OccFusion: Rendering Occluded Humans with Generative Diffusion Priors

Adam Sun*, Tiange Xiang* †, Scott Delp, Li Fei-Fei‡, Ehsan Adeli‡
Stanford University

https://cs.stanford.edu/~xtiange/projects/occfusion/

May 20, 2025

Figure 1: Reconstructing humans from monocular videos frequently fails under occlusion. In this paper, we introduce **OccFusion**, a method that combines 3D Gaussian splatting with 2D diffusion priors for modeling occluded humans. Our method outperforms the state-of-the-art in rendering quality and efficiency, resulting in clean and complete renderings free of artifacts.

Abstract

Most existing human rendering methods require every part of the human to be fully visible throughout the input video. However, this assumption does not hold in real-life settings where obstructions are common, resulting in only partial visibility of the human. Considering this, we present OccFusion, an approach that utilizes efficient 3D Gaussian splatting supervised by pretrained 2D diffusion models for efficient and high-fidelity human rendering. We propose a pipeline consisting of three stages. In the Initialization stage, complete human masks are generated from partial visibility masks. In the Optimization stage, 3D human

*Equal contribution; junior author listed first.
†Correspondence to xtiange@stanford.edu.
‡Equal mentorship.

Gaussians are optimized with additional supervision by Score-Distillation Sampling (SDS) to create a complete geometry of the human. Finally, in the Refinement stage, in-context inpainting is designed to further improve rendering quality on the less observed human body parts. We evaluate OccFusion on ZJU-MoCap and challenging OcMotion sequences and find that it achieves state-of-the-art performance in the rendering of occluded humans.

1 Introduction

Rendering 3D humans from monocular in-the-wild videos has been a persistent challenge, with significant implications in virtual/augmented reality, healthcare, and sports. Given a video of a human moving around a scene, this task involves reconstructing the appearance and geometry of the human, allowing for the rendering of the human from novel views.

When faced with the problem of human reconstruction from monocular video, several works based on neural radiance fields (NeRFs) have achieved promising results [35, 54, 18, 8]. 3D Gaussian splatting [23] further improves upon NeRF-based rendering methods for better performance. By representing the human not as an implicit radiance field but as a set of explicit 3D Gaussians, methods like GauHuman [13] and 3DGS-Avatar [44] are able to render humans comparable in quality to NeRF methods while taking only a few minutes to train and less than a second to render.

Rendering occluded humans is a relatively new yet critically important problem. Most human rendering literature assumes that humans are in clean environments free from occlusions. However, in real-world scenes such as hospitals, sports stadiums, and construction sites, humans may frequently be occluded by all kinds of obstacles. Unfortunately, most existing human rendering methods are not able to handle such challenging real-world scenarios, producing a lot of undesirable floaters, artifacts, and incomplete body parts. On the other hand, methods proposed to address human rendering under occlusions like OccNeRF [61] and Wild2Avatar [60] are limited and impractical due to their high computational costs and long training time. This tradeoff between efficiency and rendering quality greatly limits the applicability of past approaches.

In this work, we introduce OccFusion, an efficient yet high quality method for rendering occluded humans. To gain improved training and rendering speed, OccFusion represents the human as a set of 3D Gaussians. To ensure complete and high-quality renderings under occlusion, OccFusion proposes to utilize generative diffusion priors to aid in the reconstruction process. Like almost all other human rendering methods, OccFusion assumes accurate priors such as human segmentation masks and poses are provided for each frame, which can be obtained with state-of-the-art off-the-shelf estimator [61, 60, 64].

Our approach consists of three stages: **(1)** *The Initialization Stage:* we utilize segmentation and pose priors to inpaint occluded human visibility masks into complete human occupancy masks to supervise later stages. **(2)** *The*

Optimization Stage: we initialize a set of 3D Gaussians and optimize them based on observed regions of the human, applying pose-conditioned Score-Distillation Sampling (SDS) to help ensure completeness of the modeled human body in both the posed and canonical space. **(3)** *The Refinement Stage:* we utilize pretrained generative models to inpaint unobserved regions of the human with context from partial observations and renderings from the previous stage, further improving the quality of the renderings. Despite taking only 10 minutes to train, our method outperforms the state-of-the-art in rendering humans from occluded videos.

In summary, our contributions are: ***(i)*** We propose OccFusion, the first method to combine Gaussian splatting with diffusion priors for the rendering of occluded humans from monocular videos. Multiple novel components are proposed along with a three-stage pipeline, consisting of Initialization, Optimization, and Refinement stages. ***(ii)*** We demonstrate that OccFusion achieves state-of-the-art efficiency and rendering quality of occluded humans on both simulated and real-world occlusions.

2 Related Work

2.1 Neural Human Rendering

Traditional methods to reconstruct humans usually require dense arrays of cameras [9, 3, 2] or depth information [68, 47, 3, 4], both of which are unobtainable for in-the-wild scenes. To solve this problem, Neural Radiance Fields (NeRFs) [35] have recently been used to model dynamic humans from monocular videos [54, 8, 18, 16, 69, 49]. These methods achieve high-quality novel view synthesis by parametrizing the human body using an SMPL [34] pose prior and modeling it as a radiance field. However, since NeRFs depend on large Multi-Layer Perceptrons (MLPs), they are computationally expensive, usually taking days to train and minutes to render [23, 13, 44]. To speed up NeRF-based models, multi-resolution hash encoding [38, 41, 5, 17], and generalizability [39, 2, 12, 26] have been proposed. However, these methods either face a rendering bottleneck [13] or an expensive pre-training process, both of which affect their efficiency.

Point-based rendering methods like 3D Gaussian splatting [23] greatly accelerate the rendering of static and dynamic scenes. Recently, there have been an abundance of works applying 3D Gaussian splatting to human rendering tasks [44, 13, 25, 36, 71, 32, 65, 19, 28, 27, 40, 11]. Like NeRF-based approaches, Gaussian splatting-based approaches represent the human in a canonical space and use Linear Blend Skinning (LBS) to transform the human into the posed space. Gaussian splatting methods achieve state-of-the-art performance of dynamic humans with fast training times and real-time rendering, causing them to be the more desired method [44, 13].

2.2 Occluded Human Rendering

Rendering humans in occluded settings is a relatively new problem. Sun *et al.* [48] utilize a layer-wise scene decoupling model to decouple humans from occluding objects. OccNeRF [61] combines geometric and visibility priors with surface-based rendering to train a human NeRF model. Wild2Avatar [60] proposes an occlusion-aware scene parametrization scheme to decouple the human from the background and occlusions. While these works provide decent renderings of humans free of occlusions, they are slow and impractical. A concurrent work to ours is OccGaussian [64], which also proposes to model occluded humans with 3D Gaussians by performing an occlusion feature query in occluded regions. We provide comparisons to their published results in Table 1.

2.3 Generative Diffusion Priors

Inferring the appearance of unobserved regions of 3D scenes requires the usage of generative models. The recent success of 2D diffusion models has made them the preferred model to use for generation [51, 20, 33, 45, 31]. To lift 2D diffusion models for 3D content generation, DreamFusion [43] proposed Score Distillation Sampling (SDS), a commonly used method for utilizing a pre-trained 2D diffusion model to supervise 3D content generation [29, 67, 50, 52].

Diffusion models can also be used as priors for training NeRFs and Gaussian splatting, combining reconstruction with generation [58, 74, 75, 59, 63, 70]. ReconFusion [58] uses SDS in conjunction with multi-view conditioning to synthesize the appearance of unobserved regions of a scene from sparse views, while BAGS [74] utilizes SDS to supervise a Gaussian splatting model.

3 Preliminaries

Before we introduce our method, we first overview important basics of 3D human modeling with SMPL (subsection 3.1). Then, we discuss 3D Gaussian splatting, and how it can be applied to human modeling (subsection 3.2). Finally, we propose OccGauHuman, a simple improvement of GauHuman [13] that is better designed for occluded human rendering (subsection 3.3)

3.1 3D Human Modeling

SMPL [34] is a model that parametrizes the human body with a 3D surface mesh. To transform between the canonical space to a pose space, the Linear Blend Skinning (LBS) algorithm is used. Given a 3D point $\mathbf{x_c}$ in the canonical space and the shape β and pose θ parameters of the human, a point in the posed space can be calculated as:

$$\mathbf{x_p} = \sum_{k=1}^{K} w_k \left(G_k(\mathbf{J}, \theta)\mathbf{x_c} + b_k(\mathbf{J}, \theta, \beta) \right), \tag{1}$$

where J contains K joint locations, G_k and b_k are the transformation matrix and translation vector, and $w_k \in [0, 1]$ are a set of skinning weights. The SMPL representation is commonly used as a geometric prior for human rendering [61, 60, 44, 13, 54, 18, 69].

3.2 Human Rendering with 3D Gaussian Splatting

3D Gaussian splatting. 3D Gaussian splatting [23] models a scene as a set of 3D Gaussians Π. Each Gaussian is defined by its 3D location $\mathbf{p_i}$, opacity $o_i \in [0, 1]$, center $\mu_\mathbf{i}$, covariance matrix Σ_i, and spherical harmonic coefficients. The i-th Gaussian is defined as $o_i e^{-\frac{1}{2}(\mathbf{p}-\mu_i)^T \Sigma_i^{-1}(\mathbf{p}-\mu_i)}$. During rendering, these 3D Gaussians are mapped from the 3D world space and projected to the 2D image space via α-blending, with the color of each pixel being calculated across the N Gaussians as:

$$C = \sum_{j=1}^{N} c_j \alpha_j \prod_{k=1}^{j-1} (1 - \alpha_k), \tag{2}$$

where c_j is the color and α is the z-depth ordered opacity. During the training process, Gaussians are adaptively controlled via densification (splitting and cloning) and pruning until they achieve the optimal density to adequately represent the scene.

GauHuman [13] . In the line of work that uses 3D Gaussian splatting for human rendering [44, 25, 28, 11], GauHuman is a representative approach due to its balance of efficiency and rendering quality. After initializing Gaussians on the vertices of the SMPL mesh, GauHuman learns a representation of the human in canonical space and utilizes LBS to transform each individual Gaussian into the posed space. A pose refinement module $MLP_{\Phi_{pose}}$ and an LBS weight field module $MLP_{\Phi_{lbs}}$ are used to learn the LBS transformation, and a merge operation based on KL divergence is used along with splitting, cloning, and pruning to help the Gaussians reach convergence.

We base our method on GauHuman due to its fast training and state-of-the-art representative ability. GauHuman's code is distributed under the S-Lab license and can be accessed here.

3.3 OccGauHuman: A Simple Upgrade for GauHuman for Occlusion Handling

In common human rendering tasks, videos are captured in a clean environment, with every pixel in the image belonging to either the human or the background. By using a semantic segmentation model such as SAM [24] to preprocess a video, we can train the human rendering model only on pixels labeled as "human". However, occlusions in the videos may lead to sparse observations of the human. As a result, fitting NeRF-based human rendering models on only the visible human pixels results in an incomplete geometry with lots of artifacts [61, 60].

Figure 2: **OccFusion** achieves occluded human rendering via three sequential stages. In the **Initialization Stage**, we recover complete binary human masks $\{\hat{\mathbf{M}}\}$ from occluded partial observations $\{\mathbf{I}\}$ with the help of segmentation priors $\{\mathbf{M}\}$ and pose priors $\{\mathbf{P}\}$. $\{\hat{\mathbf{M}}\}$ will be further used to help optimize human Gaussians Π in subsequent stages. In the **Optimization Stage**, we apply $\{\mathbf{P}\}$ conditioned SDS on both posed human and canonical human to enforce the human occupancy to remain complete. In the **Refinement Stage**, we use the coarse human renderings $\{\hat{\mathbf{I}}\}$ from the Optimization Stage to help generate missing RGB values in $\{\mathbf{I}\}$ through our proposed in-context inpainting. Through this process, both the appearance and geometry of the human are fine-tuned to be in high fidelity. Training of all three stages takes only **10 minutes** on a single Titan RTX GPU.

Gaussian splatting-based rendering models [23] are especially suitable for human modeling tasks due to their explicit geometry and point-based representation. In this section, we present three straightforward tweaks of GauHuman [13] to make it perform better on videos with occlusions: *(1)* Firstly, as discussed above, we train the model on visible human pixels only. *(2)* We adjust the loss weights to put more weight on the mask loss computed between rendered human occupancy maps and the segmentation masks — we found that this helps render more crisp human boundaries. *(3)* We disable the densification and pruning of Gaussians during training — this helps maintain a rather complete human geometry based on the SMPL initialization. Benefits brought by our updates compared to the original GauHuman are presented in Table 1 and Figure 7.

4 OccFusion

In our approach, we train a Gaussian splatting-based human rendering model on the visible pixels of a human. However, recovering occluded content for a dynamically moving human is not trivial — humans are usually in challenging poses, and complex occlusions can cause additional issues. It is also essential to preserve consistency of the human appearance and geometry across different frames. Considering these challenges, we propose our method OccFusion in multiple separate stages. In the Initialization stage (section 4.1), we inpaint occluded binary human masks for more reliable geometric guidance. In the

Optimization stage (section 4.2), we use the inpainted masks to train a human rendering model based on GauHuman [13] while using Score Distillation Sampling (SDS) constraints on both the posed space and canonical space. In the Refinement stage (section 4.3), we fine-tune the trained model from the Optimization Stage with in-context inpainting to further refine the appearance of the human. An overview of our OccFusion is shown in Figure 2.

4.1 Initialization Stage: Recovering Human Geometry from Partial Observations

Generative diffusion models [45] have demonstrated promise to be used as priors for different tasks [21, 50]. The most straightforward method for occluded human rendering is to utilize a precomputed segmentation prior $\mathbf{M}$ and pose prior $\mathbf{P}$ to condition a diffusion prior Φ [37, 72] to inpaint $1 - \mathbf{M}$ — the image regions that are not occupied by the human. However, there are two significant barriers to such a straightforward approach, as articulated below.

Conditioned human generation cannot handle challenging poses. It is true that a conditioned diffusion prior Φ is able to generate detailed images while staying consistent to the condition. However, since diffusion models are usually overfitted on more commonly seen poses, Φ usually fails to generate reasonable images when conditioned on challenging poses (see Figure 3 middle column). We attribute this limitation to the inappropriate 2D representation of $\mathbf{P}$ — when joints self-occlude each other, it is impossible to tell which joints are closest to the camera when they are projected to 2D. So, we propose to simplify the 2D representation of $\mathbf{P}$. We apply a Z-buffer test on the depth map rendered from the SMPL mesh [34] and then calculate the distance d between its z-axis location and the corresponding 2D z-buffer. Given a pre-defined threshold σ, we deem a joint is self-occluded if $d > \sigma$. Self-occluded joints are ignored when projecting 3D joints onto the 2D canvas for conditioning Φ (see Figure 3 right column). Our simplification improves the generation quality of Φ for challenging poses.

Figure 3: Stable Diffusion 1.5 generations [45] conditioned on a challenging pose $\mathbf{P}$. While conditioning on the original pose results in multiple limbs and other abnormalities, our method of simplifying pose by removing self-occluded joints results in more feasible generations.

Per-frame inpainting cannot guarantee cross-frame consistency. Compared to image generation models, video generation models [10, 57, 6] are less

accessible and much more expensive to run. Without an explicit modeling of object motion in the video, frame-by-frame generation with an image generative model leads to cross-frame inconsistency, which is not desirable for human reconstruction (see Figure 4 middle column). Instead of inpainting the occluded parts of the human directly with Φ, we claim that it is more feasible to inpaint binary human masks since small variations in the human silhouette are more acceptable (see Figure 4 right column). We first inpaint the RGB image $\mathbf{I}$ and then rely on an off-the-shelf segmentation model [22] to obtain the inpainted binary human masks $\{\hat{\mathbf{M}}\}$, which is used to assist the training of the rendering model in subsequent stages.

Figure 4: While generative models provide inconsistent inpainting results, the binary masks that can be extracted from these generated images are much more consistent.

4.2 Optimization Stage: Enforcing Human Completeness with SDS Regularization

After obtaining the inpainted masks $\{\hat{\mathbf{M}}\}$ that outline a reasonable human silhouette, we build a Gaussian splatting model similar to the one described in section 3.2 for human rendering. The 3D Gaussians Π are initiated as the SMPL mesh vertices, which are able to be deformed to adapt to different poses through SMPL-based LBS (Equation 3.1). With the help of $\{\hat{\mathbf{M}}\}$, the training of Π consists of multiple photometric loss terms $\mathcal{L}_{photo}$:

$$\lambda_{rgb}L_1(\mathbf{M}\cdot\mathbf{I}, \mathbf{M}\cdot\mathbf{I}')+\lambda_{mask}L_2(\hat{\mathbf{M}}, \mathbf{A})+\lambda_{ssim}\texttt{SSIM}(\mathbf{M}\cdot\mathbf{I}, \mathbf{M}\cdot\mathbf{I}')+\lambda_{lpips}\texttt{LPIPS}(\mathbf{M}\cdot\mathbf{I}, \mathbf{M}\cdot\mathbf{I}'), \tag{3}$$

where L_1 is the L-1 loss, L_2 is the L-2 loss, $\texttt{SSIM}(\cdot)$ is the SSIM function [53], $\texttt{LPIPS}$ is the VGG-based perceptual loss [73], $\mathbf{I}'$ is the rendered image from Π, and $\mathbf{A}$ is the rendered human occupancy map. Each of the loss terms is scaled by a weight hyperparameter λ.

Even with the supervision of $\{\hat{\mathbf{M}}\}$, geometry inconsistency still exists. Although inconsistent human masks affect the training of Π much less than inconsistent images, human completeness cannot be guaranteed without further steps.

Using diffusion priors to enforce human completeness. We build off of the insights from [50, 56, 67] and apply Score Distillation Sampling (SDS) [43] to improve the quality of human renderings and reduce artifacts. Instead of applying SDS on RGB images $\mathbf{I}'$, which causes appearance inconsistency, we apply it directly to the rendered human occupancy maps $\mathbf{A}$ so that diffusion scores are propagated to encourage complete $\mathbf{A}$:

$$\mathcal{L}_{\text{SDS}}^{(\mathbf{P})} = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi(\mathbf{A}; t, \mathbf{P}) - \epsilon\right)\frac{\partial \mathbf{A}}{\partial \Pi}\right], \tag{4}$$

where t is a scheduled time stamp, $w(\cdot)$ is a weighting function, $\epsilon(\cdot)$ is the UNet noise estimator in Φ, and ϵ is the injected Gaussian noise.

Using diffusion priors to regularize canonical pose. In-the-wild videos often involve very sparse observations of the human, with only incomplete regions of the human visible in each frame. To further enforce completeness, we propose to render the human in the canonical Da-pose $\hat{\mathbf{P}}$ with the human oriented at a random angle $\in \{k\frac{\pi}{9}, k \in \mathbb{Z}\}$. Applying SDS on the canonical renderings serves as regularization and is randomly activated during training. Overall, at each training step in the Optimization Stage, the Gaussians Π are optimized towards:

$$\nabla_\Pi\left[\mathcal{L}_{photo} + \rho \cdot \lambda_{pose}\mathcal{L}_{\text{SDS}}^{(\mathbf{P})} + (1-\rho)\cdot\lambda_{can}\mathcal{L}_{\text{SDS}}^{(\hat{\mathbf{P}})}\right], \tag{5}$$

where ρ is a random variable that has a 75% chance to be 1 and 0 otherwise. With this formulation, the Optimization stage results in a complete and coherent geometry regardless of the viewing angle.

4.3 Refinement Stage: Refining Human Appearance via In-context Inpainting

As shown in Figure 6 Exp. C and D, applying diffusion priors on rendered human occupancy maps is not able to recover the missing appearances of the human. This motivates the need for a subsequent stage that keeps refining Π for better appearance.

The refinement of the appearance of 3D objects is not a new topic [50, 30, 67]. However, no existing generative models are capable of handling the consistency of appearance of a human across different frames and poses. We attribute this difficulty to the denoising process used in generative priors — random noise is injected to rendering at each SDS step which leads to uncertain results. This is infeasible for reconstruction tasks, which require frame-consistent representations that agree with all observations.

Our approach involves generating inpainted images of the occluded human offline to use as references. We first identify the occluded regions to be inpainted $\mathbf{R}$ by using the rendered human occupancy masks $\mathbf{A}$ from the Optimization Stage and pre-computed human visibility masks $\mathbf{M}$: $\mathbf{R} = (1 - \mathbf{M}) \cdot \mathbf{A}$. In order to encourage the generated regions to be more consistent with the partial observations, we propose in-context references inspired by in-context learning

Table 1: Quantitative comparison on the ZJU-MoCap and OcMotion datasets. LPIPS values are scaled by $\times 1000$. We color cells that have the best and second best metric values.

Methods	ZJU-MoCap			OcMotion		
	PSNR↑	SSIM↑	LPIPS↓	PSNR*↑	SSIM*↑	LPIPS*↓
HumanNeRF	20.67‡	0.9509‡	-	-	-	-
3DGS-Avatar	17.29†	0.9410†	63.25†	9.788†	0.7203†	188.1†
GauHuman [1]	21.55	0.9430	55.88	15.09	0.8525	107.1
OccNeRF [2]	22.40‡	0.9562‡	43.01‡	15.71	0.8523	82.90
OccGaussian	23.29‡	0.9482‡	41.93‡	-	-	-
Wild2Avatar [3]	-	-	-	14.09§	0.8484§	93.31§
OccGauHuman	22.71	0.9492	54.60	18.85	0.8863	86.53
OccFusion	23.96	0.9548	32.34	18.28	0.8875	82.42

* Metrics calculated on **visible pixels** only.
† Model trained for 5k iterations with $\times 3$ **training time**.
‡ Results taken from OccGaussian, using $\times 5$ **training frames**.
§ Model trained under the default setting using $\times 2$ **training frames**.

of language models [1]. Although renderings from the Optimization Stage lack sharp and high-fidelity details, they resemble complete human geometries and possess good enough features that can be used as a coarse reference to guide Φ to inpaint similar contents at occluded body regions. To achieve this, we stack $\hat{\mathbf{I}}$ and $\mathbf{I}$ together as a single image input to Φ with an additional prompt phrase — "the same person standing in two different rooms".

We use the inpainted RGB images $\{\tilde{\mathbf{I}}\}$ along with other priors to finetune Π via photometric losses. Since diffusion models still tend to be somewhat inconsistent, we smooth training by putting more weight on perceptual loss terms and use L1 loss for the pixel-wise loss terms for its high robustness to variance:

$$\nabla_{\Pi}\left[\lambda_{rgb}L_1(\mathbf{M}\cdot\mathbf{I},\mathbf{M}\cdot\mathbf{I}')+\lambda_{mask}L_2(\hat{\mathbf{M}},\mathbf{A})+\lambda_{gen}L_1(\tilde{\mathbf{I}},\mathbf{R}\cdot\mathbf{I}')+\lambda_{lpips}\texttt{LPIPS}(\mathbf{I},\mathbf{I}')\right]. \tag{6}$$

We train our entire pipeline for *only 10 minutes on a single TITAN RTX GPU.*

5 Experiments

In this section, we conduct quantitative and qualitative evaluation of our approach against state-of-the-art methods. Then, we conduct ablation studies of our entire pipeline, demonstrating that each stage is necessary for optimal performance.

5.1 Datasets and Evaluation

ZJUMoCap. ZJU-MoCap [42] is a dataset consisting of 6 dynamic humans captured with a synchronized multi-camera system. Since the humans are in a

lab environment free of occlusions, we follow OccNeRF's [61] protocol to simulate occlusion of the human, masking out the center 50% of the human pixels for the first 80 % of frames. To challenge OccFusion on videos with even sparser frames, we use only 100 **frames** from the first camera with a sampling rate of 5 to train the models and use the other 22 cameras for evaluation.

OcMotion. OcMotion [14] comprises of 48 videos of humans interacting with real objects in indoor environments. Experiments are conducted on the same 6 sequences adopted by Wild2Avatar [60], which are selected to provide a diverse coverage of real-world occlusions. We form sparser sub-sequences by sampling only 50 frames to train the models.

Evaluation. We compare our OccFusion to OccNeRF [61], OccGaussian [64], and Wild2Avatar [60], the state-of-the-art in occluded human rendering. We also compare our results to GauHuman [13], HumanNeRF [54], and 3DGS-Avatar [44], popular human rendering methods not designed for occlusion. For fairness of comparison, all methods use the same set of segmentation masks and pose priors. We evaluate the methods both quantitatively and qualitatively. For our quantitative evaluations, we calculate the Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) metrics against the ground truth images. Since no ground truth is provided for OcMotion, we calculate the metrics on visible pixels only. For qualitative evaluations, we render the human from novel views and assess the quality of the renderings.

5.2 Results on Simulated and Real-world Occlusions

We provide quantitative metrics averaged over all the sequences in Table 1. Overall, methods designed for occluded human rendering tend to outperform their traditional counterparts. Among those methods, OccFusion consistently performs up to par or better than the state-of-the-art on both datasets while significantly beating all the baselines on LPIPS.

Qualitative results on novel view synthesis can be found in Figure 5. OccNeRF [61] has trouble generating unseen regions and renders significant discoloration and floaters when faced with occlusion. OccGauHuman's renderings are blurry and occasionally incomplete. We observe that OccFusion is the only method to consistently render sharp and high-quality renderings free of occlusions.

5.3 Implementation Details

OccFusion requires several priors. We run SAM [24] to get all the human masks $\{\mathbf{M}\}$. While we follow previous work [61, 60] and use the ground truth poses provided by ZJU-MoCap and OcMotion, pose priors $\mathbf{P}$ can be obtained via occlusion-robust SMPL prediction/optimization methods such as HMR 2.0 [7] and SLAHMR [66] for in-the-wild videos. We use the pre-trained Stable Diffusion 1.5 model [45] with ControlNet [72] plugins for content generation in all the stages. Improving the quality of priors is not the focus of this work.

Table 2: Ablation results on the ZJU-MoCap [42] dataset. LPIPS values are scaled by $\times 1000$.

Exp.	Methods	PSNR↑	SSIM↑	LPIPS↓	Train time
-	GauHuman [13]	21.55	0.9430	55.88	10 mins
A	OccGauHuman	22.54	0.9457	54.88	2 mins
B	+ *Init Stage* generated masks $\{\hat{\mathbf{M}}\}$	23.52	0.9516	52.35	5 mins
C	+ Posed space SDS	23.90	0.9510	55.47	7 mins
D	+ Canonical space SDS (*Optim Stage*)	23.91	0.9514	55.35	7 mins
E	+ *Refinement Stage*	**23.96**	**0.9548**	**32.34**	10 mins

In the *Initialization Stage*, instead of inpainting incomplete human masks directly, we run the pretrained diffusion model to inpaint RGB images with 10 inference steps and 1.0 ControlNet conditioning scale. We use the positive prompt — *"clean background, high contrast to background, a person only, plain clothes, simple clothes, natural body, natural limbs, no texts, no overlay"* and the negative prompt — *""multiple objects, occlusions, complex pattern, fancy clothes, longbody, lowres, bad anatomy, bad hands, bad feet, missing fingers, cropped, worst quality, low quality, blurry"*. After inpainting the RGB images, we then run SAM-HQ [22] with $\mathbf{P}$ as the prompts to get $\{\hat{\mathbf{M}}\}$.

In the *Optimization Stage*, we train the 3D human Gaussian Π from scratch by following the objective Equation 5. We set $\lambda_{rgb} = 1e^4$, $\lambda_{mask} = 2e^4$, $\lambda_{ssim} = 1e^3$, and $\lambda_{lpips} = 1e^3$. At each training step, we random switch the SDS regularization on either posed human space or the canonical Da-pose space with a probability of 75% and 25%. When applying SDS regularization on the canonical human space, we rotate the human by θ radians around its vertical axis, where θ is uniformly sampled from $\{k\frac{\pi}{9}, k \in \mathbb{N}\}$. We set the SDS loss weights as $\lambda_{pose} = 2e^5$ and $\lambda_{can} = 2e^5$. We use the same negative prompt but no positive prompt. In this stage, we train Π for 1200 steps.

In the *Refinement Stage*, we first generate the RGB human inpaintings via the proposed in-context inpainting method. We condition the pretrained diffusion model on $\mathbf{M}$, with 10 inference steps and 0.3 ControlNet conditioning scale. We do not use positive prompts for inpainting but use the same negative prompts as in the *Optimization Stage*. During training, we set the loss weights as $\lambda_{rgb} = 1$ and $\lambda_{mask} = 0.1$, $\lambda_{gen} = 0.1$, and $\lambda_{lpips} = 0.2$. In this stage, we finetune Π for another 1800 steps with Gaussian densification and pruning enabled for the first 1000 steps.

5.4 Ablation Studies

In this section, we study the effect of each of our proposed components by adding them one by one and report average metrics on ZJU-MoCap in Table 2. Each stage plays a part towards optimal performance. Qualitative results on our ablations are included in Figure 6. We can see that the Initialization Stage helps

enforce completeness for the initially incomplete human. The SDS regularization provided in the Optimization Stage helps remove floaters and artifacts in the posed and canonical space, further improving the shape of the human. Finally, the Refinement Stage helps make the renderings more detailed in less observed regions, greatly improving rendering quality.

5.5 Additional Studies

Does the proposed OccGauHuman perform better than GauHuman [13] in rendering occluded humans? In section 3.3, we present a simple upgrade for the state-of-the-art 3DGS based human rendering model GauHuman [13] to help it better handle occlusions. Our improvements are straightforward but effective. We show quantitative results in Figure 1 (Left) and Table 1. Here we present further qualitative comparisons in Figure 7 on both datasets to validate the improvements. As shown in the figure, our improved OccGauHuman reconstructs a more complete human body than the vanilla GauHuman. Our final model OccFusion yields the best rendering quality with both complete geometry as well as high fidelity appearance.

Can existing generative models recover an occluded human? While there are works for using generative diffusion models to render 3D humans conditioned on single [55] and multiple [46] images, none are able to condition on a monocular video of the person.

Since [55] has not released code, we include results from InstantMesh [62]. We use the provided segmentation mask to mask the least occluded frame onto a white background and use it as conditioning. Novel view synthesis results are included in Figure 8. InstantMesh is unable to recover a complete human geometry and fails to generate reasonable appearance from the single image.

Effectiveness of in-context inpainting. We provide comparisons of the human in the Refinement Stage with and without in-context inpainting and provide qualitative comparisons in Figure 9. While renderings from the Optimization stage are less detailed in occluded areas, they imply sufficient guidance for further inpainting. By using them as references, our proposed in-context inpainting is able to generate the missing content and greatly increase the rendering quality in these areas.

6 Discussions and Conclusion

Limitations. Recovering occluded dynamic humans is challenging. As mentioned in section 4.3, reconstructing a 3D human requires adhering to multiple consistencies. However, even with the state-of-the-art generative models, it is still impossible to perfectly maintain those consistencies for 4D content (3D + motion) generation. Although our proposed methods are specifically designed to eliminate potential variances when using generative priors, we can still observe some generations are less coherent (e.g. Figure 4 and Figure 9), which may hurt the training of the rendering model on all stages. Moreover, we found

that conditioning generative models with 2D poses is weak — the pose of the generated human does not always align with the condition pose, which may introduce even more uncertainty for training. In future work, we hope to train our own consistency-aware diffusion model specifically finetuned on human data.

Societal Impacts. Being able to reconstruct a human from an occluded monocular video can have a great societal impact. For example, having a high-fidelity 3D reconstruction of a human can help telemedicine practitioners become more immersed in the 3D space. While our research could lead to privacy concerns if humans are reconstructed without their consent, we believe that the benefits can be harnessed responsibly with appropriate safeguards.

Conclusion. In this work, we propose OccFusion, one of the first works that utilize 3D Gaussian splatting for occluded human rendering. Our approach consists of three stages: the Initialization, Optimization, and Refinement stages. By combining the efficiency and representative ability of 3D Gaussian splatting with the generation capabilities of diffusion priors, our method achieves state-of-the-art in occluded human rendering quality as measured by the PSNR, SSIM, and LPIPS metrics while only taking around 10 minutes to train. We hope our work inspires further exploration into the capabilities of diffusion priors to aid in human reconstruction.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *European Conference on Computer Vision*, pages 222–239. Springer, 2022.

[3] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.

[4] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

[5] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023.

[6] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023.

[9] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019.

[10] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

[11] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. *arXiv preprint arXiv:2312.02134*, 2023.

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

[13] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:2312.02973*, 2023.

[14] Buzhen Huang, Yuan Shu, Jingyi Ju, and Yangang Wang. Occluded human body capture with self-supervised spatial-temporal motion prior. *arXiv preprint arXiv:2207.05375*, 2022.

[15] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022.

[17] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023.

[18] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022.

[19] HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir Navab, and Benjamin Busam. Deformable 3d gaussian splatting for animatable human avatars. *arXiv preprint arXiv:2312.15059*, 2023.

[20] Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22976–22985, 2023.

[21] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023.

[22] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[25] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023.

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

[27] Mengtian Li, Shengxiang Yao, Zhifeng Xie, Keyu Chen, and Yu-Gang Jiang. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting. *arXiv preprint arXiv:2401.09720*, 2024.

[28] Mingwei Li, Jiachen Tao, Zongxin Yang, and Yi Yang. Human101: Training 100+ fps human gaussians in 100s from 1 view. *arXiv preprint arXiv:2312.15258*, 2023.

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[30] Yuanze Lin, Ronald Clark, and Philip Torr. Dreampolisher: Towards high-quality text-to-3d generation via geometric diffusion. *arXiv preprint arXiv:2403.17237*, 2024.

[31] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.

[32] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. *arXiv preprint arXiv:2311.16482*, 2023.

[33] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

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

[36] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. *arXiv preprint arXiv:2311.17113*, 2023.

[37] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[39] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. Transhuman: A transformer-based human representation for generalizable neural human rendering. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3544–3555, 2023.

[40] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. *arXiv preprint arXiv:2312.05941*, 2023.

[41] Bo Peng, Jun Hu, Jingtao Zhou, Xuan Gao, and Juyong Zhang. Intrinsicngp: Intrinsic coordinate based hash encoding for human nerf. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[43] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[44] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. *arXiv preprint arXiv:2312.09228*, 2023.

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

[46] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *European Conference on Computer Vision*, pages 702–720. Springer, 2022.

[47] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 246–264. Springer, 2020.

[48] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4651âĂŞ4660, New York, NY, USA, 2021. Association for Computing Machinery.

[49] Wenzhang Sun, Yunlong Che, Han Huang, and Yandong Guo. Neural reconstruction of relightable human model from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–407, 2023.

[50] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[51] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023.

[52] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.

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

[54] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022.

[55] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Single-view 3d human digitalization with large reconstruction models. *arXiv preprint arXiv:2401.12175*, 2024.

[56] Zhenzhen Weng, Zeyu Wang, and Serena Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image. *arXiv preprint arXiv:2305.16411*, 2023.

[57] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[58] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981*, 2023.

[59] Jamie Wynn and Daniyar Turmukhambetov. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4180–4189, 2023.

[60] Tiange Xiang, Adam Sun, Scott Delp, Kazuki Kozuka, Li Fei-Fei, and Ehsan Adeli. Wild2avatar: Rendering humans behind occlusions. *arXiv preprint arXiv:2401.00431*, 2023.

[61] Tiange Xiang, Adam Sun, Jiajun Wu, Ehsan Adeli, and Li Fei-Fei. Rendering humans from object-occluded monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3239–3250, 2023.

[62] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[63] Xiaofeng Yang, Yiwen Chen, Cheng Chen, Chi Zhang, Yi Xu, Xulei Yang, Fayao Liu, and Guosheng Lin. Learn to optimize denoising scores for 3d generation: A unified and improved diffusion prior on nerf and 3d gaussian splatting. *arXiv preprint arXiv:2312.04820*, 2023.

[64] Jingrui Ye, Zongkai Zhang, Yujiao Jiang, Qingmin Liao, Wenming Yang, and Zongqing Lu. Occgaussian: 3d gaussian splatting for occluded human rendering. *arXiv preprint arXiv:2404.08449*, 2024.

[65] Keyang Ye, Tianjia Shao, and Kun Zhou. Animatable 3d gaussians for high-fidelity synthesis of human motions. *arXiv preprint arXiv:2311.13404*, 2023.

[66] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21232, 2023.

[67] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models, 2024.

[68] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021.

[69] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023.

[70] Zhongrui Yu, Haoran Wang, Jinze Yang, Hanzhang Wang, Zeke Xie, Yunfeng Cai, Jiale Cao, Zhong Ji, and Mingming Sun. Sgd: Street view synthesis with gaussian splatting and diffusion prior. *arXiv preprint arXiv:2403.20079*, 2024.

[71] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. *arXiv preprint arXiv:2312.11461*, 2023.

[72] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[73] 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.

[74] Tingyang Zhang, Qingzhe Gao, Weiyu Li, Libin Liu, and Baoquan Chen. Bags: Building animatable gaussian splatting from a monocular video with diffusion priors. *arXiv preprint arXiv:2403.11427*, 2024.

[75] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023.

Figure 5: Qualitative comparisons on **simulated occlusions** in the ZJU-MoCap dataset [42] (left column) and **real-world occlusions** in the OcMotion dataset [15] (right column). ON denotes OccNeRF [61] and OGH denotes OccGauHuman.

Figure 6: Qualitative ablation studies. Please see Table 2 for corresponding experiments. Major differences are highlighted by red arrows.

Figure 7: Qualitative comparisons on **simulated occlusions** in the ZJU-MoCap dataset [42] (left column) and **real-world occlusions** in the OcMotion dataset [15] (right column). GH denotes GauHuman [13] and OGH denotes OccGauHuman.

Figure 8: Novel view synthesis results from InstantMesh [62] conditioned on the least occluded frame. Discrepancies are circled in red.

Figure 9: Comparison of the inpainted human in the Refinement Stage with and without using the proposed in-context inpainting technique. Major differences are highlighted with red arrows.