

# 3D<sup>2</sup>-Actor: Learning Pose-Conditioned 3D-Aware Denoiser for Realistic Gaussian Avatar Modeling

Zichen Tang<sup>1</sup>, Hongyu Yang<sup>1,2\*</sup>, Hanchen Zhang<sup>3</sup>, Jiaxin Chen<sup>3</sup>, Di Huang<sup>3</sup>

<sup>1</sup>School of Artificial Intelligence, Beihang University, Beijing, China

<sup>2</sup>Shanghai Artificial Intelligence Laboratory, Shanghai, China

<sup>3</sup>School of Computer Science and Engineering, Beihang University, Beijing, China  
{zctang, hongyuyang, zhanghanchen, jiaxinchen, dhuang}@buaa.edu.cn

## Abstract

Advancements in neural implicit representations and differentiable rendering have markedly improved the ability to learn animatable 3D avatars from sparse multi-view RGB videos. However, current methods that map observation space to canonical space often face challenges in capturing pose-dependent details and generalizing to novel poses. While diffusion models have demonstrated remarkable zero-shot capabilities in 2D image generation, their potential for creating animatable 3D avatars from 2D inputs remains underexplored. In this work, we introduce 3D<sup>2</sup>-Actor, a novel approach featuring a pose-conditioned 3D-aware human modeling pipeline that integrates iterative 2D denoising and 3D rectifying steps. The 2D denoiser, guided by pose cues, generates detailed multi-view images that provide the rich feature set necessary for high-fidelity 3D reconstruction and pose rendering. Complementing this, our Gaussian-based 3D rectifier renders images with enhanced 3D consistency through a two-stage projection strategy and a novel local coordinate representation. Additionally, we propose an innovative sampling strategy to ensure smooth temporal continuity across frames in video synthesis. Our method effectively addresses the limitations of traditional numerical solutions in handling ill-posed mappings, producing realistic and animatable 3D human avatars. Experimental results demonstrate that 3D<sup>2</sup>-Actor excels in high-fidelity avatar modeling and robustly generalizes to novel poses.

## Introduction

Reconstructing animatable 3D human avatars is essential for applications in VR/AR, the Metaverse, and gaming. However, the task is challenging due to factors like non-rigid complex motions and the stochastic nature of subtle clothing wrinkles, which complicate realistic human actor modeling.

Traditional methods (Collet et al. 2015; Dou et al. 2016; Bogo et al. 2015; Shapiro et al. 2014) are often hindered by labor-intensive manual design and the difficulty of acquiring high-quality data, limiting their applicability in real-world scenarios. Recent advances in neural implicit representations and differentiable neural rendering (Sitzmann et al. 2020; Wang et al. 2021; Lombardi et al. 2019; Gao et al. 2021; Park et al. 2021; Pumarola et al. 2021) have

opened new avenues for character reconstruction and animation from sparse multi-view RGB videos. While techniques like Neural Radiance Field (NeRF) (Mildenhall et al. 2021) excel in synthesizing static scenes, achieving high-fidelity results for dynamic human avatars remains a significant challenge.

One prominent approach involves using a deformation field to map the observation space to a canonical space, as demonstrated in methods like (Su et al. 2021; Peng et al. 2021a). Although learning this backward mapping is relatively straightforward, its generalization to novel poses is often limited due to its reliance on the observation state. Alternatives that employ forward mapping (Wang et al. 2022; Li et al. 2022), utilizing techniques like differentiable root-finding, have been proposed to address these generalization challenges. Additionally, Neural Body (Peng et al. 2021b) introduces a conditional NeRF approach that anchors local features on SMPL (Loper et al. 2015) vertices, which serve as a scaffold for the model.

Despite these advancements, current methods face limitations in handling the complex dynamics of human bodies. For instance, the high stochasticity of clothing, characterized by delicate wrinkles that appear and disappear, poses a significant challenge. Approaches such as (Peng et al. 2021b,a; Wang et al. 2022) optimize per-frame latent codes to capture this variability but struggle to adapt to novel poses due to the limited expressivity of these latent codes. More recent works, such as (Hu and Liu 2023; Qian et al. 2023a; Li et al. 2024), have integrated 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) into their pipelines, delivering significantly improved results in terms of both rendering efficiency and fidelity. However, these approaches do not fully consider finer visual details during novel pose synthesis.

In contrast to these 3D-based approaches, recent 2D generative diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) have demonstrated significant advantages in terms of visual quality. However, the absence of a 3D representation presents a challenge when extending 2D diffusion models to maintain spatial and temporal consistency, particularly in human-centric scenarios. Several works have attempted to achieve 3D-consistent generation by incorporating additional control inputs (Liu et al. 2023) or integrating a 3D representation into the workflow (Liu et al. 2024b; Anciukevicius et al. 2023; Karnewar et al. 2023). These

\*Corresponding author.

methods mainly focus on single-scene generation for general objects, often neglecting temporal consistency, which makes them inadequate for dynamic human modeling.

In this work, we tackle the challenge of reconstructing and animating high-fidelity 3D human avatars with controllable poses by introducing **3D<sup>2</sup>-Actor**, a novel approach featuring a 3D-aware denoiser composed of interleaved pose-conditioned 2D denoising and 3D rectifying steps. Our method uniquely combines the strengths of 3DGS and 2D diffusion models to achieve superior performance in human-centric tasks. Specifically, the 2D denoiser is conditioned on pose clues to generate detailed multi-view images, which are essential for providing rich features supporting the following high-fidelity 3D reconstruction and rendering process. Additionally, the 2D denoiser enhances intricate details from preceding 2D or 3D steps, thereby improving the overall fidelity of the avatar. Complementing the 2D denoiser, our 3D rectifier employs a novel two-stage projection strategy combined with a mesh-based local coordinate representation. The rectifier queries positional offsets and other 3D Gaussian attributes from input images to produce structurally refined multi-view renderings. The integration of 3D Gaussian Splatting ensures high morphological integrity and consistent 3D modeling across various views. To address the temporal discontinuity in animated avatar videos, we propose a Gaussian consistency sampling strategy. This technique utilizes Gaussian local coordinates from previous frames to determine current positions, enabling smooth inter-frame transitions without the need for additional temporal smoothing modules. Our key contributions include:

- **Novel 3D-Aware Denoiser:** We propose a 3D-aware denoiser tailored for reconstructing animatable human avatars from multi-view RGB videos. This method integrates the generative capabilities of 2D diffusion models with the efficient rendering of 3D Gaussian Splatting.
- **Advanced 3D Rectifier:** Our 3D rectifier incorporates a two-stage projection module and a novel local coordinate representation to render structurally refined frames with high multi-view consistency.
- **Gaussian Consistency Sampling Strategy:** We propose a simple yet effective sampling strategy that ensures inter-frame continuity in generated avatar videos. This approach preserves temporal consistency and enhances the overall quality of animated sequences.

## Related Work

### Animatable 3D Human Avatars

In recent years, significant advancements in neural scene representations and differentiable neural rendering techniques have demonstrated high effectiveness in synthesizing novel views for both static (Mildenhall et al. 2021; Sitzmann et al. 2020) and dynamic scenes (Gao et al. 2021; Park et al. 2021; Pumarola et al. 2021). Building upon these studies, various methods attempt to realize 3D human **reconstruction** from sparse-view RGB videos.

Among these approaches, a common line of works involve learning a backward mapping to project points from

the observation space to the canonical space. A-Nerf (Su et al. 2021) constructs a deterministic backward mapping using bone-relative embeddings. Animatable NeRF (Peng et al. 2021a) trains a backward LBS network, yet it encounters challenges in generalizing to poses beyond the distribution. ARAH (Wang et al. 2022) and TAVA (Li et al. 2022), in contrast, utilize a forward mapping to transfer features from the canonical space to the observation space. While the generalizability to novel poses has been improved by these methods, the computational cost of their differentiable root-finding algorithm is quite high.

Another line of works focus on creating a conditional NeRF for modeling dynamic human bodies. Neural Body (Peng et al. 2021b) attaches structured latent codes to posed SMPL vertices and diffuses them into the adjacent 3D space. Despite its capability for high-quality view synthesis, this method performs suboptimally with novel poses. NPC (Su, Bagautdinov, and Rhodin 2023) employs points to store high-frequency details and utilizes a graph neural network to model pose-dependent deformation based on skeleton poses.

A key focus of human avatar **animation** lies in how to transform input poses into changes in appearance. PoseVocab (Li et al. 2023b) proposes joint-structured pose embeddings to encode dynamic human appearance, successfully mapping low-frequency SMPL-derived attributes to high-frequency dynamic human appearances. However, it neglects the fact that identical poses in different motions can result in varying appearances. Some methods (Peng et al. 2021b,a; Wang et al. 2022) employ a per-frame global latent vector to encode stochastic information but this representation cannot generalize well to novel poses. In contrast, our method directly models the distribution of appearances under various poses in image space, enabling a more effective capture of high-frequency visual details.

The advent of 3D Gaussian splatting (Kerbl et al. 2023) has unlocked new possibilities for high-fidelity avatar reconstruction with real-time rendering. A number of concurrent methods (Hu and Liu 2023; Jung et al. 2023; Li et al. 2023a; Qian et al. 2023a,b; Liu et al. 2024a) have investigated the integration of 3D Gaussian with SMPL models for constructing a 3D Gaussian avatar. While the majority of them try to improve rendering efficiency by substituting the neural implicit radiance field with 3D Gaussian representation, our work focuses more on improving the modeling of detailed appearances to enhance image quality.

### 3D Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020) have demonstrated superior performance in 2D image generation. Due to the absence of a standardized 3D data representation, the expansion of 2D diffusion models into the 3D domain remains an unresolved issue. Some studies (Nichol et al. 2022; Gupta et al. 2023; Müller et al. 2023) employ 3D supervision to achieve direct generation of 3D content. However, their practical effectiveness is constrained by the limited size and diversity of the available training data (Po et al. 2023).

Inspired by 3D GANs, various approaches (Anciukevičius et al. 2023; Karnewar et al. 2023; Chen et al. 2023; Szymanowicz, Rupperecht, and Vedaldi 2023) have been pro-

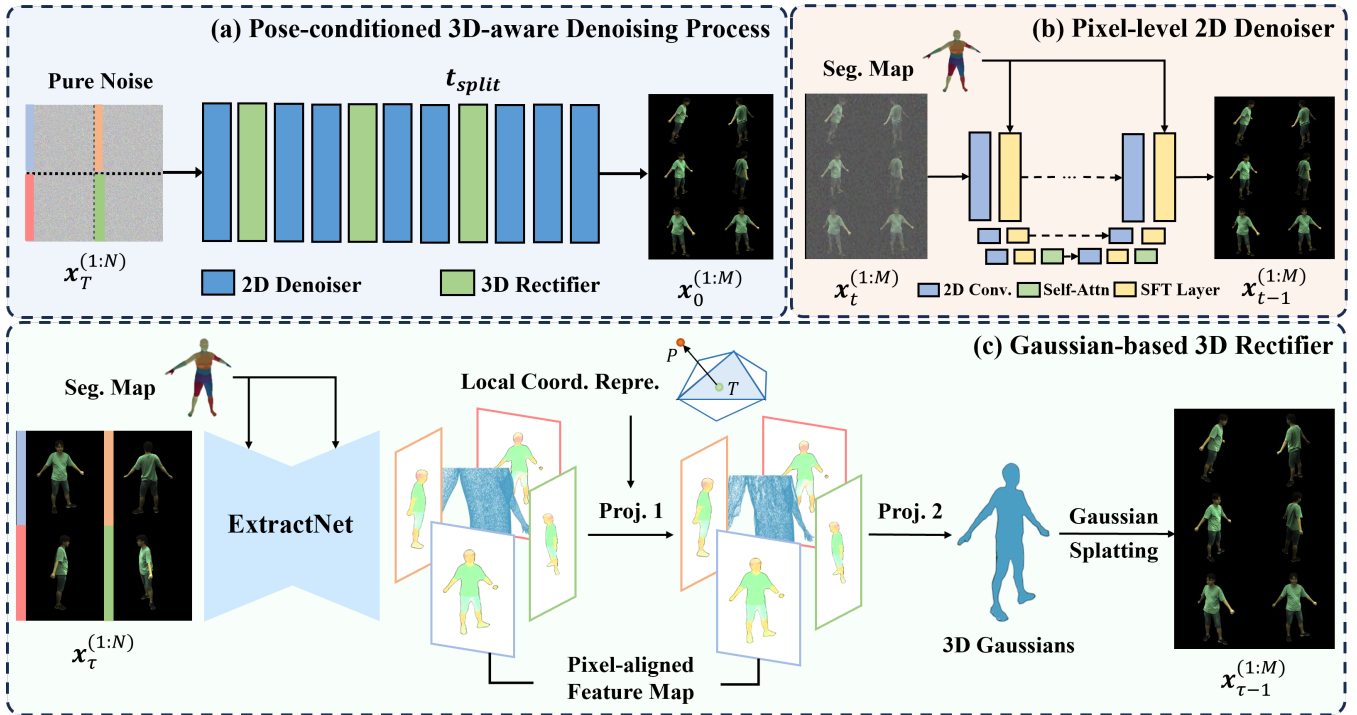


Figure 1: Illustration of the 3D-aware denoising process. (a) The 3D-aware denoising pipeline consists of interlaced 2D and 3D steps. It begins with pure noise input, progressively generating realistic multi-view images of the human avatar with the assistance of pose information. (b) Guided by body segmentation maps as pose cues, the 2D denoiser (blue box in (a)) transforms noised images from the previous 2D or 3D steps into clean ones with enhanced intricate details. It also provides clean images for the subsequent 3D rectifier to achieve accurate 3D human avatar modeling. (c) Given clean images from  $N$  anchor views, the 3D rectifier (green box in (a)) performs a two-stage projection leveraging a mesh-based Gaussian local coordinate representation to reconstruct 3D Gaussians, enabling the rendering of multi-view human images with high 3D consistency.

posed to directly train a diffusion model using 2D image datasets. RenderDiffusion (Anciukevicius et al. 2023) builds a 3D-aware denoiser by incorporating tri-plane representation and neural rendering, predicting a clean image from the noised 2D image. Building upon this research, Viewset Diffusion (Szymanowicz, Rupprecht, and Vedaldi 2023) extends it to multi-view settings. In contrast to these approaches, we take a step further to achieve pose-conditioned human-centric generation. We also present a meticulously designed sampling strategy, enabling the smooth generation of dynamic human videos without introducing extra temporal modules, which is not achieved by current works.

## Method

**Problem statement.** Given multi-view RGB videos of a single human actor as training data, a model should be trained to reconstruct a realistic 3D avatar of the actor and generate high-fidelity and temporal-smoothing videos when performing avatar animation. In the following sections, preliminary will be present first. Next, we will introduce our 3D-aware denoising process combined with 2D denoiser and 3D rectifier. Following that, a simple yet effective inter-frame sampling strategy will be detailed. Finally, we will elucidate the training objectives of our proposed 3D and 2D modules.

## Preliminary

**Diffusion and denoising process** is the core of diffusion models. In this work, we extend this process to our multi-view setting, data  $\mathbf{x} := \mathbf{x}^{(1:N)}$  represents a set of  $N$  images that consistently depict a 3D human avatar. To establish the correlation between the noise distribution and the data distribution, a hierarchy of variables is defined as  $\mathbf{x}_t^{(1:N)}$ ,  $t = 0, \dots, T$ , where  $\mathbf{x}_T^{(1:N)} \sim \mathcal{N}(0, \mathbf{I})$  and  $\mathbf{x}_0^{(1:N)}$  is the set of generated multi-view images. Leveraging the properties of the Gaussian distribution, the forward diffusion process that gradually introduces Gaussian noises to clean data  $\mathbf{x}_0^{(1:N)}$  can be rewritten as:

$$q(\mathbf{x}_t^{(1:N)} | \mathbf{x}_0^{(1:N)}) = \mathcal{N}(\mathbf{x}_t^{(1:N)}; \sqrt{\bar{\alpha}_t} \mathbf{x}_0^{(1:N)}, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (1)$$

where  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$  and  $\alpha_i$  denotes the predefined schedule constant. Correspondingly, the inverse process can be formulated as:

$$p(\mathbf{x}_{t-1}^{(1:N)} | \mathbf{x}_t^{(1:N)}) = \mathcal{N}(\mathbf{x}_{t-1}^{(1:N)}; \mu_\theta(\mathbf{x}_t^{(1:N)}, t), \Sigma_\theta(\mathbf{x}_t^{(1:N)}, t)), \quad (2)$$

where the mean and variance can be estimated through a U-Net  $D_\theta$  trained with loss  $L$  to reconstruct the clean data  $\mathbf{x}_0^{(1:N)}$  from the noised counterpart  $\mathbf{x}_t^{(1:N)}$ :

$$L = \|D_\theta(\mathbf{x}_t^{(1:N)}, t) - \mathbf{x}_0^{(1:N)}\|^2. \quad (3)$$

**3D Gaussian Splatting** (Kerbl et al. 2023) is an effective point-based representation consisting of a set of anisotropic Gaussians. Each 3D Gaussian is parameterized by its center position  $\boldsymbol{\mu} \in \mathbb{R}^3$ , covariance matrix  $\boldsymbol{\Sigma} \in \mathbb{R}^7$ , opacity  $\alpha \in \mathbb{R}$  and color  $\mathbf{c} \in \mathbb{R}^3$ . By splatting 3D Gaussians onto 2D image planes, we can perform point-based rendering:

$$G(\mathbf{p}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \exp\left(-\frac{1}{2}(\mathbf{p} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{p} - \boldsymbol{\mu}_i)\right),$$

$$\mathbf{c}(\mathbf{p}) = \sum_{i \in \mathcal{K}} \mathbf{c}_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \alpha'_i = \alpha_i G(\mathbf{p}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \quad (4)$$

Here,  $\mathbf{p}$  is the coordinate of the queried point.  $\boldsymbol{\mu}_i$ ,  $\boldsymbol{\Sigma}_i$ ,  $\mathbf{c}_i$ ,  $\alpha_i$ , and  $\alpha'_i$  denote the center, covariance, color, opacity, and density of the  $i$ -th Gaussian, respectively.  $G(\mathbf{p}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$  represents the value of the  $i$ -th Gaussian at position  $\mathbf{p}$ .  $\mathcal{K}$  is a sorted list of Gaussians in this tile.

### 3D-Aware Denoising Process

To facilitate human avatar reconstruction (given seen poses) and animation (given novel poses), we innovatively propose a generative 3D-aware denoising process which takes as input pure noise from  $N$  anchor views and SMPL (Loper et al. 2015) pose information and outputs high-quality clean images of the clothed human body. Due to the fact that the 2D denoiser denoises images from different views independently at each step, it fails to ensure consistency in human geometry and texture across views. To address this issue,  $k$  3D rectifying steps are inserted between the 2D denoising steps to maintain the 3D consistency of generated images. Considering that the overall denoising process generates large-scale global structure at early stages and finer details at later stages (Huang et al. 2023), and that 3D consistency among multi-view images is mainly reflected in large-scale features, we merely insert 3D steps in the early stages. This approach aims to improve 3D consistency without jeopardizing the quality of fine texture generation. As illustrated in Fig. 1, we first apply the initial 3D rectifying step after the first 2D denoising step. Then, we select a timestep  $t_{split}$  as the split point between the early and later stages of denoising and insert the final 3D rectifying step at this point. Subsequently,  $k - 2$  3D rectifying steps are evenly inserted between these 2D steps. Finally, a few 2D steps are appended to the overall denoising process, further optimizing the local delicate textures. The details of the 2D denoiser and the 3D rectifier will be introduced below.

### Pixel-Level 2D Denoiser

The 2D denoiser is a fundamental component of our 3D-aware denoising process. Basically, it functions as a refiner that enhances local details in the output images from prior 2D or 3D steps. It can also provide clean images for the subsequent 3D rectifying step. Our 2D denoiser acts like a U-Net (Ronneberger, Fischer, and Brox 2015), taking noisy images, human body segmentation maps and the denoising timestep  $t$  as inputs to predict the denoised clean image at each step. To effectively incorporate pose cues, we draw inspiration from SFTGAN (Wang et al. 2018) and introduce

an SFT layer into each U-Net block to modulate the output of the 2D convolution layer. Given that the 3D rectifier can render images from any camera view, our 2D denoiser is trained on frames with varying views to ensure robustness.

### Gaussian-Based 3D Rectifier

The 3D rectifier plays an essential role in our 3D-aware denoising process. It takes in clean images  $\mathbf{I}^{(1:N)}$  from  $N$  anchor views produced by the previous 2D denoiser and reconstructs the current 3D Gaussians of the avatar. Then, real-time rendering of structure-aligned multi-view images with higher 3D consistency (than the previous 2D step) can be achieved. Note that the 3D rectifier outputs clean images, which can be regarded as “ $\mathbf{x}_0$ ”. Therefore, we can naturally integrate it with the next denoising step leveraging the DDIM (Song, Meng, and Ermon 2020) sampling trick:

$$\mathbf{x}_{t-1}^{(1:N)} = \sqrt{\alpha_{t-1}} \hat{\mathbf{x}}_0^{(1:N)} + c_t \hat{\boldsymbol{\epsilon}}_t^{(1:N)} + \sigma_t \boldsymbol{\epsilon}_t^{(1:N)}, \quad (5)$$

where  $c_t = \sqrt{1 - \alpha_{t-1} - \sigma_t^2}$  and  $\sigma_t$  are necessary coefficients,  $\hat{\mathbf{x}}_0^{(1:N)}$  is the output of the 3D rectifier and  $\boldsymbol{\epsilon}_t^{(1:N)}$  is sampled random noise.

Specifically, we start by rendering body segmentation maps  $\mathbf{S}^{(1:N)} = \mathcal{R}(\mathcal{M}, \mathbf{c}^{(1:N)})$ , where  $\mathcal{M}$  and  $\mathbf{c}^{(1:N)}$  denote the current posed SMPL model and camera poses, respectively, and  $\mathcal{R}$  is the mesh rasterizer. They serve as pose conditions to aid the neural network  $f_{ext}$  in feature extraction for perceiving the 3D actor. Similar to the 2D denoiser, we also insert SFT layers into each U-Net block, effectively leveraging the pose guidance. The entire process of extracting pixel-aligned features can be formulated as:

$$\mathbf{F}_{pix}^{(1:N)} = f_{ext}(\mathbf{I}^{(1:N)}, \mathbf{S}^{(1:N)}, t). \quad (6)$$

After pixel-aligned features are fetched, a key question is how to build the 3D representation of the avatar. Considering the flexibility and efficiency of 3D Gaussian Splatting, we choose it as our 3D representation. Different from current works (Li et al. 2024; Jiang et al. 2024) which use regressed 2D maps to store Gaussian attributes, we seek to predict these attributes with a two-stage projection strategy which fully exploits the 3D spatial information.

**Stage 1: query Gaussian local coordinates.** For any Gaussian  $P$  in the 3D space, its projection  $T$  on the nearest triangle mesh can be represented by barycentric coordinates  $(\lambda_1, \lambda_2, \lambda_3)$ , where  $\lambda_1 + \lambda_2 + \lambda_3 = 1$ . Therefore,  $P$  can be easily described by a local coordinate quaternion  $\boldsymbol{\xi} = (\lambda_1, \lambda_2, \lambda_3, m)$ , where  $m = |\mathbf{T}\mathbf{P}|$  and we can derive the actual position of  $P$  by  $P = \lambda_1 A + \lambda_2 B + \lambda_3 C + m\mathbf{n}$ , where  $A, B, C$  and  $\mathbf{n}$  are the vertex positions and the normal vector of the triangle mesh, respectively. After Gaussian positions are initialized by sampling uniformly on the SMPL mesh, we project each Gaussian onto  $\mathbf{F}_{pix}^{(1:N)}$  to query their position displacements. Rather than directly querying their displacements in observation space, we choose to query their local coordinates instead. This trick constrains the Gaussian movement within a reasonable range, helps model subtle clothes wrinkles and facilitates our inter-frame

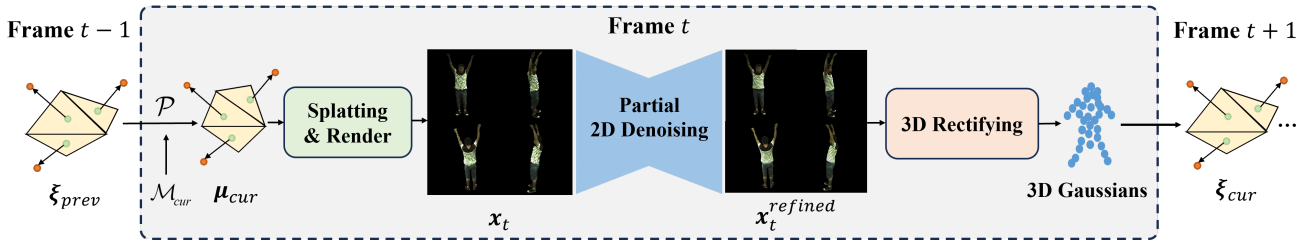


Figure 2: An illustration of the inter-frame Gaussian consistency sampling strategy for improving temporal continuity.

sampling strategy. Specifically, We project each Gaussian onto  $\mathbf{F}_{pix}^{(1:N)}$ , apply bilinear interpolation to obtain feature vectors for each view, and subsequently concatenate them along the feature dimension. If a Gaussian is not visible under a certain view, the corresponding vector is set to zero. Afterwards, a light-weight MLP takes Gaussian positions in the canonical pose space and the previously obtained projected features as input to predict Gaussian local coordinates. **Stage 2: query other Gaussian attributes.** After deriving actual Gaussian positions with local coordinates, we apply another projection to fetch the offsets of the remaining Gaussian attributes. Finally, clean multi-view images with higher 3D consistency can be rendered rapidly.

### Inter-Frame Gaussian Consistency Sampling

From the previous discussion, we know that the 3D-aware denoising process can generate highly realistic single-frame renderings from pure noise. However, applying this independently to each frame in video generation can cause noticeable inconsistencies, severely affecting the visual quality. Adding a temporal module is a possible way to address this issue, but this comes with an increased computational cost for training. In contrast, we design a novel inter-frame Gaussian consistency sampling strategy during inference to ensure seamless inter-frame transitions when synthesizing videos. The core idea is to use information from the previous frame to generate a rough image of the current frame, then perform several late-stage 2D denoising steps to correct visual artifacts. To obtain the current frame’s Gaussians, we should propagate the SMPL pose change to the change of Gaussian positions. Fortunately, this can be achieved easily using our mesh-based local coordinates. As depicted in Fig. 2, given Gaussian local coordinates  $\xi_{prev}^{(1:n)}$  of the last frame and the current frame’s SMPL mesh  $\mathcal{M}_{cur}$ , the current Gaussian positions  $\mu_{cur}^{(1:n)}$  can be derived by:

$$\mu_{cur}^{(1:n)} = \mathcal{P}(\xi_{prev}^{(1:n)}, \mathcal{M}_{cur}), \quad (7)$$

where  $n$  is the number of Gaussians and  $\mathcal{P}$  is an operation that transforms Gaussian local coordinates to their actual positions in the observation space. However, rendering images directly using these Gaussians may lead to noticeable artifacts. To mitigate this, we add slight noise to the rendered images and perform several 2D denoising steps from a smaller timestep, yielding more plausible results. Finally, we apply an additional 3D step to obtain the 3D Gaussians of the current frame. Adopting this sampling strategy for video

generation offers distinct advantages in terms of inter-frame continuity compared to generating each frame separately. It also provides computational efficiency, as denoising is only performed partially from a relatively small timestep.

### Training Objective

The complete training process includes two separate training workflows for the 3D rectifier  $G_{3D}$  and the 2D denoiser  $D_{2D}$ .  $G_{3D}$  is trained with a loss function that includes both photometric loss and mask loss. Given clean video frames  $\mathbf{I}_f^{(1:N)}$  from  $N$  anchor views and conditional SMPL segmentation maps  $\mathbf{S}_f^{(1:N)}$  of frame  $f$ , the training objective is to reconstruct accurate 3D Gaussians from the given input to achieve consistent 3D rendering from  $M$  specified views  $\mathbf{c}^{(1:M)}$ . The loss function measures the similarity between the rendered multi-view images and the ground-truth images, including the  $L_2$  loss for the RGB images:

$$L_{rgb} = \|\mathbf{G}_{3D}(\mathbf{I}_f^{(1:N)}, \mathbf{S}_f^{(1:N)}, \mathbf{c}^{(1:M)}) - \mathbf{I}_f^{(1:M)}\|^2, \quad (8)$$

and the  $L_2$  loss  $L_{mask}$  for the masks, which is omitted for brevity. The overall loss of  $G_{3D}$  can be represented as:

$$L_{3D} = \lambda_{rgb} L_{rgb} + \lambda_{mask} L_{mask}. \quad (9)$$

In terms of the 2D denoiser  $D_{2D}$ , given a clean video frame  $\mathbf{I}_f$ , the corresponding SMPL segmentation map  $\mathbf{S}_f$  and timestep  $t$ , we only apply RGB loss to train the model:

$$L_{2D} = \|D_{2D}(\mathbf{I}_f, \mathbf{S}_f, t) - \mathbf{I}_f\|^2. \quad (10)$$

## Experiments

### Implementation Details

The 3D rectifier takes clean images at a resolution of  $512 \times 512$  from  $N = 4$  anchor views as input, reconstructs 3D avatar Gaussians, and renders  $M = 8$  multi-view images at the same resolution. The number of Gaussians sampled on the SMPL (Loper et al. 2015) mesh is  $n = 373056$ . We train this model with a learning rate of  $5 \times 10^{-5}$ . The 2D denoiser functions at a resolution of  $512 \times 512$ , consistent with the resolution of the ground-truth images. This model is trained with a learning rate of  $4 \times 10^{-4}$ . When conducting single frame novel pose synthesis, our 3D-aware denoising process has 20 denoising steps in total. All experiments are conducted on NVIDIA RTX 3080 Ti GPUs.

Method	313			315			377			386		
	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
ARAH	24.3	0.097	32.6	<b>20.9</b>	0.102	31.4	<b>24.8</b>	0.109	32.6	<b>27.9</b>	0.152	54.6
PoseVocab	23.3	0.101	27.5	20.6	0.100	27.2	24.1	0.091	<b>25.8</b>	<u>26.8</u>	0.134	31.9
Ours	<u>23.5</u>	<b>0.080</b>	<b>19.5</b>	<u>20.7</u>	<b>0.090</b>	<b>20.2</b>	<u>24.4</u>	<b>0.090</b>	<u>26.4</u>	<u>26.8</u>	<b>0.123</b>	<b>28.6</b>

Table 1: Quantitative comparison of single-frame novel pose synthesis against ARAH (Wang et al. 2022) and PoseVocab (Li et al. 2023b) on 4 sequences of the ZJU-MoCap dataset. Bold indicates the best, while underline denotes the second-best.

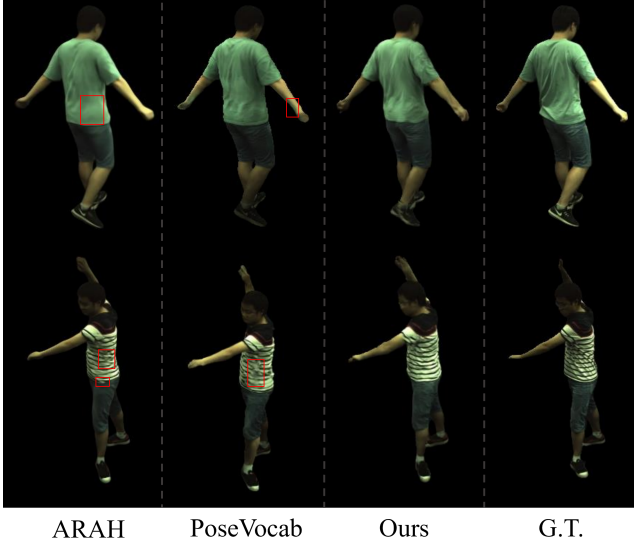


Figure 3: Qualitative comparison of single-frame novel pose synthesis results against ARAH (Wang et al. 2022) and PoseVocab (Li et al. 2023b) on sequences 313 and 315 of the ZJU-MoCap dataset. Please zoom in for better observation.

**Dataset.** Our experiments are conducted on the ZJU-MoCap (Peng et al. 2021b) dataset, which includes 9 sequences captured with 23 calibrated cameras. Each sequence features a video of an individual performing a specific action. We utilize 80% of frames from each sequence for training and left the remaining frames for testing.

**Metrics.** We adopt Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), and Frechet Inception Distance (FID) (Heusel et al. 2017) for quantitative evaluation.

**Baselines.** We compare our method with two state-of-the-art counterparts suitable for ZJU-MoCap dataset: ARAH (Wang et al. 2022) and PoseVocab (Li et al. 2023b). We re-trained the baseline methods using their officially released code to align the training/test set split for all the methods.

### Single-Frame Novel Pose Synthesis

To evaluate the visual quality of the generated frames and the pose generalization performance of our method, testing is specifically performed on novel poses that are not included in the training dataset. It is important to note that due to the stochasticity introduced by factors such as clothing wrin-



Figure 4: Consecutive frame generation results. Top row shows results using the proposed sampling strategy; bottom row displays results from independent sampling.

kles, for the unseen poses, ground-truth images represent only *one possible scenario*. Therefore, metrics emphasizing pixel-level correspondence, such as PSNR, may not comprehensively evaluate the fidelity of the generated images. In our evaluation, we primarily employ LPIPS and FID, metrics describing perceptual similarity, while we also provide experimental results with PSNR (in light font).

**Quantitative results.** When performing novel pose synthesis on different IDs, we search for the best  $t_{split}, k$  pair for each ID regarding their various clothes wrinkles and action dynamics. Tab. 1 presents a quantitative comparison among ARAH, PoseVocab and our method across four sequences of the ZJU-MoCap dataset, we report the average values of these metrics across all test frames. It can be illustrated that our method achieves the best or second-best results across the four sequences. While ARAH achieves the highest PSNR, our approach offers a balanced performance, excelling in LPIPS and FID, which indicates that our method has good generalizability and generative capability given novel poses.

**Qualitative results.** The results of the qualitative experiments are depicted in Fig. 3. ARAH tends to produce relatively blurry images in these scenarios. In contrast, images generated by PoseVocab exhibit relatively clear texture details, albeit with some issues. For instance, the stripes on the T-shirt appear somewhat blurry, and there is color “bleeding” from the green clothing onto the arms. Contrarily, the images produced by our method show clearer finer details and a higher sense of realism in clothing wrinkle details.

Method	313			315		
	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
Ours-2D	22.7	<u>0.086</u>	<b>16.8</b>	19.7	<u>0.098</u>	<b>19.8</b>
Ours-3D	<b>24.2</b>	0.148	115.8	<b>21.6</b>	0.150	79.7
Ours-overall	<u>23.5</u>	<b>0.080</b>	<u>19.5</u>	<u>20.7</u>	<b>0.090</b>	<u>20.2</u>

Table 2: Quantitative ablation study on our 3D-aware denoising process. Ours-2D uses only 2D denoisers, while Ours-3D retains the initial 3D rectifier but omits later 2D or 3D steps.

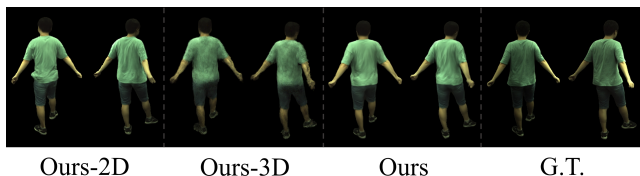


Figure 5: Novel pose synthesis results with different designs of our 3D-aware denoiser.

### Continuous Video Synthesis

Fig. 4 illustrates the experimental results of generating a sequence of consecutive frames using our method, which is also performed on novel poses. When the Gaussian consistency sampling strategy is not utilized, and instead, sampling begins with pure Gaussian noise for each frame, the resulting frames experience pronounced inter-frame inconsistency. On the contrary, deriving the 3D Gaussians for the current frame with the Gaussian local coordinates from the previous frame first and then proceeding with subsequent 2D denoising processes significantly enhances the continuity between the output frames.

### Ablation Studies

**3D-aware denoising process.** The complete 3D-aware denoising process consists of two submodules: 3D rectifier and 2D denoiser. LPIPS and FID metrics in Tab. 2 indicates that conducting only the 2D process yields relatively better performance, while images obtained solely through the 3D counterpart are the least satisfactory. The results in Fig. 5 reveal that images generated solely through 2D denoising display richer local details but have limitations in overall modeling. In contrast, the 3D process excels in producing reasonable global attributes but fails to model wrinkles and occlusions. Consecutively performing these two processes allows for a synergistic combination of their strengths, as depicted in “ours”. We further analyze the impact of varying split point  $t_{split}$  and the insertion counts of 3D rectifier  $k$  on the result. As presented in Tab. 3, as  $t_{split}$  increases, the 2D denoiser applies stronger corrections, resulting in lower LPIPS and FID values, indicating improved image realism. Moreover, the number of 3D rectifying steps has little impact on image quality when  $t_{split}$  is fixed.

**Gaussian local coordinate representation.** We conduct an ablation study comparing the image generation quality

$t_{split}$ $k$	313		315		377		386		
	LPIPS $\downarrow$	FID $\downarrow$	LPIPS $\downarrow$	FID $\downarrow$	LPIPS $\downarrow$	FID $\downarrow$	LPIPS $\downarrow$	FID $\downarrow$	
200	2	0.082	21.5	0.093	21.9	0.094	29.4	0.127	31.8
3	0.082	23.1	0.093	21.7	0.093	28.6	0.127	31.9	
4	0.082	21.7	0.093	21.4	0.094	28.5	0.127	32.3	
300	2	<b>0.080</b>	<b>19.5</b>	0.092	21.2	<u>0.091</u>	<u>27.2</u>	<b>0.123</b>	<b>28.6</b>
3	<u>0.081</u>	20.0	<u>0.091</u>	<u>20.5</u>	<b>0.090</b>	<b>26.4</b>	<u>0.124</u>	<u>29.2</u>	
4	<u>0.081</u>	<u>19.7</u>	<b>0.090</b>	<b>20.2</b>	<u>0.091</u>	27.6	<b>0.123</b>	29.5	

Table 3: Comparison of image generation quality under varying  $t_{split}$  and  $k$ .

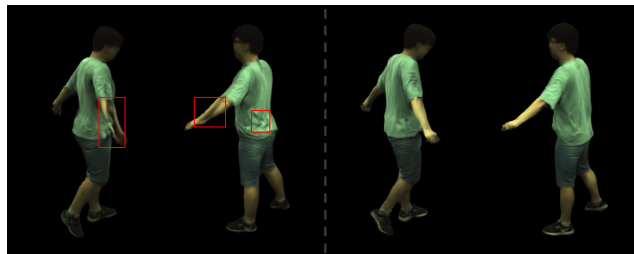


Figure 6: Novel pose synthesis results without (w/o) and with (w/) our mesh-based local coordinate representation.

of our framework with and without our mesh-based Gaussian local coordinate representation. From Fig. 6, we can see that this representation allows 3D Gaussians to move flexibly within a reasonable range, resulting in more realistic modeling of body and clothes details, such as wrinkles and occlusions. In contrast, a 3D-aware denoiser without this representation lacks the ability to express details reasonably. As a consequence, the generated images tend to exhibit unnatural clothing wrinkles and fail to accurately model limbs, leading to a significant decline in visual quality.

## Conclusion

We present 3D<sup>2</sup>-Actor, an innovative pose-conditioned 3D-aware denoiser designed for the high-fidelity reconstruction and animation of 3D human avatars. Our approach employs a 2D denoiser to refine the intricate details of noised images from the previous step, generating high-quality clean images that facilitate the 3D reconstruction process in the subsequent 3D rectifier. Complementing this, our 3D rectifier employs a two-stage projection strategy with a novel local coordinate representation to render multi-view images with enhanced 3D consistency by incorporating 3DGS-based techniques. Additionally, we introduce a Gaussian consistency sampling strategy that improves inter-frame continuity in video synthesis without additional training overhead. Our method achieves realistic human animation and high-quality dynamic video generation with novel poses.

## Acknowledgments

This work is partly supported by the National Key R&D Program of China (2022ZD0161902), Beijing Municipal Natural Science Foundation (No. 4222049), the National Natural Science Foundation of China (No. 62202031), and the Fundamental Research Funds for the Central Universities.

## References

- Anciukevičius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12608–12618.
- Bogo, F.; Black, M. J.; Loper, M.; and Romero, J. 2015. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proceedings of the IEEE international conference on computer vision*, 2300–2308.
- Chen, H.; Gu, J.; Chen, A.; Tian, W.; Tu, Z.; Liu, L.; and Su, H. 2023. Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruction. In *ICCV*.
- Collet, A.; Chuang, M.; Sweeney, P.; Gillett, D.; Evseev, D.; Calabrese, D.; Hoppe, H.; Kirk, A.; and Sullivan, S. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4): 1–13.
- Dou, M.; Khamis, S.; Degtyarev, Y.; Davidson, P.; Fanello, S. R.; Kowdle, A.; Escolano, S. O.; Rhemann, C.; Kim, D.; Taylor, J.; et al. 2016. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4): 1–13.
- Gao, C.; Saraf, A.; Kopf, J.; and Huang, J.-B. 2021. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5712–5721.
- Gupta, A.; Xiong, W.; Nie, Y.; Jones, I.; and Oğuz, B. 2023. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Hu, S.; and Liu, Z. 2023. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:2312.02973*.
- Huang, Y.; Wang, J.; Shi, Y.; Tang, B.; Qi, X.; and Zhang, L. 2023. Dreamtime: An Improved Optimization Strategy for Diffusion-guided 3D Generation. In *The Twelfth International Conference on Learning Representations*.
- Jiang, Y.; Liao, Q.; Li, X.; Ma, L.; Zhang, Q.; Zhang, C.; Lu, Z.; and Shan, Y. 2024. UV Gaussians: Joint Learning of Mesh Deformation and Gaussian Textures for Human Avatar Modeling. *arXiv preprint arXiv:2403.11589*.
- Jung, H.; Brasch, N.; Song, J.; Perez-Pellitero, E.; Zhou, Y.; Li, Z.; Navab, N.; and Busam, B. 2023. Deformable 3d gaussian splatting for animatable human avatars. *arXiv preprint arXiv:2312.15059*.
- Karnewar, A.; Vedaldi, A.; Novotny, D.; and Mitra, N. J. 2023. Holodiffusion: Training a 3D diffusion model using 2D images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18423–18433.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).
- Li, M.; Tao, J.; Yang, Z.; and Yang, Y. 2023a. Human101: Training 100+ fps human gaussians in 100s from 1 view. *arXiv preprint arXiv:2312.15258*.
- Li, R.; Tanke, J.; Vo, M.; Zollhöfer, M.; Gall, J.; Kanazawa, A.; and Lassner, C. 2022. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, 419–436. Springer.
- Li, Z.; Zheng, Z.; Liu, Y.; Zhou, B.; and Liu, Y. 2023b. PoseVocab: Learning Joint-structured Pose Embeddings for Human Avatar Modeling. In *ACM SIGGRAPH Conference Proceedings*.
- Li, Z.; Zheng, Z.; Wang, L.; and Liu, Y. 2024. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19711–19722.
- Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9298–9309.
- Liu, X.; Wu, C.; Liu, X.; Liu, J.; Wu, J.; Zhao, C.; Feng, H.; Ding, E.; and Wang, J. 2024a. GEA: Reconstructing Expressive 3D Gaussian Avatar from Monocular Video. *arXiv preprint arXiv:2402.16607*.
- Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2024b. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *The Twelfth International Conference on Learning Representations*.
- Lombardi, S.; Simon, T.; Saragih, J.; Schwartz, G.; Lehrmann, A.; and Sheikh, Y. 2019. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*.
- Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6).
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Müller, N.; Siddiqui, Y.; Porzi, L.; Buló, S. R.; Kotschieder, P.; and Nießner, M. 2023. Diffrrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4328–4338.

- Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.
- Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.
- Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; and Bao, H. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14314–14323.
- Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9054–9063.
- Po, R.; Yifan, W.; Golyanik, V.; Aberman, K.; Barron, J. T.; Bermano, A. H.; Chan, E. R.; Dekel, T.; Holynski, A.; Kanazawa, A.; et al. 2023. State of the Art on Diffusion Models for Visual Computing. *arXiv preprint arXiv:2310.07204*.
- Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.
- Qian, S.; Kirschstein, T.; Schoneveld, L.; Davoli, D.; Giebenhain, S.; and Nießner, M. 2023a. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *arXiv preprint arXiv:2312.02069*.
- Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; and Tang, S. 2023b. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. *arXiv preprint arXiv:2312.09228*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.
- Shapiro, A.; Feng, A.; Wang, R.; Li, H.; Bolas, M.; Medioni, G.; and Suma, E. 2014. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4): 201–211.
- Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33: 7462–7473.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Su, S.-Y.; Bagautdinov, T.; and Rhodin, H. 2023. NPC: Neural Point Characters from Video. In *Proceedings of the IEEE/CVF International conference on computer vision*, 14795–14805.
- Su, S.-Y.; Yu, F.; Zollhöfer, M.; and Rhodin, H. 2021. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34: 12278–12291.
- Szymanowicz, S.; Rupprecht, C.; and Vedaldi, A. 2023. Viewset Diffusion: (0-)Image-Conditioned 3D Generative Models from 2D Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8863–8873.
- Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *Advances in Neural Information Processing Systems*, 34: 27171–27183.
- Wang, S.; Schwarz, K.; Geiger, A.; and Tang, S. 2022. Arah: Animatable volume rendering of articulated human sdf. In *European conference on computer vision*, 1–19. Springer.
- Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.