter P_{ref}, and the parameters of a human template, *i.e.*, the SMPL [33] model, we adopt single-view generalizable human NeRF [17] to synthesize an image corresponding to the target camera parameters P_{tar} and target template parameters consists of pose $\boldsymbol{\theta}_{tar}$ and shape $\boldsymbol{\beta}_{tar}$.

To render the target image, a point $\boldsymbol{x}$ is sampled along the cast rays in the target space. Then it is transformed to point $\boldsymbol{x_c}$ in the the SMPL canonical space via inverse LBS, where the features associated with $\boldsymbol{x_c}$ are queried from the observation space. Specifically, pixel-aligned features and human template-conditioned features are obtained and fused together, denoted as $\boldsymbol{p}$. The density $\boldsymbol{\sigma}$ and color $\boldsymbol{c}$ are obtained by a multi-layer perception (MLP) network $\mathcal{F}$:

$$\boldsymbol{\sigma}(\boldsymbol{x}), \boldsymbol{c}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{x}, \boldsymbol{p}, \gamma_d(\boldsymbol{d})), \tag{1}$$

where γ_d is the positional encoding of viewing direction $\boldsymbol{d}$.

The target image I_{tar} , corresponding to the desired view and pose, is generated using volume rendering as described in [34]. We emphasize that this human NeRF model is designed to generalize, enabling the synthesis of human appearances across a range of novel views and poses. However, due to its mean-seeking property, producing sharp renderings, particularly in occluded or unseen regions, remains challenging. This limitation motivates our introduction of generative priors to learn a distribution over the human views and poses, as discussed in the next section.

3.3. Synthesis as Video Generation Condition

This section describes how we leverage a video diffusion prior to reformulate novel view and pose synthesis as a unified video generation task, conditioned on human radiance fields and geometric templates. While previous methods typically treat view and pose synthesis as separate problems, our approach bridges them by conditioning the diffusion model on NeRF-rendered images. This dense and appearance-rich signal leads to more consistent outputs. In contrast, conditioning solely on sparse human templates, as in prior work [64], often results in inconsistent appearance across views.

Novel View Synthesis. Given a reference image I_{ref} and a camera trajectory $\{P_1, P_2, \cdots, P_T\}$, we render the corresponding images $\boldsymbol{I}_{nerf} = \{I_1, I_2, \cdots, I_T\}$ from the NeRF, which servers as an input to our video diffusion model, Stable Video Diffusion (SVD) [2]. SVD has a spatio-temporal attention module and 3D residual convolution in the diffusion UNet. For the single input image I_{ref}, we extract its feature with CLIP [39] and repeat it for T times, denoted as $\boldsymbol{h}_{\text{clip}}$, which is then added to the video diffusion model through the cross attention. Meanwhile, we use the VAE encoder to encode our input image I_{ref} and obtain its latent feature C_{vae}.

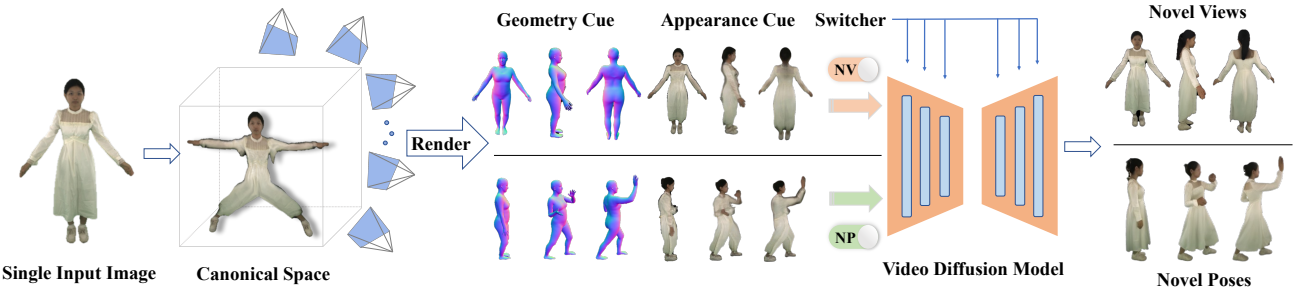

Figure 2. **Overview of GAS.** Starting from a single input image, GAS uses a generalizable human NeRF to map the subject into a canonical space, then reposes and renders the 3D NeRF model to extract detailed appearance cues (i.e., NeRF renderings). These are paired with geometry cues (i.e., SMPL normal maps) and fed into a video diffusion model. A switcher module disentangles the tasks, enabling the model to generate either multi-view consistent novel views or temporally coherent pose animations.

To introduce NeRF renderings $\boldsymbol{I}_{nerf}$ to our video diffusion model, we feed $\boldsymbol{I}_{nerf}$ to the VAE encoder, after which it is further encoded by a small convolutional neural network. We denote the output latent feature as C_{nerf}.

In practice, we observe that solely relying on NeRF renderings in our video diffusion model is insufficient, as these renderings may sometimes exhibit artifacts, particularly due to inaccurate SMPL fittings or occlusions. Such artifacts can corrupt the guidance and hinder the diffusion model from learning meaningful conditional distributions. To address this limitation, we further integrate a geometric cue, *i.e.*, the SMPL model, to provide additional structural guidance. The SMPL model captures essential geometry information, which leads to the enhanced spatial consistency as well as the robust human shape recovery.

In order to integrate this information with the NeRF rendered features in the 2D pixel space, similar to [64], we render the human template mesh into 2D normal maps. Specifically, we render the SMPL normal maps $\boldsymbol{M} = \{M_1, M_2, \cdots, M_T\}$ under the camera trajectory $\{P_1, P_2, \cdots, P_T\}$. Then a set of 2D convolution layers are utilized to extract the features, denoted by C_{smpl}.

To effectively fuse C_{nerf} and C_{smpl}, we combine them through element-wise addition. The resulting fused feature is then added to the output of the first convolutional layer of the UNet in the video diffusion model.

Novel Pose Synthesis. Given a sequence of SMPL poses $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_T\}$ and a fixed camera parameter, we render corresponding images from the NeRF and SMPL normal maps. They can be used as conditions to the video diffusion model in the same manner as illustrated above.

Now we can formulate the learning objective of our diffusion model. The diffusion UNet $\mathcal{U}_\Theta$ predicts the noise ϵ for each diffusion step t, and our training objective is

$$\mathcal{L}_{\mathcal{U}_\theta} = \mathbb{E}\left[\|\epsilon - \mathcal{U}_\theta(Z_t, t, \boldsymbol{h}_{\text{clip}}, C_{\text{vae}}, C_{\text{nerf}}, C_{\text{smpl}})\|\right] \quad (2)$$

where $Z_t = \alpha_t Z + \sigma_t \epsilon$. Here, Z is the ground-truth latent, $\epsilon \sim \mathcal{N}(0, I)$, and α_t and σ_t define the noise at timestep t. Θ is the learnable parameters of the UNet $\mathcal{U}$.

3.4. Switcher for Disentangled Synthesis

Joint learning of human view synthesis and pose synthesis presents inherent challenges for feed-forward methods. Our proposed framework addresses these challenges by unifying both tasks into a single video generation task, leveraging the capability of video diffusion model to effectively model each task under the same representation. Under this framework, a straightforward approach would be to train the video diffusion model on both multi-view and dynamic video data simultaneously. However, our empirical findings reveal that dynamic motions embedded within view synthesis videos can disrupt view consistency (see Figure 9). This issue arises due to the inherent modality differences between static view synthesis videos and dynamic animation videos. To mitigate this problem, we introduce a switcher mechanism within our video diffusion model that explicitly controls and disentangles these two modes, ensuring task-specific consistency and performance.

In particular, we introduce the switcher $\boldsymbol{s}$, which is a one-hot vector that labels each of the two modalities, as the additional condition to the video diffusion model. This allows us to extend the formulation in Equation 2 as follows:

$$\mathcal{L}_{\mathcal{U}_\theta} = \mathbb{E}\left[\|\epsilon - \mathcal{U}_\theta(Z_t, t, \boldsymbol{h}_{\text{clip}}, C_{\text{vae}}, C_{\text{nerf}}, C_{\text{smpl}}, \boldsymbol{s})\|\right] \quad (3)$$

To incorporate the domain switcher $\boldsymbol{s}$, we first apply positional encoding and then concatenate it with the time embedding. This combined encoding is subsequently fed into the UNet within the video diffusion model.

3.5. Training and Inference

Training. The entire training process of our pipeline consists of two stages. During the first stage, we train the generalizable human NeRF $\mathcal{F}$ model on the multi-view datasets. Following [17], we randomly sample observation and target image pairs from each subject. The prediction of the target view is supervised by minimizing the loss objective $\mathcal{L} = \mathcal{L}_2 + \lambda_{\text{ssim}} \cdot \mathcal{L}_{\text{ssim}} + \lambda_{\text{lpips}} \cdot \mathcal{L}_{\text{lpips}} + \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}}$ where $\mathcal{L}_2, \mathcal{L}_{\text{ssim}}, \mathcal{L}_{\text{lpips}}$ are photometric L_2 loss, SSIM loss [50]

and LPIPS [59] loss between the predicted image and the ground truth. $\mathcal{L}_{\text{mask}}$ is the L_2 difference between the accumulated volume density and the ground truth human binary mask. $\lambda_{\text{ssim}}, \lambda_{\text{lpips}}, \lambda_{\text{mask}}$ are weights of each loss to balance their contributions to the final loss function.

In the second stage, we freeze the generalizable NeRF model and train our video diffusion model. We train the full spatio-temporal UNet and feature encoders for NeRF and SMPL normal renderings following our training objective in Equation 3. The second stage is trained on our complete dataset to ensure the generalization.

Inference. We apply classifier-free guidance (CFG) [15] to inference from the video diffusion model. The two tasks are done with different CFG schedules according to the task properties. For the novel view synthesis task, we utilize a triangular CFG scaling [46], where we linearly increase CFG from 1 at the front view to 2 at the back view, then linearly decrease it back to 1 at the front view. For the novel pose synthesis task, we fix the CFG scale to be 2.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

3D Scans. We use THuman2.1 [57] and 2K2K dataset [13] with around 4500 3D scans in total. THuman2.1 comprises 2445 high-quality 3D scans and texture maps. We randomly sample 2345 subjects for training and the remaining 100 subjects for testing. 2K2K dataset consists of 2000 train and 50 test subjects, totally 2050 scans. For both datasets, RGB images are rendered from 20 uniformly distributed views around the scan, at the resolution of 1024×1024. SMPL parameters are estimated using off-the-shelf method for multi-view SMPL fitting [1].

Multi-view Videos. MVHumanNet [52] is a multi-view video dataset featuring a large number of diverse identities and everyday clothing. We use a subset of 944 human captures, each consisting of synchronized 16-view videos per subject. We reserve 48 subjects for evaluation and use the remaining subjects for training. The dataset also includes SMPL parameters, optimized from multi-view images.

Monocular Videos. For in-the-wild datasets, we use the TikTok dataset [19] and an additional collection of internet videos. The TikTok dataset includes 350 dance videos, from which we processed and filtered 289 valid video sequences for training. Following the protocol in [6, 48], we use subjects from 335 to 340 for testing. To further diversify the data, we selected 122 video sequences from Champ [64] training dataset, originally sourced from reputable online platforms such as TikTok and YouTube. For both datasets, we obtain the foreground human masks using Grounded-SAM [40] and the SMPL parameters using 4DHumans [10].

4.1.2. Implementation Details

We trained the generalizable human NeRF model [17] on MVHumanNet dataset. To accelerate the video diffusion training, instead of creating and rendering the human NeRF on-the-fly, we choose to store the NeRF renderings for all datasets offline. For the video diffusion model, we initialize it with the pre-trained Stable Video Diffusion 1.1 image-to-video model [1] [2]. We resize all images to a resolution of 512×512. Each batch consists of 20 frames. We train the model for 150k iterations with an effective batch size of 8 and a learning rate of 10^{-5}. We utilize 8 A100 GPUs and the total training time is 3 days.

4.1.3. Baselines and Metrics

Baselines. We benchmark our method against state-of-the-art generative human rendering methods including Champ [64] and Animate Anyone [16]. As for Animate Anyone [16], we use the implementations from Moore Threads [2].

Metrics. We evaluate the fidelity and consistency of our results using both image-level and video-level metrics. For image-level comparisons, we report peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [50], and learned perceptual image patch similarity (LPIPS) [59]. For video-level evaluation, we use the Fréchet Video Distance (FVD) [45] metric.

4.2. Comparisons on Novel View Synthesis

We compare our approach with two leading generative human synthesis methods—Animate Anyone [16] and Champ [64]. For a fair evaluation, we fine-tuned both models on our complete 3D scan dataset for 10,000 iterations.

Quantitative results. We employ a rigorous evaluation protocol for single-image human novel view synthesis. For each subject, we sample four orthogonal input views and report the average performance across all corresponding novel views. Notably, Animate Anyone [16] tends to generate noisy backgrounds. To ensure a fair comparison that emphasizes the synthesized human subject, we apply the ground truth masks to remove backgrounds in the THuman dataset. In contrast, for the 2K2K dataset, we evaluate the generated outputs without background removal.

Quantitative results are presented in Table 1. Results show that our method achieves state-of-the-art performance across all evaluation metrics, highlighting the advantages of our generalizable human radiance field in preserving intricate details across viewpoint changes.

Qualitative results. Figure 3 presents our qualitative comparisons with baseline methods, focusing on challenging reference views and novel views with less overlap.

[1] https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt-1-1

[2] https://github.com/MooreThreads/Moore-AnimateAnyone

Method	PSNR ↑		SSIM ↑		LPIPS ↓		FVD ↓	
	THuman	2K2K	THuman	2K2K	THuman	2K2K	THuman	2K2K
Animate Anyone	22.48	18.48	0.927	0.557	0.061	0.263	460.3	1422.1
Champ	20.96	22.14	0.909	0.910	0.074	0.075	470.3	480.3
Animate Anyone*	25.20	26.22	0.938	0.936	0.046	0.050	302.7	286.4
Champ*	23.89	25.66	0.928	0.935	0.054	0.052	296.1	279.3
Ours	**26.77**	**28.82**	**0.943**	**0.954**	**0.041**	**0.039**	**194.8**	**191.3**

Table 1. **Quantitative comparison for novel view synthesis on THuman and 2K2K dataset.** For all the methods, we report the average score on 20 views using four orthogonal input views (front, back, and side views). * indicates methods fine-tuned on our 3D scan dataset.

Figure 3. **Qualitative comparisons for novel view synthesis on the THuman dataset.** For the first subject, our method generates cleaner garment textures and sharper facial details, achieving better realism and consistency across views (e.g., the hair style in our generated front view is faithful to the reference image). For the second subject, baseline methods exhibit inconsistencies across views, marked with circles (e.g., hair misalignment between generated front and back views). * denotes fine-tuning on our full 3D scan dataset.

4.3. Comparisons on Novel Pose Synthesis

We compare GAS on novel pose synthesis task with Animate Anyone [16] and Champ [64]. For a fair comparison, we also further fine-tuned the baseline methods on our complete video dataset for 10k iterations.

Quantitative results. To evaluate on the TikTok dataset, we randomly sample a frame as the reference and generate the subsequent 100 frames. Quantitative results are presented in Table 2. Our method consistently outperforms the baseline methods, even after additional fine-tuning.

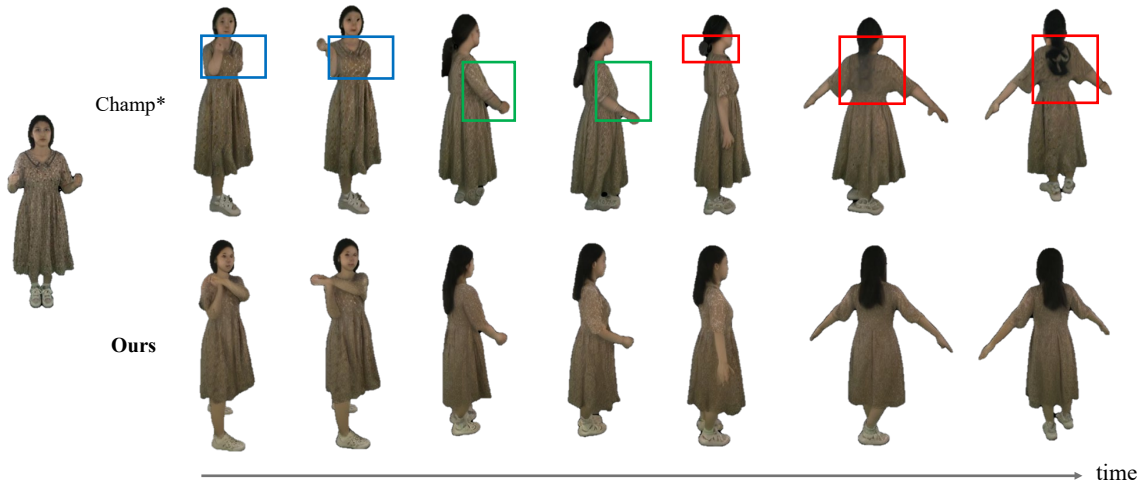

Figure 4. **Qualitative comparisons for novel pose synthesis on MVHumanNet dataset.** In the first row of Champ [64] results, blue rectangles mark disappearing arms, green rectangles show varying sleeve lengths, and red rectangles indicate inconsistencies in hair appearance. * denotes a method further fine-tuned using our complete animation video dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Animate Anyone	17.21	0.762	0.225	1274.1
Champ	18.48	0.806	0.182	585.0
Animate Anyone*	17.83	0.791	0.204	840.5
Champ*	18.57	0.797	0.187	893.7
Ours	**19.11**	**0.833**	**0.176**	**362.0**

Table 2. **Quantitative comparisons for novel pose synthesis on TikTok dataset.** * Indicates methods fine-tuned on our multi-view video and monocular video dataset.

Qualitative results. Qualitative results are shown in Figure 4, where we compare our method with Champ [64] on the novel view animation task for the MVHumanNet dataset. Our method synthesizes temporally consistent animations, even from novel views. It highlights the advantage of leveraging human NeRF's dense and 3D consistent appearance cue for guidance.

4.4. Ablation Studies and Analyses

We perform ablation studies to assess different variants of our proposed method. Additional quantitative and qualitative results can be found in the supplementary material.

Generalizable geometry and appearance cues. To study the effect of both geometry (*i.e.*, SMPL normal map) and appearance cues (*i.e.*, human radiance field), we train a variant with the geometry cue completely removed and a variant with appearance cue removed. We present quantitative comparisons for novel view synthesis on the THuman dataset and novel pose animation on the MVHumanNet dataset. As shown in Table 3, our approach consistently improved performance across both tasks and datasets.

Our hypothesis for the generalizable human NeRF is that it provides rich and consistent appearance cues across different views and times. To validate this, we present an additional qualitative study, as suggested in Figure 5.

Method	PSNR ↑		SSIM ↑		LPIPS ↓		FVD ↓	
	NVS	NPS	NVS	NPS	NVS	NPS	NVS	NPS
w.o. geo. cue	26.07	28.33	0.938	0.943	**0.045**	0.044	227.1	210.6
w.o. appear. cue	26.38	27.66	0.938	0.941	0.041	0.043	207.5	234.6
Ours	**26.77**	**28.74**	**0.943**	**0.945**	0.041	**0.040**	**194.8**	**188.5**

Table 3. **Quantitative ablation study on the geometry and appearance cues.** NVS denotes novel view synthesis and NPS denotes novel pose synthesis. The same notation applies to Table 4.

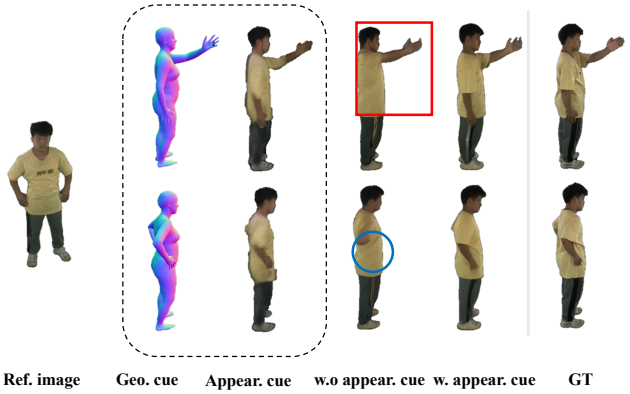

Figure 5. **Ablation study on the appearance cue.** Without the appearance cue, artifacts include incorrect arm raises (red rectangles) and distorted hand placement on the waist (blue circles), both resolved with the appearance cue.

Occlusions, which are frequent in in-the-wild videos, often introduce artifacts in generalizable human NeRF due to

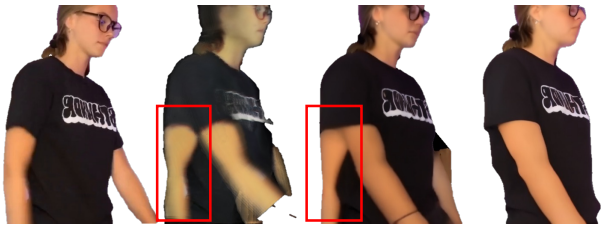

Figure 6. **Ablation study on the geometry cue.** Without the geometry cue, occlusion leads human NeRF to misinterpret the arm as clothing texture, which further misleads diffusion generation (red rectangles).

Figure 7. **Novel view synthesis on real-world human subjects.**

pixel-aligned feature extraction and warping. As illustrated in Figure 6, the absence of clean geometric guidance can mislead the diffusion model, resulting in noticeable distortions.

Video diffusion model. The video diffusion model is suggested to refine the generalizable human NeRF renderings and produce shaper results with fewer artifacts. We ablate the effect of video diffusion model in Table 4 by comparing the generalizable human NeRF renderings and the diffusion refined results. Results show significant improvement across all metrics on both tasks.

	PSNR ↑		SSIM ↑		LPIPS ↓		FVD ↓	
	NVS	NPS	NVS	NPS	NVS	NPS	NVS	NPS
Before diffusion	24.25	18.37	0.925	0.809	0.073	0.233	517.7	1255.8
After diffusion	**28.82**	**19.11**	**0.954**	**0.833**	**0.039**	**0.176**	**191.3**	**362.0**

Table 4. **Quantitative ablation on the video diffusion model.**

Mixed training with 3D/multi-view captures and internet videos. We evaluate the effectiveness of our training strategy that incorporates diverse data sources to enhance generalization. As illustrated in Figure 7, our model exhibits robust performance in novel view synthesis across real-world scenarios. Additionally, Figure 8 shows an ablation study highlighting the contribution of internet video data to improved generalization.

Figure 8. **Ablation on involving internet videos for training.**

Figure 9. **Ablation on the switcher for disentangling static view and dynamic motion synthesis.** Without the switcher, undesired clothing deformations, such as dress swinging, are involved in the novel view generation, corrupting the view consistency.

Effect of switcher for disentangled synthesis. In our unified framework for static view synthesis and dynamic motion modeling, we incorporate a switcher module to explicitly separate the two tasks. To assess its impact, we train a variant without the switcher. As shown in Figure 9, omitting the switcher leads to unintended motion artifacts in sequences of novel views, highlighting its importance for task disentanglement.

5. Conclusion

We propose a unified framework for avatar synthesis from a single image that tackles two major challenges: maintaining appearance consistency and generalizing to in-the-wild scenarios. Our method combines a regression-based human radiance field with a video diffusion model, leveraging dense conditioning to reduce the mismatch between driving signals and target outputs. The framework supports large-scale training on diverse, real-world human data, while cleanly separating the modeling of static novel views and dynamic motion. This combination enables high-fidelity avatar generation with realistic deformations and strong generalization across varied environments.

Acknowledgements

This work was supported by the Shanghai Artificial Intelligence Laboratory. The authors would like to thank Aviral Chharia, Jialu Gao, Jianjin Xu, and Wenbo Gou for their valuable feedback on the manuscript.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 5

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 5

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 3

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[6] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023. 2, 5

[7] Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, et al. X-dyna: Expressive dynamic human image animation. *arXiv preprint arXiv:2501.10021*, 2025. 2

[8] Junting Dong, Qi Fang, Yudong Guo, Sida Peng, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. In *Advances in Neural Information Processing Systems*, 2022. 2

[9] Junting Dong, Qi Fang, Tianshuo Yang, Qing Shuai, Chengyu Qiao, and Sida Peng. ivs-net: Learning human view synthesis from internet videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22942–22951, 2023. 2

[10] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 5

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[13] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)*, 2023. 5

[14] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. *arXiv preprint arXiv:2408.14211*, 2024. 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[16] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2, 5, 6

[17] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9352–9364, 2023. 1, 2, 3, 4, 5

[18] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2

[19] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 5

[20] Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. *arXiv preprint arXiv:2502.07785*, 2025. 3

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[22] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 1, 2

[23] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural image-based avatars: Generalizable radiance fields for human avatar modeling. *arXiv preprint arXiv:2304.04897*, 2023. 2

[24] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-

Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. *arXiv preprint arXiv:2407.12777*, 2024. 1, 2

[25] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[26] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 2

[27] Boyi Li, Jathushan Rajasegaran, Yossi Gandelsman, Alexei A Efros, and Jitendra Malik. Synthesizing moving people with 3d control. *arXiv preprint arXiv:2401.10889*, 2024. 2

[28] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024. 2

[29] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16008–16018, 2025. 3

[30] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019. 2

[31] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 2

[32] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models. *arXiv preprint arXiv:2409.02851*, 2024. 3

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

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

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

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

[38] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. *arXiv preprint arXiv:2412.02684*, 2024. 2, 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[40] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[42] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[43] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: Free-view human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024. 2, 3

[44] Mingze Sun, Junhao Chen, Junting Dong, Yurun Chen, Xinyu Jiang, Shiwei Mao, Puhua Jiang, Jingbo Wang, Bo Dai, and Ruqi Huang. Drive: Diffusion-based rigging empowers generation of versatile and expressive characters. *arXiv preprint arXiv:2411.17423*, 2024. 2

[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5

[46] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 5

[47] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all that nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 2

[48] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic

human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 2, 5

[49] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 2

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 5

[51] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 3

[52] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024. 5

[53] Yuhang Yang, Fengqi Liu, Yixing Lu, Qin Zhao, Pingyu Wu, Wei Zhai, Ran Yi, Yang Cao, Lizhuang Ma, Zheng-Jun Zha, et al. Sigman: Scaling 3d human gaussian generation with millions of assets. *arXiv preprint arXiv:2504.06982*, 2025. 3

[54] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3

[55] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2

[57] 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. 5

[58] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 3

[59] 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. 5

[60] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, 2022. 2

[61] Yizhou Zhao, Hengwei Bian, Kaihua Chen, Pengliang Ji, Liao Qu, Shao-yu Lin, Weichen Yu, Haoran Li, Hao Chen, Jun Shen, et al. Metric from human: Zero-shot monocular metric depth estimation via test-time adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[62] Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-space camera and human reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1216–1226, 2024. 3

[63] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19680–19690, 2024. 2

[64] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 5, 6, 7

[65] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. *arXiv preprint arXiv:2412.14963*, 2024. 2