# Guess The Unseen: Dynamic 3D Scene Reconstruction from Partial 2D Glimpses

Inhee Lee Byungjun Kim Hanbyul Joo
Seoul National University
{ininin0516, byungjun.kim, hbjoo}@snu.ac.kr
https://snuvclab.github.io/gtu/

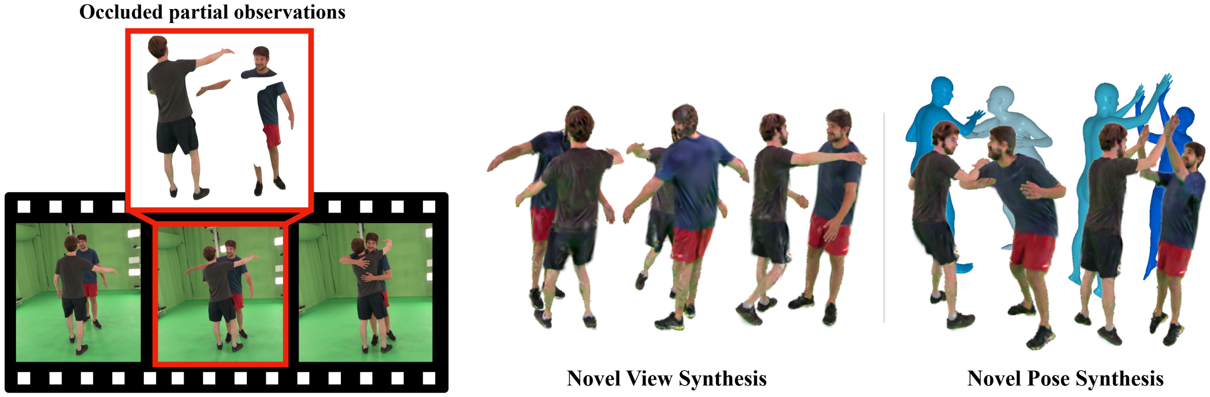



Figure 1. We present a method to reconstruct dynamic scenes from a monocular video capturing partial 2D observations. As a key advantage, our method can estimate the unseen body parts by leveraging a pre-trained diffusion model [40] via SDS method [37]. The reconstructed scenes can be rendered to any viewpoint and each human body can be transformed into any body posture controlled by SMPL [27] parameters.

## Abstract

*In this paper, we present a method to reconstruct the world and multiple dynamic humans in 3D from a monocular video input. As a key idea, we represent both the world and multiple humans via the recently emerging 3D Gaussian Splatting (3D-GS) representation, enabling to conveniently and efficiently compose and render them together. In particular, we address the scenarios with severely limited and sparse observations in 3D human reconstruction, a common challenge encountered in the real world. To tackle this challenge, we introduce a novel approach to optimize the 3D-GS representation in a canonical space by fusing the sparse cues in the common space, where we leverage a pre-trained 2D diffusion model to* ***synthesize*** *unseen views while keeping the consistency with the observed 2D appearances. We demonstrate our method can reconstruct high-quality animatable 3D humans in various challenging examples, in the presence of occlusion, image crops, few-shot, and extremely sparse observations. After reconstruction, our method is capable of not only rendering the scene in any novel views at arbitrary time instances, but also editing the 3D scene by removing individual humans or applying different motions for each human. Through various experiments, we demonstrate the quality and efficiency of our methods over alternative existing approaches.*

## 1. Introduction

The process of digitizing our world in 3D necessitates the reconstruction of not only static environmental elements but also dynamic objects, particularly humans. Given the limited availability of multi-view camera setups, a groundbreaking leap can be achieved by developing a 4D reconstruction method capable of rendering the scenes at novel views and arbitrary times just using a monocular video input. Reconstructing static components (e.g., buildings) from monocular video benefits from the well-established multi-view geometry principles [12], where epipolar geometry constraints across different views still hold at different times. Recent advances have further enhanced the quality of these reconstructions by leveraging implicit 3D representations, as demonstrated by NeRF [31], NeuS [49], and Gaussian Splatting (3D-GS) [19], resulting in more realistic renderings.

However, the same approach is not directly applicable to dynamically moving components, specifically humans. Early work addresses this problem within the context of general non-rigid structure-from-motion approaches [1, 35]. More recent breakthroughs leverage human prior models, such as SMPL [27, 54], as a way to canonicalize the non-rigid observations from multiple views of the monocular video and transform back into the posed spaces [7, 8, 11, 16, 51].

Yet, these approaches often assume the scenarios where the camera focuses on the human subject, capturing their entire body while the target person revolves around the camera's field of view. While this approach is suitable for intentionally digitizing a specific individual, they encounter substantial challenges in in-the-wild video scenarios, where humans are captured in partial, occluded, cropped, and sparse observed conditions. See the examples shown in Fig. 1 and Fig. 4. Moreover, reconstructing and rendering multiple individuals along with 3D backgrounds within the existing approaches is non-trivial, mainly due to the complexities of integrating multiple neural radiance field models [34, 59].

In this paper, we present a method to reconstruct both the static world and multiple dynamically moving humans in 3D from a monocular video input, specifically focusing on scenarios with extremely limited and sparse observations. To address this challenge, we represent both the world and multiple humans in a common Gaussian splatting 3D representation [19]. Our approach allows to efficiently compose them for novel view rendering or scene editing. Notably, to tackle the scenarios with extremely limited and sparse observations in 3D human reconstruction, we introduce a novel approach to optimize the 3D GS representation in a canonical space by fusing the sparse cues in the common space. As a core idea, we leverage a pre-trained 2D diffusion model, equipped with Texture Inversion [10], to synthesize unseen views without losing the consistency with the observed 2D appearances [37, 40]. Via thorough evaluations, we demonstrate that our approach successfully reconstructs high-quality animatable 3D human avatars of dynamically moving individuals from sparse and partial observations. The animatable nature of our 3D human reconstruction outputs enables us to replay the observed motions of the humans in novel views and edit the postures of the humans with arbitrary motions as well. Our contribution is summarized as: (1) representing both a 3D world and multiple humans in a common 3D GS representation for efficient composing and rendering; (2) reconstructing and canonicalizing animatable 3D humans from sparse and partial 2D observations by incorporating SDS loss with a diffusion model and textual inversion; and (3) introducing efficient 4D scene reconstruction and editing pipeline.

## 2. Related Work

**Human Reconstruction from Monocular Video.** There has been a series of approaches [15, 17, 36, 51, 55, 58, 58] to reconstruct 3D human avatars from a monocular video capturing moving humans. Mostly, they tackle the problem by finding the correspondences between each frame and optimizing them in a common canonical space. To find the correspondences across the frames, diverse prior knowledge is leveraged such as parametric model [2, 27], pixel-aligned features [33], or optical flow [56]. After the success of NeRF [31], recent methods [15, 36, 51, 58] use NeRF and its variants to reconstruct a human by leveraging a parametric model SMPL [27]. HumanNeRF [51] and SelfRecon [15] improve the reconstruction quality by correcting the inaccurate canonicalization originated by non-rigid deformation. Vid2Avatar and Neuman [11, 17] reconstruct a human without a mask by learning a background jointly. InstantAvatar [16] reduces the required time of optimization from a few hours into a minute leveraging iNGP [32]. OccNeRF [53] proposes a method that can reconstruct people even with occlusion, using surface-based rendering and visibility attention. However, unlike our method, all of the above except OccNeRF assumes the person is not occluded and most of the body is shown in the video, which is rare in in-the-wild videos.

**Diffusion on 3D Tasks.** After the recent breakthroughs of diffusion models on image generation task [13], several methods suggest a way to use diffusion model on 3D tasks [9, 23, 26, 37, 45, 48, 62]. For example, RGBD2 [23] trains an RGB-D diffusion model to complete the unobserved area of a room using diffusion inpainting approach [28] and DiffuStereo [45] trains diffusion-stereo network for higher reconstruction quality in sparse multi-view settings. In particular, the SDS [37] method which leverages a pretrained text-to-image diffusion model [43] has been applied widely, such as text-to-3D [37, 48, 50] or single image-to-3D generation tasks [25, 26, 30]. However, the vanilla SDS loss is not 3D consistent itself and prone to artifacts like the Janus effect. To improve the SDS loss, many techniques are proposed such as leveraging 3D prior [26], fine-tuning diffusion [50], giving better conditions [30] and using advanced optimization schemes [5, 6, 6, 25]. SDS Loss is also applied to make better 3D avatars recently [14, 24]. However, there has been no trial on completing an imperfect reconstruction of a human with a diffusion.

**Compositional Human-Scene Reconstruction.** Separating 4D scenes into static backgrounds and dynamic objects is a common approach in the 4D reconstruction problem. The static-dynamic separation can be done by prior knowledge on the targets [22, 34], such as cars and pedestrians are dynamic, or can be performed automatically by minimizing a certain energy term [52]. Recent human reconstruction methods also use compositional approaches to reconstruct a human from videos [11, 17, 46, 59]. For monocular reconstruction, Neuman [17] and Vid2avatar [11] use two separate implicit functions each of which represents background and person respectively. While they allow the model to reconstruct a person regardless of the person's mask quality, these models cannot handle occlusion and multiple people at once. In a multi-view setting, *Shuai et al.* [46] tackles more practical situations where multiple people interact with objects. It models each person, object, and background with a NeuralBody [36] and NeRF [31] respectively and renders them compositionally through ST-NeRF [59] pipeline. However, it requires 8 multi-view videos as input, which is unavailable in in-the-wild situations.

## 3. Preliminaries

In our method, we use 3D Gaussian Splatting [19] to represent 4D scenes, and use Score Distillation Sampling (SDS) as a tool to estimate unseen human body parts. Here, we provide an overview of these preliminary concepts.

### 3.1. 3D Gaussian Splatting

3D Gaussian Splatting (3D-GS) is an explicit 3D representation to model a radiance field of a static 3D scene with a set of 3D Gaussians and their attributes [19]. A 3D static scene can be modeled by a set of 3D Gaussians $\{G_i\}_{i=1}^{M}$ where the $i$-th Gaussian is represented by $G_i = \{\boldsymbol{\mu}_i, \boldsymbol{q}_i, \boldsymbol{s}_i, \boldsymbol{c}_i, o_i\}$, where $\boldsymbol{\mu} \in \mathbb{R}^3$ is the Gaussian center, $\boldsymbol{s} \in \mathbb{R}^3$ and $\boldsymbol{q} \in SO(3)$ are respectively the scaling factor and the rotation represented in quaternion to define the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, $\boldsymbol{c_i} \in \mathbb{R}^3$ is the color, and $o_i \in \mathbb{R}$ is opacity. For a 3D query location $\boldsymbol{x} \in \mathbb{R}^3$, its Gaussian weight $\mathbf{g}(\boldsymbol{x})$ is represented as:

$$\mathbf{g}(\boldsymbol{x}) = e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{1}$$

where the symmetric 3D covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$ is represented by

$$\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{R}^T. \tag{2}$$

$\boldsymbol{R} = \text{quat2rot}(\boldsymbol{q})$ is a rotation matrix converted from $\boldsymbol{q}$, and $\boldsymbol{S} = \text{diag}(\boldsymbol{s})$ is a diagonal matrix from scaling factor $\boldsymbol{s}$.

3D-GS rasterizes these 3D Gaussians $\{G_i\}_{i=1}^{M}$ by sorting them in depth order in camera space and projecting them to the image plane. If $N$ number of Gaussians are projected on 2D location $\boldsymbol{p} \in \mathbb{R}^2$, the pixel color $C(\boldsymbol{p})$ is given by $\alpha$-blended rendering as follows:

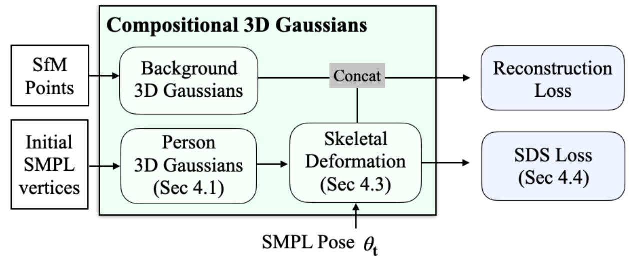


Figure 2. **Method overview.** Overview of our pipeline. (Sec. 4).

$$C(\boldsymbol{p}) = \sum_{i\in N} c_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{3}$$

$$\alpha_i = \mathbf{g}_i^{2D}(\boldsymbol{p}) \cdot o_i, \tag{4}$$

where $\mathbf{g}_i^{2D}$ is the weight after the 2D projection of 3D Gaussian $\mathbf{g}_i$ to the image plane, and we use the Jacobian of the affine approximation of the projective transformation, following previous approaches [19, 63]. As the output of 3D scene reconstruction, we obtain the parameters of 3D Gaussians $\mathcal{G} = \{G_i\}_{i=1}^{M}$ by optimizing them with reconstruction loss calculated from rendering Eq. (3).

### 3.2. Score Distillation Sampling

Score Distillation Sampling (SDS) method [37] is an approach that leverages the prior underlying text-to-image (T2I) diffusion models to generate 3D content. SDS optimizes any differentiable 3D representation $\boldsymbol{\Theta}$ by aligning rendered output $\{\boldsymbol{I}\}$ from arbitrary views to be on the distribution of diffusion model $\phi$. This can be achieved by minimizing the residual between noise $\boldsymbol{\epsilon}$, which perturbs $\boldsymbol{z}$ into $\boldsymbol{z}_t$, and predicted noise $\boldsymbol{\epsilon}(\boldsymbol{z}_\tau; y, \tau)$ where $\boldsymbol{z}$ is a latent of $\{\boldsymbol{I}\}$ encoded by VAE of latent diffusion model [41]. By omitting gradient through diffusion model $\phi$, the SDS loss can be written as follows:

$$\nabla_\Theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\omega(\tau)(\boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t; y, \tau) - \boldsymbol{\epsilon})\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{I}}\frac{\partial \boldsymbol{I}}{\partial \Theta}\right], \tag{5}$$

where $\boldsymbol{z}_\tau$ is a noised latent with time step $\tau$ and $y$ is text prompt conditioning diffusion model $\phi$. The $\omega(\tau)$ denotes the weighting function defined by the scheduler of the diffusion model.

## 4. Method

### 4.1. Overview

Our model reconstructs dynamically moving multi-humans and the static background jointly from a casually captured monocular video. Our system takes, as input, $T$ frames of images $\{I_t\}_{t=1...T}$ with corresponding camera parameters $\{\mathbf{P}_t\}_{t=1...T}$, and outputs the 3D scenes in the representation of 3D Gaussian Splatting $\mathcal{G}^{BG}$ and $\{\mathcal{G}_j^h\}_{j=1...N}$,

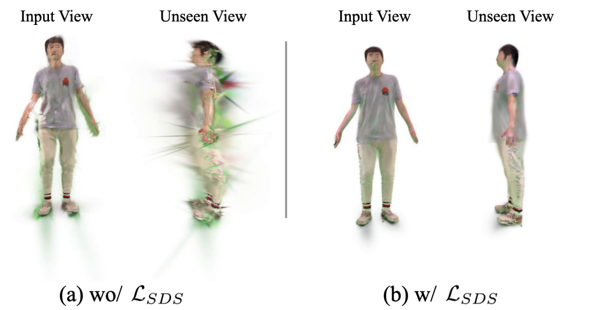



Figure 3. **Failure examples of optimizing 3D-GS naively.** (a) shows that naively optimizing 3D-GS suffers from artifacts shaped like a hedgehog in unseen view and input view. (b) shows that our SDS loss effectively removes the artifacts observed in both input and unseen views.

where $\mathcal{G}^{BG}$ is to represent the 3D background and $\mathcal{G}_j^h$ is for the $j$-th human.

Importantly, we represent the individual human in a canonical space mapped to the rest pose (or A-pose) of SMPL model [27], which can be transformed to any "posed" space parameterized by SMPL pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{72}$. Thus, the appearance of the $j$-th human at time $t$ can be represented as $\mathcal{G}_j^h(\boldsymbol{\theta}_{j,t})$ by inputting the corresponding SMPL parameters $\boldsymbol{\theta}_{j,t}$ for $j$-th humans at time $t$. Note that within our representation, the posture of each individual can be controlled independently from other scene parts. This provides us the flexibility to edit people's body motions using various available motion capture data [29].

Since we build both dynamic humans and backgrounds in the same 3D-GS representation, we can effectively render the whole scene by compositing 3D-GS representations, $\mathcal{G}^{\text{All}} = \{\mathcal{G}^{BG}\} \cup \{\mathcal{G}_j^h\}_{j=1...N}$, where we can use the same rendering function Eq. (3) without any modification. This shows the major advantage over the alternative approaches such as NeRF-based representation [11, 51], where compositing multiple humans is not trivial. Our method is much more convenient and efficient, showing much faster rendering speed (e.g., 40 times) than the competing approaches as demonstrated in our experiments.

Notably, we mainly focus on reconstructing the 3D humans from sparse monocular observations, in the presence of severe occlusions, cropped views, and few shots, which are commonly observable in the wild. We address this challenge by fusing the observed cues into the canonical spaces, for which we introduce the transformation between the posed space to the canonical space, while we leverage 3D-GS representation (Sec. 4.3). As a core contribution, we also present a solution to include 2D diffusion prior as a way to synthesize the missing and unobserved part of target human while keeping the consistency to the observed parts (Sec. 4.4), where we further enhance the quality by incorporating Texture Inversion technique [21] to better preserve the target identity. The Fig. 2 shows the overall pipeline of our optimization.

## 4.2. Initializing and Densifying Gaussians

To represent the background and humans via 3D-GSs, we initialize each representation via the available cues. The background Gaussians $\mathcal{G}^{BG}$ are initialized with point cloud obtained by Structure-from-Motion (SfM) [44]. In the cases of a fixed camera input where SfM cannot be applicable, we assume that the background is a large 3D sphere with background texture, centered on the mean position of humans. We represent a $j$-th human via 3D-GS in a canonical space (denoted as subscript $c$), $\mathcal{G}_{j,c}^h$, which is initialized by the vertices of A-posed SMPL mesh $\mathcal{V}(\boldsymbol{\beta}_j, \boldsymbol{\theta_c})$, where $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta_c}$ are the SMPL shape and canonical pose parameters respectively, regressed using a monocular 3D pose regressor [38, 47]. The color and opacity are set in grey and $0.9$ respectively. We assume the SMPL shape parameter $\beta_j$ from the pose regressor is fixed for each human while training.

To capture the fine details of the background and human, we densify the initial Gaussians adaptively [19] every $N_{den}$ iteration. The Gaussians to be densified are chosen based on the accumulated gradients $\nabla \boldsymbol{\mu}_i$, by summing the gradient on the center of Gaussians $\boldsymbol{\mu}_i$ computed in each iteration. If the accumulated gradients $\sum_{N_{den}} \nabla \boldsymbol{\mu}_i$ is bigger than a predefined threshold, we densify the Gaussian $G_i$ by cloning or splitting it.

## 4.3. Canonicalizing Dynamic Humans

To fuse the cues of an individual with different poses across different frames, we model each person with a single canonical model $\mathcal{G}_c^h$ (for brevity, we drop the person index $j$ in the subscript). We also model the deformation function $\mathcal{G}_d^h(t) = F_d^h(\mathcal{G}_c^h, \boldsymbol{\theta}_t)$ which transforms Gaussians in canonical into posed $\mathcal{G}_d^h(t)$ at time $t$, following the SMPL pose parameter $\boldsymbol{\theta}_t$. Our deformation function $F_d^h : \mathbb{R}^3 \rightarrow SE(3)$, which maps canonical space into posed space, is defined via the Linear Blend Skinning (LBS) based on the SMPL [27]. It translates the center of Gaussian $\boldsymbol{x}_i$ and rotates the covariance matrix $\boldsymbol{\Sigma}_i$ as follows:

$$\boldsymbol{x}_{i,p} = \sum_{k=1}^{N_{joint}} w_k(\boldsymbol{x}_{i,c})(\boldsymbol{R}_k \boldsymbol{x}_{i,c} + \boldsymbol{T}_k), \tag{6}$$

$$\boldsymbol{\Sigma}_{i,p} = \boldsymbol{R}_{wei} \boldsymbol{\Sigma}_{i,c} \boldsymbol{R}_{wei}^T, \tag{7}$$

where $\boldsymbol{R}_k$ and $\boldsymbol{T}_k$ are rotation and translation of $k^{th}$ joint of SMPL skeleton which is computed from $\boldsymbol{\theta}$ and $\boldsymbol{\beta}_j$. The $\boldsymbol{R}_{wei}$ is a derivation of LBS equal to the weighted sum of rotations $\{\boldsymbol{R}_k\}_{k=1}^{N_{joint}}$ as follows:

$$\boldsymbol{R}_{wei} = \sum_{k=1}^{N_{joint}} w_k(\boldsymbol{x}_{i,c}) \boldsymbol{R}_k. \tag{8}$$

The skinning weight $w_k(\boldsymbol{x})$ is calculated from the aligned SMPL vertices defined in canonical space. Here we get

$w_k(\boldsymbol{x})$ by summing the skinning weight of the 30 nearest vertices with Inverse Distance Weight (IDW). To accelerate rendering speed, we pre-calculate and store the skinning weight in a voxel grid, similar to SelfRecon [15]. In each rendering time, we obtain the skinning weight by trilinear interpolation on the weight grid instead of searching the nearest SMPL vertices.

### 4.4. Diffusion-Guided Reconstruction

We employ the 2D diffusion prior [40] in optimizing 3D-GS to represent a human model, as a key idea to overcome sparse observations for the target human in input videos. The intuition behind our approach is that the quality of a 3D human model can be measured by assessing the realism of the rendered images in novel views. For example, naively optimizing 3D Gaussians $\mathcal{G}$ from sparse and occluded views results in artifacts or missing parts, as shown in Fig. 3 (a). The use of the diffusion model, particularly the SDS loss [37], can be beneficial to improving the quality of the desired 3D model by enforcing realism in the rendered images at novel views. The SDS loss can be considered as additional virtual cameras guided by the pre-trained 2D diffusion model [40].

For each iteration, we make rendering $R_v^h$ of the target person's 3D-GS $\mathcal{G}^h$ with a virtual camera $v$ which is randomly sampled from a sphere that is centered on each person and viewing the full body of target person. To give more diversity for the rendering, we also randomly sample the body pose $\theta$ of the person and transform the 3D-GS $\mathcal{G}^h$ into the posed space. We randomly sample $\boldsymbol{\theta}$ either among observed poses or the canonical A-pose($\boldsymbol{\theta}_c$), which can be written as $\{\boldsymbol{\theta}_t\}_{t\in[1\ldots T]} \cup \{\boldsymbol{\theta}_c\}$. With the rendered image $R_v^h$ at viewpoint $v$, we compute the SDS loss [37], which is proportional to the difference between added noise $\epsilon$ and estimated noise $\epsilon_\phi$ by diffusion $\phi$:

$$\nabla\mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\Big[\omega(\tau)(\epsilon_\phi(\boldsymbol{z}_\tau;\boldsymbol{y},\tau)-\epsilon)\frac{\partial \boldsymbol{z}}{\partial R_v^h}\frac{\partial R_v^h}{\partial \mathcal{G}^h}\Big] \quad (9)$$

, where $\boldsymbol{z}_\tau$ is a noised latent of rendering $R_v^h$, $\tau \in [0,1]$ is a time-step of noise and $\boldsymbol{y}$ is conditions applied on diffusion. The SDS loss mitigates the artifacts in our 3D-GS human model by enforcing the rendered output $R_v^h$ in a novel pose to be plausible.

**Textual Inversion on SDS.** We further improve the efficacy of our SDS method by leveraging the concept of Texture Inversion (TI) [10], as a way to make the SDS loss to synthesize the human appearance similar to our target identity, rather than generating arbitrary appearance. Applying the SDS loss only with text prompts, such as "A photo of a person", may easily converge to a random human appearance due to the diversity of diffusion prior model [40]. Ideally, we want to specify the target individual observed from our input images, to encourage the diffusion model to synthesize the consistent human appearance at the virtual viewpoints. To incorporate this idea, we leverage the Textual Inversion (TI) by finding the text-embedding specific to our target human [10, 42]. Here, we use CustomDiffusion [21] to invert observations of the target individual into the text token `<person-j>`, where we collect the images of the target human from input frames by cropping and masking the target person only. Together with Textual Inversion, CustomDiffusion fine-tunes the attention and text embedding layer of the diffusion $\phi$ which makes a person-specific diffusion $\phi_j$. By adding the inverted text-token to the text-prompt, we can get diffusion-generated images consistent with the observations. We perform the diffusion fine-tuning and Textual Inversions per person separately and apply the subject-specific fine-tuned version of SDS for each target person.

Furthermore, we also utilize the OpenPose ControlNet [60] to align the body pose of a person generated by diffusion and our person Gaussians $\mathcal{G}^h$. For this, when we compute the SDS loss, we project the 3D SMPL joints $J_{smpl}(\boldsymbol{\theta}, \boldsymbol{\beta}_j)$ into the viewpoint, convert to OpenPose [4] format, and query them into the ControlNet. We additionally add a view-augmented language prompt [37] for stable optimization. We sample the noise time-step $\tau$ from $\mathcal{U}[0.2, 0.98]$ for the first 2000 iterations and then we decay the maximum time-step from 0.98 to 0.3 for 2000 iterations. Also to enhance the fine details of the body, we randomly sample camera $v_j$ which zooms in the face, lower body, and upper body. Refer to our supp. mat. for more details.

### 4.5. Training Objectives

For every iteration, we render the image $R_t$ corresponding to frame $t$ and calculate MSE loss, SSIM loss, and LPIPS loss by comparing it with the ground truth image $I_t$:

$$\begin{aligned}\mathcal{L}_{recon} = \lambda_{rgb}\text{MSE}(R_t, I_t) + \lambda_{ssim}\text{SSIM}(R_t, I_t) \\ + \lambda_{lpips}\text{LPIPS}(R_t, I_t). \quad (10)\end{aligned}$$

Then we compute SDS Loss $\mathcal{L}_{SDS}$ for each person's Gaussians $\mathcal{G}_j^h$ with person-specific diffusion $\phi_j$:

$$\mathcal{L}_{sds} = \sum_{j=1}^{N}\mathcal{L}_{SDS}(\mathcal{G}_j^h, \phi_j). \quad (11)$$

To avoid transparent artifacts, we additionally add hard-surface regularization on human rendering following LOL-NeRF [39]:

$$\mathcal{L}_{hard} = -\log(\exp^{-|\alpha|} + \exp^{|\alpha|}) + \text{const}. \quad (12)$$

, where $\alpha$ is the rendered alpha map of person Gaussian $\mathcal{G}_j^h$. Our final training objective is as follows:

$$\mathcal{L}_{tot} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{sds}\mathcal{L}_{sds} + \lambda_{hard}\mathcal{L}_{hard}. \quad (13)$$

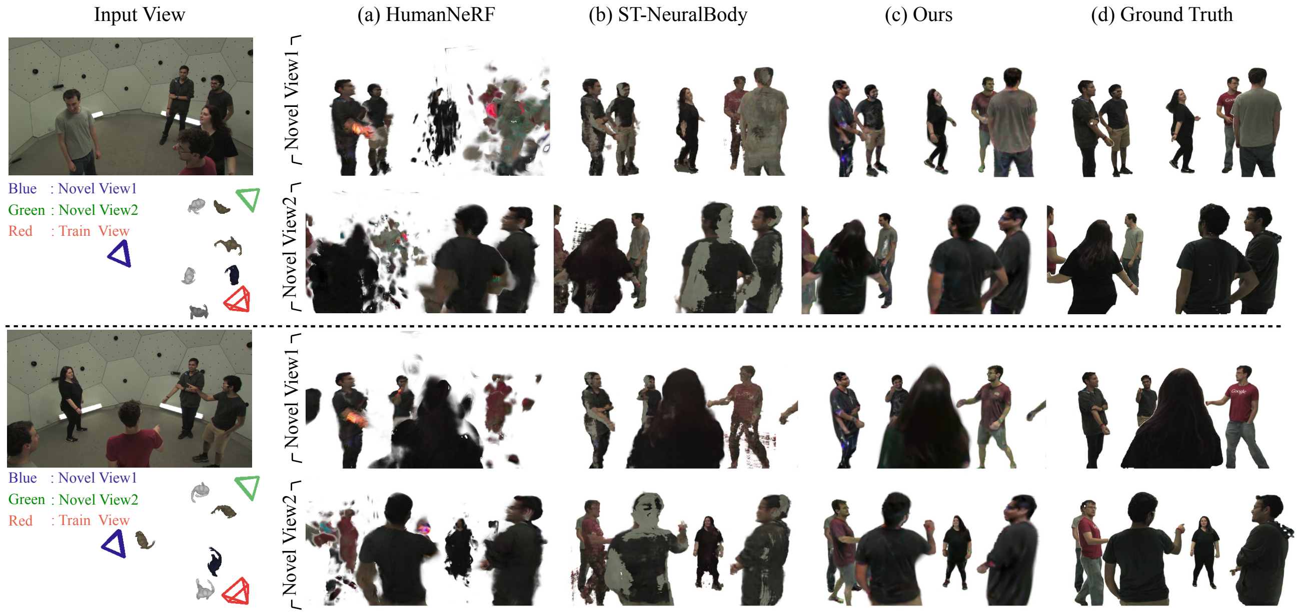



Figure 4. **Novel view synthesized results of Panoptic Dataset [18]**

## 5. Experiments

We perform rigorous quantitative and qualitative evaluations to show the strengths of our method. We first test our method on challenging scenes capturing multiple people with sparse 2D observations, to show the major advantage of our method. We also apply our method to existing single human reconstruction benchmarks. Additionally, by performing ablation studies, we demonstrate the efficacy of each module within our pipeline. We also show multiple qualitative experiments on challenging in-the-wild videos.

### 5.1. Dataset

**Panoptic Dataset [18]** This dataset captures socially interacting multiple individuals via a multi-view camera system. We simulate the sparse 2D view scenario by choosing a single camera view as the input for the reconstruction and using other views for GT views. We use an ultimatum sequence with 6 people (540 frames in ultimatum 160422 sequence), and choose an HD camera view in a common view angle as the input video. We select other 7 HD cameras in diverse novel viewpoints as the GTs for the evaluations, as shown in Fig. 4. As a pre-processing, we fit SMPL models on the provided pseudo-GT 3D skeletons of the dataset using a pose-prior [3]. See the supp. mat. for more details of processing.

**Hi4D [57]** This dataset contains multiple individuals at close distances, which are captured with synchronized 8 cameras. We consider challenging two sequences `pair00-dance` and `pair01-hug`. Similar to Panoptic DB, we choose the video from a camera (camera 76) as input where only a single side of the people is visible and use all other 7 views as novel GT views for the evaluation. We use the provided pseudo-GT SMPL parameters.

**Single Human Benchmarks** We also conduct experiments on an existing single human benchmark ZJU-Mocap [36] dataset, to check the performance of our model on the single human reconstruction task with sufficient observations. We follow the evaluation pipeline used in baselines [16, 51].

### 5.2. Baseline and Evaluation Metrics

We compare our model with three methods, **HumanNeRF** [51], **InstantAvatar** [16] and **Shuai et al.** [46]. **HumanNeRF** [51] shows SOTA quality on the monocular human reconstruction task. As HumanNeRF cannot handle multiple people at once, we optimize it separately for each person and merge them in the final evaluation. We get a rendering of each individual separately and accumulate them in an $\alpha$ blending manner. The order of accumulation is determined by the distance of the SMPL pelvis from the center of the camera. Because HumanNeRF requires a foreground mask for processing, we use GT masks if available (Hi4D and ZJU-Mocap), or use an off-the-shelf method [20]. **InstantAvatar** [16] is the most efficient method to reconstruct a human from a monocular video with high quality. We evaluate the InstantAvatar using the same pipeline applied for HumanNeRF evaluation, as it's only capable of mono-human cases. **Shuai et al.** [46] is a compositional method similar to ours, which optimizes implicit functions of people and background from sparse view input [46]. It represents each person with NeuralBody [36] and background with NeRF [31] and renders them together by using the compositional rendering pipeline of ST-NeRF [59]. Different from the original paper, we optimize it using only a single train view.

| Methods | PSNR↑ | SSIM↑ | LPIPS* ↓ |
|---|---|---|---|
| HumanNeRF [51] | 19.59 | 0.6514 | 38.69 |
| InstantAvatar [16] | 15.03 | 0.4163 | 65.95 |
| *Shuai et al.* [46] | 15.79 | **0.8370** | 25.77 |
| Ours | **23.60** | 0.8323 | **25.41** |

Table 1. **Quantitative results on Panoptic dataset.** LPIPS* = 100 × LPIPS. Our method shows better performance in PSNR and LPIPS and comparable results in SSIM.

| Method | `pair00-dance` | | `pair01-hug` | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| HumanNeRF [51] | 18.79 | 0.8552 | 21.40 | 0.8238 |
| InstantAvatar [16] | 18.60 | 0.8646 | 19.07 | 0.8254 |
| *Shuai et al. [46]* | 20.78 | 0.9165 | 19.72 | 0.9078 |
| Ours | **23.76** | **0.9328** | **25.14** | **0.9289** |

Table 2. **Quantitative results on Hi4D dataset [57].** Ours show better performance on both PSNR and SSIM in situations when people closely interact like hugging or dancing together

We compare the quality of human rendering by masking out the background of GT images. For evaluation metric, we use peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and perceptual similarity (LPIPS) [61] following the prior work [51].

### 5.3. Implementation Details

To stabilize optimization, we first train background Gaussians $\mathcal{G}^{BG}$ solely with background images where people are masked out. We then optimize human Gaussians $\mathcal{G}_{p_i,c}$ without $\mathcal{L}_{SDS}$ for the first 1000 iterations and then, optimize human Gaussians $\mathcal{G}_c^{p_i}$ and background Gaussians $\mathcal{G}_{bg}$ simultaneously together with $\mathcal{L}_{SDS}$ for 10k iterations. Refer to our supp. mat. for more details.

### 5.4. Evaluations on Multiple People Reconstruction

**Panoptic dataset.** We show the qualitative comparison between our results and the outputs of baselines in Fig. 4. As shown, our method reconstructs the appearances of people even in the presence of severe occlusions and image cropping. In contrast, both HumanNeRF and *Shuai et al.* fail to reconstruct details showing noticeable artifacts since many body parts are not visible in the input view. In particular, HumanNeRF suffers from severe occlusions because it requires accurate foreground masks containing whole human shapes, which cannot be obtained due to the occluder in front of the target individual. In contrast, our method does not suffer from such issues.

We also quantify the output qualities in Tab. 1 by rendering the reconstructed scenes into unseen novel views. As shown in the table, our method outperforms the baselines in PSNR and LPIPS.

| Method | PSNR↑ | SSIM↑ | LPIPS*↓ |
|---|---|---|---|
| HumanNeRF [51] | 30.23 | 0.9554 | 3.36 |
| InstantAvatar [16] | 28.55 | 0.9282 | 10.60 |
| Ours *wo* $\mathcal{L}_{SDS}$ | 30.10 | 0.9529 | 4.68 |
| Ours | 30.13 | 0.9535 | 4.50 |

Table 3. **Quantitative results on 6 subjects in ZJU-Mocap dataset [36].** LPIPS* = 100 × LPIPS. Our method shows comparable quality compared to HumanNeRF [51] even if the observation is sufficient.

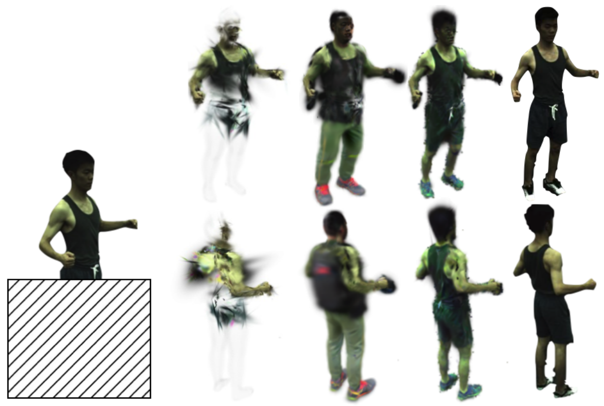

(a) HumanNeRF (b) Ours (c) GT (d) Camera Position

Figure 5. **Novel view synthesis output of Hi4D *pair00-dance* sequence.** While HumanNeRF [51] fails to reconstruct a face, ours synthesizes a plausible face guided by diffusion model [40]. (d) plots camera position relative to the front viewing female body. As shown here, the majority of the rendered output has been never observed in the train view.

**Hi4D dataset.** We show the quantitative results in Tab. 2 and qualitative examples in Fig. 5 from `pair00-dance` sequence. As shown in Tab. 2, our method outperforms baselines in all metrics. Interestingly, in this dataset, some body parts of the individual are never observed in the input view due to the severe occlusions, e.g., the face of the female person in Fig. 5, where our method "hallucinates" realistic human face without any observations.

### 5.5. Evaluations on Single Person Reconstruction

We show the quantitative comparisons on ZJU-Mocap [36] dataset. As shown in Tab. 3, our method shows comparable performance over baselines when sufficient 2D observations are available. This result demonstrates that our pipeline based on 3D-GS with SDS loss does not negatively affect the single-person reconstruction scenarios while showing its major strengths in the case of sparse observations.



Input (a) wo/ $\mathcal{L}_{SDS}$ (b) wo/ TI (c) Full GT

Figure 6. **Ablation studies.** From (a) shows that using only reconstruction loss suffers from artifacts (like hedgehogs) even in shown regions. (b) shows that textual inversion is essential to generate contextually similar appearances on unseen regions.

| w/ Textual Inversion | w/ $\mathcal{L}_{SDS}$ | w/ $\mathcal{L}_{recon}$ | PSNR↑ | SSIM↑ | LPIPS* ↓ |
|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | **26.03** | **0.9435** | **7.00** |
| | ✓ | ✓ | 23.43 | 0.9342 | 8.36 |
| | | ✓ | 24.51 | 0.9323 | 10.54 |

Table 4. **Ablation Studies with lower-body occluded ZJU-Mocap [36] 377 subject.** LPIPS* = 100×LPIPS. As shown, SDS loss with textual inversion token shows the biggest improvement.

### 5.6. Ablation Studies

We conduct an ablation study to demonstrate the importance of the proposed modules of our framework. As a way for the quantitative evaluations, we use ZJU-Mocap [36] dataset and simulate a challenging scenario by (1) completely occluding the lower body, and (2) using a few frames only for the input (10% of frames from the ZJU-Mocap 377 subject). An example of the artificially occluded images is shown in Fig. 6 with test results.

Our *Full* method can successfully reconstruct the whole part of the humans. Interestingly, it also synthesizes the completely unseen short pants and shoes of the target individual, while the appearance and colors are different from the GT. As expected, the output without $\mathcal{L}_{SDS}$ fails to reconstruct the unseen lower part, and also shows poor quality on the upper body due to the insufficient image frames. The output without Texture Inversion (TI) reconstructs the unseen lower parts as well, but, interestingly, the output appearance is very different from the GT. This result directly demonstrates the importance of our Texture Inversion process in applying SDS loss, helpful in preserving the identity of the target individual. We also show the quantification results in Tab. 4, where the importance of each module is also clearly demonstrated.

| Method | Rendering Speed (FPS ↑) | | VRAM ↓ |
|---|---|---|---|
| | 512 × 512 | 1024 × 1024 | |
| InstantAvatar [16] | 18.03 | 7.09 | 4972MiB |
| Ours (15k) | **361.01** | **277.01** | **1496MiB** |

Table 5. **Novel pose rendering speed.** We compared novel pose rendering speed between InstantAvatar [16] and ours with ZJU-Mocap [36] 377 subject. It shows that our method consists of a 15k Gaussians renders faster on both low and high resolutions, with less VRAM consumption.

### 5.7. Rendering Efficiency

We show the computational efficiency of our method by comparing the rendering time of novel pose synthesis to InstantAvatar [16], which is known as the most efficient existing method and also the fastest among our competitors. All reported times here are measured with a single GeForce RTX 3090 GPU. By taking advantage of 3D-GS representations, our method achieves more than real-time rendering speed with 300 FPS on 1K×1K images. As shown in Tab. 5, our method surpasses InstantAvatar [16] both in rendering speed and memory consumption while showing better rendering quality as shown in Tab. 3.

## 6. Discussion

In this paper, we present a method to reconstruct the world and dynamically moving humans in 3D from a monocular video input, particularly focusing on sparse and limited observation scenarios. We represent both the world and multiple humans via 3D Gaussian Splatting representation, enabling us to conveniently and efficiently compose and render them together. We also introduce a novel approach to optimize the 3D-GS representation in a canonical space by fusing the sparse cues in the common space, where we leverage a pre-trained 2D diffusion model to synthesize unseen views by keeping the consistency with the observed 2D appearances. Via thorough experiments, we demonstrate the high performance and efficiency of our method in various challenging examples.

Our approach, however, still has limitations such as: (1) SMPL fitting needs to be provided; (2) our method only considers humans as the dynamic target, ignoring animals, cars, or other dynamic objects; (3) the quality of the synthesized parts are still limited with visible artifacts. All these limitations can be exciting future research directions.

**Acknowledgements** This work was supported by Samsung Electronics C-Lab, NRF grant funded by the Korea government (MSIT) (No. 2022R1A2C2092724 and No. RS-2023-00218601), and IITP grant funded by the Korean government (MSIT) (No.2021-0-01343). H. Joo is the corresponding author.

## References


[1] Ijaz Akhter, Tomas Simon, Sohaib Khan, Iain Matthews, and Yaser Sheikh. Bilinear spatiotemporal basis models. *TOG*, 2012. 2

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

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 6

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 5

[5] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, 2023. 2

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 2

[7] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *TPAMI*, 2023. 2

[8] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 2

[9] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023. 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2, 5

[11] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR*, 2023. 2, 3, 4

[12] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[14] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *3DV*, 2024. 2

[15] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 2, 5

[16] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 2, 6, 7, 8

[17] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 2, 3

[18] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017. 6

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023. 2, 3, 4

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 6

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 4, 5

[22] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation. In *CVPR*, 2022. 3

[23] Jiabao Lei, Jiapeng Tang, and Kui Jia. Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In *CVPR*, 2023. 2

[24] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *3DV*, 2024. 2

[25] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 2

[26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *TOG*, 2015. 1, 2, 4

[28] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 4

[30] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023. 2

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 6

[32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *SIGGRAPH*, 2022. 2

[33] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *NeurIPS*, 2020. 2

[34] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 2, 3

[35] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 3d reconstruction of a moving point from a series of 2d projections. In *ECCV*, 2010. 2

[36] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 3, 6, 7, 8

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 2, 3, 5

[38] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location & pose. In *CVPR*, 2022. 4

[39] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *CVPR*, 2022. 5

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 5, 7

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 5

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[44] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4

[45] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, 2022. 2

[46] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH*, 2022. 3, 6, 7

[47] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 4

[48] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 2

[49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 2

[50] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2

[51] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 2, 4, 6, 7

[52] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D^ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. In *NeurIPS*, 2022. 3

[53] Tiange Xiang, Adam Sun, Jiajun Wu, Ehsan Adeli, and Li Fei-Fei. Rendering humans from object-occluded monocular videos. In *ICCV*, 2023. 2

[54] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 2

[55] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. In *SIGGRAPH*, 2018. 2

[56] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2

[57] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, 2023. 6, 7

[58] Zhengming Yu, Wei Cheng, xian Liu, Wayne Wu, and Kwan-Yee Lin. MonoHuman: Animatable human neural field from monocular video. In *CVPR*, 2023. 2

[59] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. In *SIGGRAPH*, 2021. 2, 3, 6

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5

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

[62] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2

[63] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS '01.*, 2001. 3