

---

# HumanCrafter: Synergizing Generalizable Human Reconstruction and Semantic 3D Segmentation

---

Panwang Pan<sup>1\*†</sup>, Tingting Shen<sup>2\*</sup>, Chenxin Li<sup>3\*</sup>, Yunlong Lin<sup>2</sup>,  
Kairun Wen<sup>2</sup>, Jingjing Zhao<sup>1</sup>, Yixuan Yuan<sup>3†</sup>

\* Equal contribution † Corresponding author

<sup>1</sup>ByteDance, <sup>2</sup>Xiamen University, <sup>3</sup>CUHK

<https://paulpanwang.github.io/HumanCrafter>

## Abstract

Recent advances in generative models have achieved high-fidelity in 3D human reconstruction, yet their utility for specific tasks (e.g., human 3D segmentation) remains constrained. We propose HUMANCRAFTER, a unified framework that enables the joint modeling of appearance and human-part semantics from a single image in a feed-forward manner. Specifically, we integrate human geometric priors in the reconstruction stage and self-supervised semantic priors in the segmentation stage. To address labeled 3D human datasets scarcity, we further develop an interactive annotation procedure for generating high-quality data-label pairs. Our pixel-aligned aggregation enables cross-task synergy, while the multi-task objective simultaneously optimizes texture modeling fidelity and semantic consistency. Extensive experiments demonstrate that HUMANCRAFTER surpasses existing state-of-the-art methods in both 3D human-part segmentation and 3D human reconstruction **from a single image**.

## 1 Introduction

Reconstructing high-fidelity 3D human representations and understanding human body and clothing attributes are a fundamental challenge in the 3D vision realm. The philosophy can unlock many novel and practical downstream applications, such as semantic-guided 3D reasoning, context-aware human behavior analysis, and interactive semantic editing. This capability further facilitates immersive augmented and virtual reality (AR/VR), character stylization, and cinematic production.

In light of advances in Neural Radiance Fields [1], previous attempts [2, 3, 4] have synthesized high-quality human novel views. Nevertheless, implicit representations are typically computationally intensive, as they rely on dense point querying in 3D space. Recent advances in 3D Gaussian Splatting (3DGS) [5] have provided real-time rendering for reconstructing high-quality explicit human models, which however rely on dense multi-view images [6, 7] or monocular video input [8, 9, 10] and time-consuming per-subject optimization processes, limiting their stability and feasibility in downstream applications. With the emergence of large reconstruction models [11], recent advances can directly generalize the regression of 3D representations [12, 13, 14, 15] thanks to the prevalence of large-scale 3D object datasets [16]. However, in the specific task of human reconstruction, these works collapse and fail to produce faithful and consistent novel views, primarily due to the scarcity of 3D human datasets and a lack of human prior knowledge as inductive bias in model design. Recent pioneering studies on human reconstruction [17, 18] enable generalizable and robust synthesis under sparse-view settings. However, the crucial requirement for semantic 3D segmentation is currently hindered by the lack of expansive, well-labeled 3D human downstream task datasets. One workaround is to first reconstruct and then leverage recent advances in 2D human visual foundation models [19]. This paradigm can result in extensive processing times and substantial engineering efforts. Furthermore,

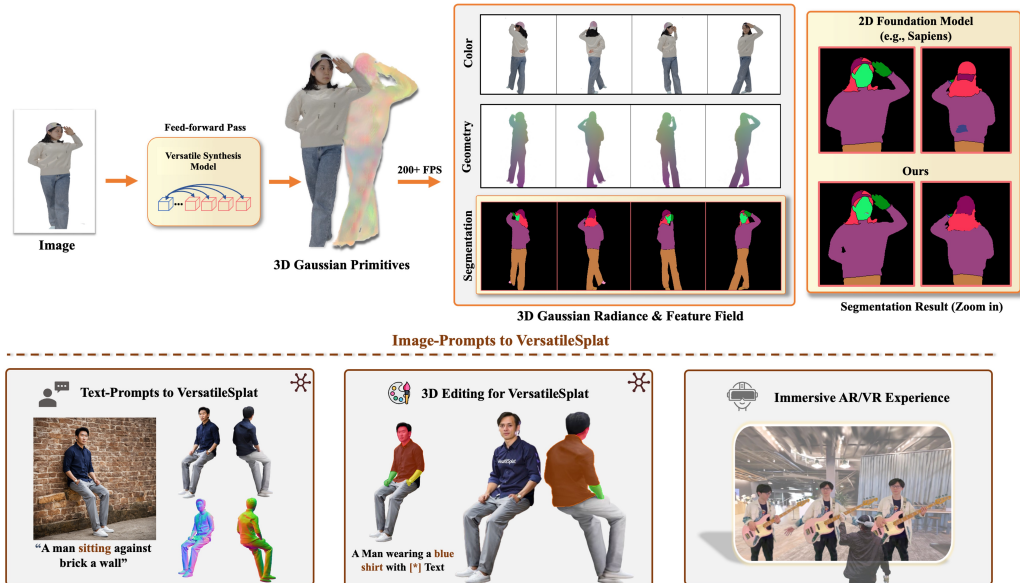


Figure 1: We introduce HUMANCRAFTER, a unified framework for simultaneous human 3D reconstruction and body-part segmentation from single images. HUMANCRAFTER introduces explicit 3D Gaussian VersatileSplats, showcasing enhanced performance over foundation models in delivering 3D-consistent segmentation outcomes. This breakthrough offers significant advantages for downstream applications.

2D operations cannot maintain 3D consistency and coherence across different viewpoints. The two-stage pipeline faces challenges including inconsistencies, high computational costs, and scalability issues, which hinder its robustness and efficiency. We hold the belief that *the best way to understand something is to reconstruct it*.

We introduce a unified synthesis model that revolutionizes 3D reconstruction and editing through innovative view synthesis and semantic understanding. Our model takes a single image as input and generates explicit 3D Gaussian Splats enriched with semantic features, enabling human 3D reconstruction and body-part segmentation, while seamlessly enabling the 3D editing task and integrating into VR devices, as illustrated in Figure 1. Our approach builds upon the incorporation of tailored human priors and the aggregation of multi-view features with camera embeddings using Transformers. Specifically, we translate the set of aggregated features to pixel-aligned 3D Gaussians as initialized geometry. We unleash a pre-trained 2D model [20] into a 3D consistent feature field and establish a weighting mechanism to propagate into multi-view, addressing the scarcity of labeled 3D semantic data. Finally, by jointly training on the constructed 3D segmentation datasets, which consist of 40,000 images from 2,500 human scans, HUMANCRAFTER enables novel view rendering and segmentation to mutually benefit from each other’s task. In summary, our contributions can be summarized as follows:

- We are first to introduce a unified 3D human representation and a holistic framework that addresses versatile novel-view synthesis in a feed-forward pass, allowing two independent tasks to mutually benefit.
- HUMANCRAFTER leverages geometric human priors and the attention mechanism to effectively bypass labor-intensive and computationally expensive steps, establishing a new paradigm in the realm of human foundation models.
- Experiments demonstrate that HUMANCRAFTER exhibits superior photorealistic 3D human Reconstruction and human-part segmentation capabilities, surpassing many state-of-the-art baselines simultaneously with real-time rendering.
- Experiments on in-the-wild images demonstrating HUMANCRAFTER’s strong generalizability to diverse image, and facilitating the potential real-world applications.

## 2 Related Work

### 2.1 Synergizing Reasoning with 3D Reconstruction.

2D human-centric models [21, 22, 19] exhibit remarkable performance through the advent of 2D foundation models and vision transformers [23, 20] and curated datasets (e.g., Humans-300M [19]). Yet, these cutting-edge methods cannot achieve simultaneous 3D modeling and coherent segmentation due to lack 3D constraints. A crucial requirement for 3D reasoning is currently hindered by the paucity of extensive, well-annotated multi-view image datasets. Conversely, early studies like Semantic NeRF [24] and Panoptic Lifting [25] successfully embedded segmentation network semantic data into 3D scenes. 3DGS with Feature Field [26, 27, 28, 29] emerged as a prominent joint training approach for 3DGS and multiple prediction tasks. Recent works [30, 31] enable training-free 2D feature lifting to 3D paradigm for large scenes. However, the human-specific domain remains largely unexplored, despite its potential applications in areas such as human editing, gaming, and film production.

### 2.2 Human Gaussian Splatting.

Recent advances [6, 32, 7] optimize photo-realistic animatable 3DGS from temporal multi-view images. In particular, HiFi4G [33] combines 3DGS with a dual-graph mechanism to maintain spatial-temporal coherence, ASH [34] employs mesh UV parameterization for real-time rendering, and Animatable Gaussians [35] utilizes StyleUNet [36] and 3DGS for high-fidelity animatable avatars. For monocular video input [8, 37, 9, 38, 39, 40, 41, 42, 43, 44], human templates and Linear Blend Skinning are commonly adopted, necessitating per-instance optimization of 3DGS in a canonical space. For sparse-view scenarios, GPS-Gaussian [17] estimates depth maps from two-view stereo and unprojects them to pixel-wise 3D Gaussians. [45] further introduces a regularization term and an epipolar attention mechanism to preserve geometry consistency between source views. EVA-Gaussian [46] employs a recurrent feature refiner to address artifacts. [18] utilizes human templates for multi-scaffold photorealistic and accurate view rendering. For single image input, [47] incorporates generative diffusion models to predict triplane NeRF and followed by a feed-forward reconstruction model. Recent advances [48, 49, 50, 51] predicts the 3D outputs from a single input image in a generalizable manner through the combination of 2D Diffusion and well-conceived human priors. Notably, Human-3Diffusion [51] introduces a groundbreaking single-image-to-3DGS framework that synergizes diffusion models with 3D reconstruction to create highly consistent 3D avatars. Our work further enhances the synergy between 3D reconstruction and semantic segmentation. Our unified pipeline utilizes efficient multi-view geometric aggregation and a refined Transformer module to significantly improve accuracy, real-time rendering, and multi-task consistency.

### 2.3 Generalizable Reconstruction Transformer.

GINA-3D, LRM and TripoSR [52, 53, 54] demonstrate that feed-forward Transformers trained on large-scale datasets [16, 55] are capable of generating 3D models in a generalizable manner by reconstructing triplane features for volumetric rendering. Real3D [56] further proposes a self-training reconstruction framework that capitalizes on both synthetic and large-scale real-world datasets. Instant3D [57] proposes a two-stage pipeline, which first generates a sparse set of four posed images, and then directly regresses the Neural Radiance Fields (NeRF) [58]. [59] integrates efficient 3DGS with triplane representations, where point clouds inferred from input images query triplane features to decode 3DGS attributes. CRM, MeshLRM, and InstantMesh [60, 61, 62] replace the triplane NeRF with FlexiCubes [63] as outputs, incorporating differentiable mesh extraction and rendering for potential applications. LGM, GRM, and GS-LRM [64, 65, 66] predict pixel-wise 3DGS parameters [67, 68] from multiple images, enabling scalable reconstruction frameworks. However, such methods lack human-specific priors as tailored inductive biases in Transformers, failing to synthesize faithful novel views, especially reconstructing human body details.

## 3 Method

**Preliminary of 3D Gaussians Splatting.** 3DGS [5] formulates the 3D representation as a collection of  $N$  Gaussian primitives  $\{\mathcal{G}_p\}_{p=1}^N$ . Each  $\mathcal{G}_p$  is characterized by an opacity  $\sigma_p \in \mathbb{R}$ , a mean location  $\mu_p \in \mathbb{R}^3$ , a scaling factor  $s_p \in \mathbb{R}^3$ , an orientation quaternion  $q_p \in \mathbb{R}^4$ , and a color feature  $c_p \in \mathbb{R}^C$

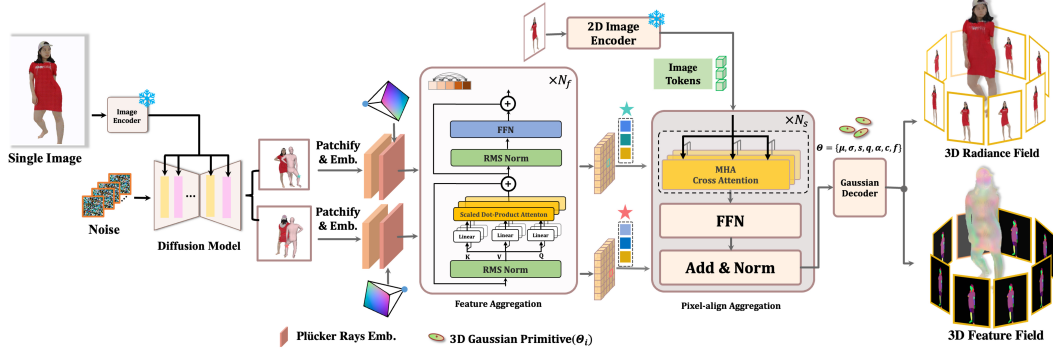


Figure 2: **The network architecture of HUMANCRAFTER.** The proposed method fully utilizes 2D diffusion priors and human body geometry features to regress pixel-aligned point maps via a generic Transformer (Sec. 3.1). Subsequently, another Transformer (Sec. 3.2) employs an attention mechanism to produce a set of semantic 3D Gaussians that encapsulate geometric, appearance, and semantic information. The entire pipeline is trained in an end-to-end manner by minimizing a loss function (Sec. 3.3) that compares the predicted outputs against ground truth data and rasterized label maps from novel viewpoints.

are maintained for rendering, where spherical harmonics (SH) can model view-dependent effects. Specifically, each Gaussian can be formulated as:  $\mathbf{G}_p(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p)\right)$ .

The rendering process involves projecting the 3D Gaussians onto the image plane as 2D Gaussians and performing alpha blending for each pixel in front-to-back depth order, thereby determining the final color, opacity, and depth maps.

**Pipeline Overview.** Given a single RGB image  $\mathbf{I} \in \mathbb{R}^{\mathbf{H} \times \mathbf{W} \times 3}$ , our objective is to jointly synthesize multiple properties at novel views using the multi-task labels from the given source views without any optimization or fine-tuning. To accommodate versatile labels, we extend the 3D Gaussians  $\mathbf{G}_p$  to construct a feature field, enabling the rendering of feature maps through a differentiable rasterizer.

To this end, we firstly propose a purely Transformer (Sec. 3.1) to aggregate multi-view input images, followed by a attention mechanism (Sec. 3.2) to facilitate downstream tasks. Furthermore, we propose a multi-task loss (Sec. 3.3) to achieve high-fidelity texture rendering. An overview of the model architecture is illustrated in Figure 2.

### 3.1 Human Prior for Feature Aggregation

**Image Diffusion Prior.** We leverage the pre-trained 2D diffusion model SV3D [69] as appearance prior. For the input image  $\mathbf{I}_0$ , we leverage a CLIP image encoder [70] to obtain the image embedding  $\mathbf{c}$  as conditions. Subsequently, we progressively denoise gaussian noises into temporal-continuous  $\mathbf{N}$  multi-view images by a spatial-temporal UNet  $D_\theta$  [71]. In this work, we also employ the SMPL [72] as the 3D human parametric model to guide the Feature Aggregation. Inspired by [73, 74], we render the side-view normal images as a guide, which are then concatenated with the input image and corresponding Plücker embeddings [75] along the channel dimension, resulting in dense pose-conditioned images. These pose-conditioned images are divided into non-overlapping patches [23] and mapped to  $d_1$  dimensional patch tokens  $\mathbf{F}_i \in \mathbb{R}^{(h \times w) \times d_1}$  by a Linear layer, where  $h = \mathbf{H}/\mathbf{P}$  and  $w = \mathbf{W}/\mathbf{P}$  represent the height and width of the feature map, respectively, and  $\mathbf{P}$  denotes the patch size.

**Cross-view Attention Module.** For feature interaction within patch tokens, we utilize  $\mathbf{N}_f$  layers of Grouped Query Attention blocks [76] with RMS Pre-Normalization, GELU activation, and feed-forward network. These features are subsequently mapped into the location of 3D Gaussian Splatting. To accurately model the positioning of Gaussians, we predict the depth map  $\mathbf{D}_i \in \mathbb{R}^{\mathbf{H} \times \mathbf{W}}$  for each input image and additional 3D positional offset  $\boldsymbol{\Delta}_i \in \mathbb{R}^{\mathbf{H} \times \mathbf{W} \times 3}$ . A pixel located at  $(u, v)$  in the image  $\mathbf{I}_i$  is unprojected from the image plane onto the 3D position  $\boldsymbol{\mu}_p \in \mathbb{R}^3$  by unproject function



$\Pi^{-1}(\mathbf{K}; \mathbf{R}_i; \mathbf{t}_i)$ :

$$\Pi^{-1}[u, v] := \mathbf{R}_i^\top \mathbf{K}^{-1} \mathbf{D}_i[u, v] - \mathbf{t}_i + \Delta[u, v] \quad (1)$$

where,  $\mathbf{D}_i[u, v]$  represents the depth value at pixel  $(u, v)$  in the  $i$ -th view’s depth map;  $\mathbf{R}_i$  is the rotation matrix of the  $i$ -th view,  $\mathbf{K}$  is the camera intrinsic matrix, and  $\mathbf{t}_i$  is the translation vector of the  $i$ -th view. The proposed Transformer architecture leverages feature aggregation to synergistically exploit human geometry priors and the information embedded within input images, effectively bridging the 2D feature space and 3D coordinates to deliver enriched geometric representations.

### 3.2 Self-Supervised Model as Inductive Bias

We employ the DINOv2 `vits14-with-registers` [20, 77] as 2D Image Encoder, which has been proven effective for 3D scene segmentation [78] and diverse downstream tasks [79]. We freeze the network parameters and extract features from the input image, denoted as  $\mathbf{f}_i \in \mathbb{R}^{(h \times w) \times d_2}$ , where  $d_2$  is the dimension of the image encoder. Furthermore, as the feature aggregation stage for the posed images has already learned the relevance between each position token, we directly extract the attention weights from the previous Transformer and perform a weighted combination of the features which is independent of the feature dimension. We leverage the tailored inductive biases learned by the preceding Transformer across the hierarchical Transformer blocks. Specifically, for each Transformer block, the feature-wise similarity is extracted by the dot product of  $\mathbf{Q}(\mathbf{F}_i)$  and  $\mathbf{K}(\mathbf{F}_i)$  corresponding to each  $\mathbf{V}(\mathbf{f}_i)$ .

$$\text{CrossAttn}(\mathbf{f}_i) := \text{SoftMax} \left( \frac{\mathbf{Q}(\mathbf{F}_i) \mathbf{K}(\mathbf{F}_i)^\top}{\sqrt{d_k}} + \mathbf{B} \right) \mathbf{f}_i \quad (2)$$

where,  $d_k$  presents the dimension of  $\mathbf{F}_i$ ,  $\mathbf{B}$  is the relative position bias. Ultimately, from each output token  $\tilde{\mathbf{f}}_i \in \mathbb{R}^{(h \times w) \times d_2}$ , we decode the attributes of pixel-aligned Gaussians  $\mathcal{G}$  in the corresponding patch using a convolutional layer with a  $1 \times 1$  kernel. The final pixel color  $\mathbf{c}$  is calculated by blending  $\mathbf{N}$  ordered Gaussians overlapping the pixels via the following rendering function:  $\mathbf{c} = \sum_{i=1}^{\mathbf{N}} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j)$ . This equation efficiently models the contributions of each Gaussian to the pixel’s final appearance, accounting for their transparency and layering order. To facilitate body-part semantic 3D representation, inspired by [26], we augment the 3D Gaussians Splats with a learnable semantic feature embedding (denoted as 3D Gaussians primitives) and rasterize onto the 2D image plane by blending Gaussians that overlap with each pixel using a feature rendering function. This implies that the novel view synthesis task and human-part segmentation task share the same 3D Gaussian parameters, where  $\mathbf{N}$  is the number of Gaussian primitives participating in the blending:  $f = \sum_{i=1}^{\mathbf{N}} \mathbf{M} \tilde{\mathbf{f}}_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j)$ . where  $f$  indicates the final rasterized feature embedding on the image plane, and  $\tilde{\mathbf{f}}_i$  represents the semantic

### 3.3 Multi-task Training Objective

The 3D reconstruction loss function is designed to minimize the rendering loss  $\mathcal{L}_{\text{render}}$  for novel viewpoints. The loss for  $k_v$  rendered multi-view images is defined as:

$$\mathcal{L}_{\text{render}} := \mathbb{E}[\mathcal{L}_{\text{mse}}(\hat{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_m \mathcal{L}_{\text{mask}}(\hat{\mathbf{M}}_i, \mathbf{M}_i) + \lambda_p \mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{I}}_i, \mathbf{I}_i)] \quad (3)$$

where  $\mathbf{I}_i$  and  $\hat{\mathbf{I}}_i$  denote the ground-truth images and rendered images via 3D Gaussian Splatting, respectively.  $\mathbf{M}_i$  and  $\hat{\mathbf{M}}_i$  represent the original and rendered foreground masks.  $\mathcal{L}_{\text{mse}}$  measures the mean squared error loss, and  $\mathcal{L}_p$  measures the perceptual loss [80].  $\lambda_m$  and  $\lambda_p$  are hyperparameters employed for balancing the respective loss terms. Building upon prior work [31], the loss function for the feature field is minimized during training by utilizing rasterized feature maps on novel views  $\mathbf{f}$  and directly inferred feature maps using ground truth images on new views  $\hat{\mathbf{f}}$ , thereby facilitating the learning of blending weights for consistent semantic field regression.  $\mathcal{L}_{\text{dist}} = 1 - \text{CosSim}(\hat{\mathbf{f}}, \mathbf{f}) = 1 - \frac{\hat{\mathbf{f}} \cdot \mathbf{f}}{\|\hat{\mathbf{f}}\| \|\mathbf{f}\|}$ , where  $\text{CosSim}(\cdot, \cdot)$  denotes the cosine similarity between the predicted and ground truth feature maps, which serves as a distance metric to be minimized during training. Finally, we construct a semantic segmentation dataset for fine-tuning. We employ 28 classes for body-part segmentation,

along with the background class, following [19]. We jointly train the network on all tasks using differentiable rendering. Our model can be optimized in an end-to-end manner:

$$\mathcal{L}(\Theta) := \mathbb{E}_{i \in \{1, \dots, k_v\}} [\mathcal{L}_{\text{render}} + \lambda_{\text{dist}} \cdot \mathcal{L}_{\text{dist}}(\mathbf{f}_i, \hat{\mathbf{f}}_i)] + \lambda_{\text{seg}} \cdot \mathbb{E}_{j \in \{1, \dots, k_s\}} [\mathcal{L}_{\text{CE}}(\mathbf{S}_j, \hat{\mathbf{S}}_j)] \quad (4)$$

where  $\Theta$  is the model parameters,  $\mathbf{S}_i$  and  $\hat{\mathbf{S}}_j$  denote the annotated and predicted semantic segmentation, respectively.  $L_{\text{CE}}$  represents the Cross Entropy Loss, and we enforce the predicted score to align with the ground truth viewpoint. Since we only have a small number of annotated viewpoints,  $L_{\text{CE}}$  is added only when  $j \in \{1, \dots, k_s\}$ . The Aggregation mechanism is independent of features. Therefore, we can leverage a 2D pretrained model to supervise the generation of rasterized novel views using  $\mathcal{L}_{\text{dist}}$ , and allow the two tasks to benefit each other through  $\mathcal{L}_{\text{CE}}$ .

## 4 Experiments

We conducted experiments on the following datasets to evaluate the results of 3D human reconstruction and segmentation. The 2D diffusion model takes approximately 6 seconds (with the number of views set to 2) to generate multi-view latent features, while the subsequent reconstruction stage takes only about 0.2 seconds.

**Datasets.** (1) THuman2.1 Dataset [81] contains approximately 2500 human scans. Specifically, we select 2300 scans for training and the rest for evaluation. (2) 2K2K Dataset [82] includes 2000 human scans. Similarly, we select 1500 scans for training and the rest for evaluation. (3) Human MVImageNet [83] approximately comprises 4000 identities and 8000 outfits, which provide the rich multi-perspectives. Consistent with the protocol established in [48], we utilize PIXIE [84] as the SMPL parameter estimator, strategically placing 36 cameras across three hierarchical levels to capture full-body, upper-body, and facial views, with all renderings resolution of  $512 \times 512$  pixels.

**Curated Dataset.** We randomly select 500 scans from the training dataset and annotate 8 semantic segmentation maps for each scan. More details are provided in Appendix A.1.

**In-the-wild Dataset.** To assess the model’s generalizability under challenging conditions, we construct a test dataset from Internet-sourced images. These images encompass diverse human poses, identities, and camera viewpoints.

**Training Details.** The hyperparameters  $\lambda_{\text{mask}}$ ,  $\lambda_p$  are set to 1 and 0.1 in this paper. The hyperparameters  $\lambda_{\text{dist}}$  and  $\lambda_{\text{dist}2}$  are both set to 0.5. We use the AdamW optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.95$ , and a weight decay of 0.05 is applied to all parameters except those in the LayerNorm layers. A cosine learning rate decay scheduler is employed, with a linear warm-up of 2,000 steps. The peak learning rate is set to  $4 \times 10^{-4}$ . The training process is divided into two stages: the model is trained for 80K iterations at  $256 \times 256$  resolution and then fine-tuned for an additional 20K iterations at  $512 \times 512$  resolution. Please refer to Appendix A.2 for more detailed procedural insights.

### 4.1 Evaluation of Human 3D Segmentation

**Comparisons.** The semantic segmentation is evaluated by class-wise intersection over union (mIoU) and average pixel accuracy (mAcc) on novel views as metrics. To provide a more comprehensive evaluation of the algorithm’s 3D consistency, we compare HUMANCRAFTER against the state-of-the-art LSM [31] baseline. Additionally, we benchmark our approach against the 2D state-of-the-art

Table 1: Comparison with feed-forward Human 3D Segmentation methods on 2K2K dataset.

Method	2K2K [82]					Runtime
	mIoU $\uparrow$	Acc. $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	
<i>Two GT Input Views</i>						
LSM* [31]	0.724	0.873	23.811	0.892	0.053	<b>108 ms / object</b>
Sapiens [19]	0.823	0.904	N/A	N/A	N/A	640 ms / frame
Ours	0.840	0.925	24.786	0.937	0.022	126 ms / object
<i>Single View</i>						
Human3Diffusion [51] + Sapiens [19]	0.781	0.851	21.832	0.891	0.069	23.21 s / object
Ours	<b>0.801</b>	<b>0.882</b>	<b>23.489</b>	<b>0.916</b>	<b>0.045</b>	<b>6.24 s / object</b>

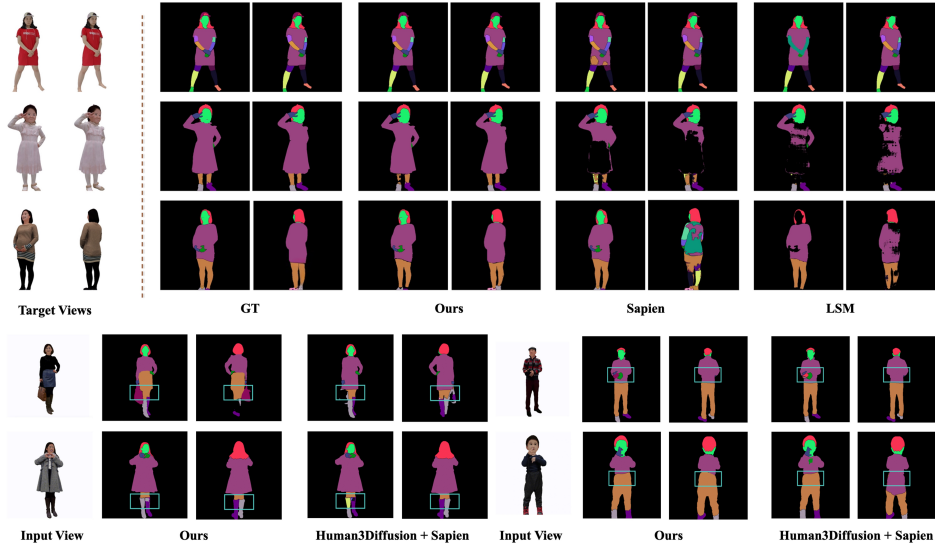


Figure 3: Qualitative Results and Comparisons on Human 3D Segmentation on THuman2.1 and 2K2K datasets. HUMANCRAFTER achieves the best precise segmentation results in terms of 3D consistency.

human segmentation algorithm proposed in [19], which is trained on large-scale human datasets. LSM\* is trained using 3D human scans and 2D semantic segmentation maps, and it takes two images as input to ensure a fair comparison. As illustrated in Table 1, HUMANCRAFTER outperforms the state-of-the-art baselines in terms of segmentation accuracy, while exhibiting comparable reconstruction times. To enhance single-image reconstruction, we employ a two-stage approach using cutting-edge algorithms [51] as a baseline. As shown in Table 1, HUMANCRAFTER surpasses baselines in segmentation accuracy while maintaining comparable reconstruction times. Furthermore, Figure 3 demonstrates the superior segmentation quality achieved by proposed method.

## 4.2 Evaluation of 3D Human Reconstruction

**Comparisons.** The quantitative assessment leverages metrics such as PSNR, SSIM [87], and VGG-LPIPS [80] to comprehensively analyze the rendering fidelity. Inspired by DiffSplat [], multi-view consistency is evaluated through COLMAP reconstructed point number [88]. We compare our approach with state-of-the-art methods. These include advanced reconstruction-based techniques for single-image-conditioned generation: LGM [64], GRM [85], InstantMesh [62], Lara [86], Human3Diffusion [51], and PSHuman [49]. For the single-view setting, our method can either replicate the same input or utilize existing multi-view diffusion models available on the shelf to introduce 2D generative priors and achieve better results, as demonstrated in the last row of Table 2. The single image-conditioned generation performance on the THuman 2.1 and 2K2K datasets is evaluated in Table 2, with Qualitative results on challenging scenarios (e.g., far camera viewpoints, complex human poses, and loose clothing) are presented in Figure 4 and Figures 7 in the Appendix. HUMANCRAFTER demonstrates superior performance compared to state-of-the-art baselines in both

Table 2: Comparison with feed-forward 3D reconstruction methods at a resolution of  $512 \times 512$ .

Method	THuman2.1 [81]				2K2K [82]			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	#Points $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	#Points $\uparrow$
LGM $\dagger$ [64]	20.106	0.859	0.196	502.05	21.685	0.850	0.166	694.84
GRM $\dagger$ [85]	20.503	0.868	0.141	602.46	21.496	0.858	0.171	722.95
InstantMesh [11]	19.997	0.875	0.128	1803.28	21.983	0.865	0.118	2028.30
LaRa [86]	18.120	0.840	0.207	3035.43	19.113	0.860	0.207	4139.94
SiFU [73]	20.164	0.842	0.088	4500.68	21.698	0.904	0.084	4560.18
PSHuman [49]	20.853	0.862	0.076	5321.23	21.932	0.892	0.076	4734.23
Human3Diffusion [51]	22.164	0.872	0.063	5123.52	22.323	0.882	0.053	5134.23
HumanCrafter	<b>23.186</b>	<b>0.907</b>	<b>0.046</b>	<b>5744.96</b>	<b>23.489</b>	<b>0.916</b>	<b>0.045</b>	<b>6453.23</b>



Figure 4: Novel-view images rendered by HUMANCRAFTER and the state-of-the-art baselines on various datasets. Our method achieves the highest rendering quality. Please refer to the zoomed-in regions for details.

rendering quality and 3D consistency metrics. Notably, LGM<sup>†</sup> and GRM<sup>†</sup> have been fine-tuned on our dataset to ensure a fair comparison.

### 4.3 Ablation Studies

we carefully investigate the effectiveness of human prior and each design choice in this subsection.

**The Effect of Human Parametric Model.** As shown in Table 3, the human geometric prior benefits significantly from the reconstruction model, particularly when  $k_v = 1$ . Additionally, in the context of human body reconstruction tasks, rendering SMPL normals can provide richer geometric cues compared to relying on coordinate or rendered depth maps. As the number of input viewpoints  $k_v$  increases, the reconstruction model effectively resolves geometric ambiguities through stereo matching, reducing its dependence on Human Pose Estimation. The statistics in Table 3 validate HUMANCRAFTER consistently outperforms other variants in terms of image quality and fidelity.

**Model Design Choices.** As demonstrated in Table 4 (b)-(d), the experiments confirm the efficacy of each crucial design decision. **i)** The HUMANCRAFTER model, in the absence of Plucker embeddings and relies solely on the SMPL-based geometric prior, demonstrates a marginal decline in performance. **ii)** When integrating the pre-trained MAE model, denoted as ViT-s [89], in place of the DINOv2 model, HUMANCRAFTER exhibits a slight performance decrease. This alteration is attributed to

Table 3: Ablation study of human geometry prior on 2K2K dataset.

	$k_v$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	#Param. $\downarrow$
w/o SMPL	1	22.203	0.890	0.064	41.2M
+ Depth	1	23.103	0.901	0.0270	42.4M
+ Coord.	1	23.324	0.917	0.047	42.6M
+ Normal (Ours)	1	23.489	0.916	0.045	42.4M
w/o SMPL	2	23.570	0.913	0.034	41.2M
with SMPL (Ours)	2	24.786	0.937	0.022	42.6M

Table 4: Ablation of model and objective design on 2K2K dataset.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
(a) Full Model (Ours)	<b>23.489</b>	<b>0.916</b>	<b>0.045</b>
(b) w/o Cam. Emb.	23.327	0.903	0.048
(c) w/o DINOv2	22.025	0.891	0.055
(d) w/o Pixel-align Aggregation	21.183	0.891	0.067
(e) w/o $\mathcal{L}_{\text{dist}}$	22.464	0.896	0.055
(f) w/o $\mathcal{L}_{\text{CE}}$	23.223	0.901	0.051

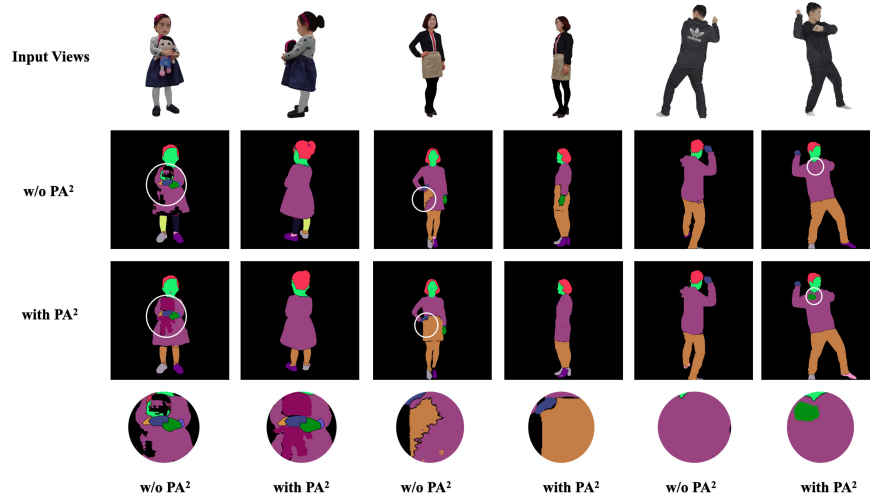


Figure 5: **Ablation of Pixel-Align Aggregation.** HUMANCRAFTER with PA<sup>2</sup> can leverage knowledge learned from novel-view synthesis task and incorporate a pre-trained 2D model, thereby boosting semantic tasks.

the superior ability of the selected pre-trained model to establish correspondences among the input view images, a crucial factor for pixel-aligned aggregation. **iii)** Furthermore, HUMANCRAFTER is solely enhanced with the Feature Aggregation Module akin to the current LGM methodology, and the learned features are directly forwarded through the GS-Decoder, a notable decline in performance is observed, as illustrated in Table 4 (d) and Figure 5 in the Appendix. This result underscores the efficacy of the Pixel-align Aggregation Module.

**The Effect of Loss Functions.** As demonstrated in Table 4 (e)-(f), without the LPIPS loss, novel view renderings are susceptible to blurriness and unnatural generations, leading to a slight performance decline. The integration of the distillation loss enhances 3D view consistency. Similarly, akin to LSM, the addition of the semantic distillation loss illustrates that integrating the human segmentation task enhances the performance of novel view synthesis.

#### 4.4 Applications: Human Editing and Immersive Exploring

Figure 1 and Figure 6 illustrate the potential scenarios enabled by the proposed model: **(1) Text to VersatileSplats:** ControlNet [90] is used to produce a human image with the human mask and text prompts. Subsequently, HUMANCRAFTER is employed to reconstruct a 3D human from the generated image. **(2) 3D Consistency Editing:** HUMANCRAFTER’s 3D coherence and precise semantic masks are employed to direct the 3D human editing process, supported by a FLUX-based inpainting model [91]. **Immersive Exploring:** HUMANCRAFTER showcases high efficiency, enabling real-time end-to-end 3D modeling. Following the generation of edited 3DGS primitives for the provided input views, seamless integration into VR devices.



Figure 6: We demonstrate the generalizability of HUMANCRAFTER with in-the-wild images in challenging scenarios.





Figure 7: We demonstrate the generalizability of HUMANCRAFTER with in-the-wild images in challenging scenarios.

As depicted in Figure 4.3, we validate that Pixel-Align Aggregation effectively utilizes information from the task of synthesizing new viewpoints. As illustrated in Figure 6, a potential application of our proposed model is its combination with the existing Text-to-Image (T2I) inpainting Diffusion model, such as FLUX.1 [91]. Notably, the 3D Gaussian primitives we generate can seamlessly integrate into VR devices. We demonstrate this through the application of watching a high-fidelity virtual concert using the PICO 4 Ultra VR headset. As depicted in Figure 7, we showcase the generalizability of our model using in-the-wild images in challenging scenarios.

## 5 Conclusion

We have introduced HUMANCRAFTER, a unified framework for 3D human reconstruction and understanding. First, we adopt tailored human priors and aggregate multi-view images from a 2D diffusion model and camera embedding features in a Transformer. Second, we translate the set of aggregated features to pixel-aligned 3D Gaussians as initialized geometry. We extend a 2D pre-trained model into a 3D consistent feature field and establish a weighting mechanism to propagate into multi-view. Extensive experiments demonstrate that HUMANCRAFTER surpasses existing methods in terms of novel view synthesis quality and downstream task performance while exhibiting robustness in complex scenarios.

**Broader Impacts** HUMANCRAFTER allows users to generate 3D human models tailored to their specific inputs, enabling a broad spectrum of downstream applications, such as AR/VR Chat, 3D cinematography, and 3D editing. However, this capability also presents potential ethical challenges, including privacy violations and racial biases. To mitigate these risks, it is imperative to establish robust ethical guidelines and enforce legal regulations.

**Limitations and Future Works.** Building on the significant acceleration our method provides for semantic 3D human reconstruction, a compelling avenue for future work is its extension to dynamic 4D scene generation with Gaussian representations [92]. Furthermore, by leveraging web-scale human datasets and relying solely on 2D supervision, extensive real-world video datasets could further unlock the potential of HUMANCRAFTER. Moreover, there is potential for misuse, such as the arbitrary distribution of digital assets. These risks can be mitigated by embedding watermarks into the 3D assets [93, 94].

## References

- [1] Ben Mildenhall, Pratul Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.
- [2] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 2020.
- [3] 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.
- [4] 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 (ICCV)*, 2023.
- [5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023.
- [6] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023.
- [7] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [8] 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.
- [9] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [10] Sen Peng, Weixing Xie, Zilong Wang, Xiaohu Guo, Zhonggui Chen, Baorong Yang, and Xiao Dong. Rmavatar: Photorealistic human avatar reconstruction from monocular video based on rectified mesh-embedded gaussians. *arXiv preprint arXiv:2501.07104*, 2025.
- [11] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.
- [12] Huiqun Wang, Yiping Bao, Panwang Pan, Zeming Li, Xiao Liu, Ruijie Yang, and Di Huang. Multi-modal relation distillation for unified 3d representation learning. In *European Conference on Computer Vision*, pages 364–381. Springer, 2024.
- [13] Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable gaussian splat generation. *arXiv preprint arXiv:2501.16764*, 2025.
- [14] Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. *arXiv preprint arXiv:2506.05573*, 2025.
- [15] Chenguo Lin, Yuchen Lin, Panwang Pan, Xuanyang Zhang, and Yadong Mu. Instructlayout: Instruction-driven 2d and 3d layout synthesis with semantic graph prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

- [17] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [18] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision(ECCV)*, pages 451–468. Springer, 2025.
- [19] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025.
- [20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [21] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17840–17852, 2023.
- [22] Yizhou Wang, Yixuan Wu, Shixiang Tang, Weizhen He, Xun Guo, Feng Zhu, Lei Bai, Rui Zhao, Jian Wu, Tong He, et al. Hulk: A universal knowledge translator for human-centric tasks. *arXiv preprint arXiv:2312.01697*, 8, 2023.
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.
- [24] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.
- [25] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kotschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023.
- [26] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suyu You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024.
- [27] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024.
- [28] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2025.
- [29] Renjie Li, Zhiwen Fan, Bohua Wang, Peihao Wang, Zhangyang Wang, and Xi Wu. Versatile-gaussian: Real-time neural rendering for versatile tasks using gaussian splatting. 2025.
- [30] Juliette Marrie, Romain Ménégaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. *arXiv preprint arXiv:2410.14462*, 2024.

- [31] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. In *Advanced in Neural Information Processing Systems (NeurIPS)*, 2024.
- [32] Arthur Moreau, Jifei Song, Helisa Dharmo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [33] Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [34] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [35] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [36] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023.
- [37] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [38] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiye Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Lan Xu. Headgap: Few-shot 3d head avatar via generalizable gaussian priors, 2024.
- [39] 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.
- [40] Xinqi Liu, Chenming Wu, Jialun Liu, Xing Liu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. Gva: Reconstructing vivid 3d gaussian avatars from monocular videos. *arXiv preprint arXiv:2402.16607*, 2024.
- [41] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [42] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [43] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [44] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. *arXiv preprint arXiv:2404.01053*, 2024.
- [45] Boyao Zhou, Shunyuan Zheng, Hanzhang Tu, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian+: Generalizable pixel-wise 3d gaussian splatting for real-time human-scene rendering from sparse views. *arXiv preprint arXiv:2411.11363*, 2024.

- [46] Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse camera settings. *arXiv preprint arXiv:2410.01425*, 2024.
- [47] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*, 2024.
- [48] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *Advanced in Neural Information Processing Systems (NeurIPS)*, 2024.
- [49] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024.
- [50] Yiyu Zhuang, Jiayi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image, 2025.
- [51] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. *arXiv preprint arXiv:2406.08475*, 2024.
- [52] Bokui Shen, Xinchen Yan, Charles R Qi, Mahyar Najibi, Boyang Deng, Leonidas Guibas, Yin Zhou, and Dragomir Anguelov. Gina-3d: Learning to generate implicit neural assets in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.
- [53] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024.
- [54] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.
- [55] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [56] Hanwen Jiang, Qixing Huang, and Georgios Pavlakos. Real3d: Scaling up large reconstruction models with real-world images. *arXiv preprint arXiv:2406.08479*, 2024.
- [57] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *International Conference on Learning Representations (ICLR)*, 2024.
- [58] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf-: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.
- [59] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [60] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *European Conference on Computer Vision (ECCV)*, 2024.



- [61] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrn: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024.
- [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] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics (TOG)*, 2023.
- [64] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *European Conference on Computer Vision (ECCV)*, 2024.
- [65] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024.
- [66] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrn: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024.
- [67] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [68] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [69] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *European Conference on Computer Vision (ECCV)*, 2024.
- [70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on Machine Learning (ICML)*, 2021.
- [71] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [72] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 2015.
- [73] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [74] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: Free-view human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024.
- [75] Julius Plücker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, 1865.
- [76] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

- [77] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *International Conference on Learning Representations (ICLR)*, 2024.
- [78] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022.
- [79] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.
- [80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [81] 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 (CVPR)*, 2021.
- [82] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [83] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024.
- [84] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021.
- [85] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024.
- [86] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields, 2024.
- [87] 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 (TIP)*, 2004.
- [88] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [89] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.
- [90] 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.
- [91] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2023.
- [92] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024.
- [93] Chenxin Li, Brandon Y Feng, Zhiwen Fan, Panwang Pan, and Zhangyang Wang. Steganerf: Embedding invisible information within neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 441–453, 2023.

- [94] Chenxin Li, Hengyu Liu, Zhiwen Fan, Wuyang Li, Yifan Liu, Panwang Pan, and Yixuan Yuan. Instantsplamp: Fast and generalizable stenography framework for generative gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [95] 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.
- [96] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- [97] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>, 2022.
- [98] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.
- [99] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.
- [100] Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiaoxiao Long, and Ping Tan. Rade-gs: Rasterizing depth in gaussian splatting, 2024.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of the work is discussed in Section ??.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?



Answer: [No]

Justification: Code will be released upon acceptance of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

#### 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the training and test details are specified in Section ?? and Section ??.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

#### 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Please refer to Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Please refer to Section B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Please refer to Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Creators or original owners of assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.



## A More Implementation Settings

We provide comprehensive implementation details in this section to facilitate the reproducibility of our work. Specifically, in Section. A.1 , we provide the details of how to construct semantic segmentation data. Section. A.2, we provide details about training details. In Section. B, we offer further explanation of the implementation for HUMANCRAFTER.

### A.1 Constructed Dataset Details

**Interactive Annotation.** Human-part Semantic Segmentation aims to classify pixels in the input image  $I_i$  into  $N_{\text{class}}$  categories while ensuring 3D consistency. Following Sapiens [19], we construct a dataset with  $N_{\text{class}} = 28$  (27 body parts and one background class). The class names are as follows: ‘Background’, ‘Apparel’, ‘Face\_Neck’, ‘Hair’, ‘Left\_Foot’, ‘Left\_Hand’, ‘Left\_Lower\_Arm’, ‘Left\_Lower\_Leg’, ‘Left\_Shoe’, ‘Left\_Sock’, ‘Left\_Upper\_Arm’, ‘Left\_Upper\_Leg’, ‘Lower\_Clothing’, ‘Right\_Foot’, ‘Right\_Hand’, ‘Right\_Lower\_Arm’, ‘Right\_Lower\_Leg’, ‘Right\_Shoe’, ‘Right\_Sock’, ‘Right\_Upper\_Arm’, ‘Right\_Upper\_Leg’, ‘Torso’, ‘Upper\_Clothing’, ‘Lower\_Lip’, ‘Upper\_Lip’, ‘Lower\_Teeth’, ‘Upper\_Teeth’, and ‘Tongue’.

To accelerate the manual annotation process, we utilize the Segment Anything Model (SAM) [95] for assisted labeling. The dataset construction process is illustrated in Figure 8. In (a), we show the instance data and segmentation pipeline, and in (b), we demonstrate how to accelerate annotation using Segment Anything Model.

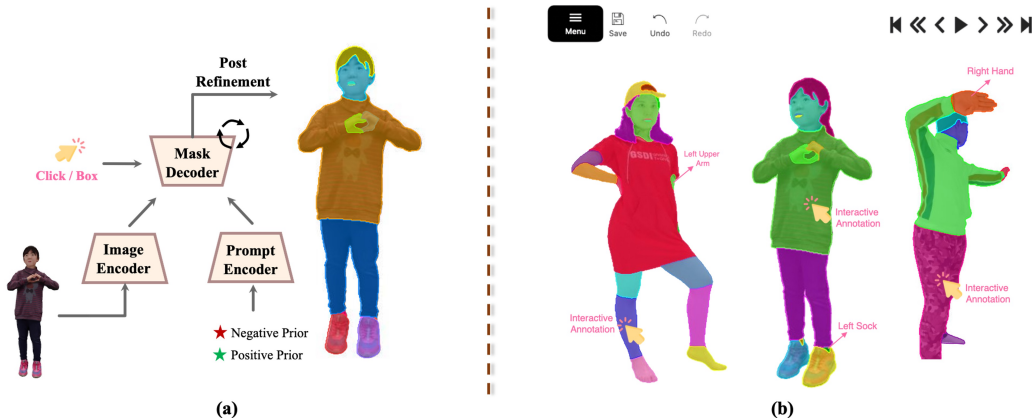


Figure 8: The dataset construction pipeline. (a) The instance and semantic segmentation annotation pipeline allows us to repeatedly reuse the features of input images and the prior negative and positive coordinates. (b) Accelerating annotation procedure by leveraging the Segment Anything Model [95].

### A.2 More Training Details

To accelerate the training process, we employ Flash-Attention-v2 [96] from the xFormers library [97], gradient checkpointing [98], and BFloat16 mixed-precision arithmetic [99]. Leveraging a pre-trained model and human geometric priors, our method takes 7 days of training on 8 NVIDIA A800 GPUs.

**Differentiable 3DGS Rasterization.** A modified 3DGS rasterization implementation<sup>1</sup> [100] supports depth, alpha, and **normal** rendering. Additionally, we extend 3DGS rasterization pipeline to incorporate feature attributes, enabling **feature** rendering from new perspectives, where all operations are differentiable. Due to limitations in GPU memory, we render 3DGS features  $\mathbf{f}$  with a dimensionality of up to 1024 at most. Initially, we filter out low-opacity 3D Gaussian splats ( $\sigma_p < 0.005$ ) to enhance rendering speed without compromising quality.

<sup>1</sup><https://github.com/BaowenZ/RaDe-GS.git>

## B More Details of HUMANCRAFTER

**Dataset Normalization.** To better learn the 3DGS attributes, we place the human scans at the origin of the coordinate system and normalize them to a unit cube, so that they are located within the bounding box  $([-1, 1]^3)$  in the world space. The camera poses of the rendered views are normalized with a global scale of 1.4.

**3D Gaussian Primitives Normalization.** As 3D Gaussians are unstructured explicit representation, the parameterization of the output parameters can affect the model’s convergence. For numerical values of Gaussian splat properties, we set to Spherical Harmonics to 3, and all attributes confined within  $[0, 1]$  in preparation of diffusion-based generation, outputs of 3DGS are all activated by the sigmoid function, except for  $\mathbf{r}$ , which is  $L_2$ -normalized to yield unit quaternions. RGB color  $c$  and opacity  $o$  are already supposed to be in  $[0, 1]$ . Raw scale  $\hat{s}$  is linearly interpolated with predefined values  $s_{\min}$  and  $s_{\max}$  [65].  $\mathbf{s} := s_{\min} \cdot \text{sigmoid}(\hat{s}) + s_{\max} \cdot (1 - \text{sigmoid}(\hat{s}))$ . Here,  $s_{\min}$  and  $s_{\max}$  are set to  $5\text{e-}4$  and  $2\text{e-}2$  respectively to represent fine-grain details.

**Ablation on Image Encoder [20].** We freeze the image encoders based on DINOv2’s best practices to leverage its pre-trained features and to maintain training efficiency by only training a lightweight decoder. We validated this design choice with an ablation study on the 2K2K dataset, as shown in Table 5. The results indicate that fine-tuning the image encoder provides only marginal gains (+0.032 PSNR).

Table 5: Ablation study on the effect of fine-tuning the DINOv2 image encoder. The experiment is conducted on the 2K2K dataset.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Fine-tuned DINOv2	<b>23.521</b>	<b>0.918</b>	0.046
Frozen DINOv2	23.489	0.916	<b>0.045</b>

**Ablation on Dual-Transformer.** The first Transformer (Feature Aggregation) focuses on the general task of aggregating multi-view geometric and appearance features. The second Transformer (Pixel-align Aggregation) is specialized, using an attention mechanism to translate these fused features into the structured parameters of our semantic 3D Gaussians. To validate this, we trained a baseline with a single, monolithic Transformer, and the results, presented in Table 6, confirm our design is superior.

Table 6: Ablation study of our two-stage Transformer architecture. The baseline “w/o Pixel-align Aggregation” uses a single, monolithic Transformer.

Architecture	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o Pixel-align Aggregation	21.183	0.891	0.067
Full Model (Ours)	<b>23.489</b>	<b>0.916</b>	<b>0.045</b>