HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors

Panwang Pan[*1], **Zhuo Su**[*†1], **Chenguo Lin**[*1,2], **Zhen Fan**[1], **Yongjie Zhang**[1], **Zeming Li**[1],
Tingting Shen[3], **Yadong Mu**[2], **Yebin Liu**[‡4]
[*] Equal contribution   [†] Project lead   [‡] Corresponding author

[1]ByteDance, [2]Peking University, [3]Xiamen University, [4]Tsinghua University

Abstract

Despite recent advancements in high-fidelity human reconstruction techniques, the requirements for densely captured images or time-consuming per-instance optimization significantly hinder their applications in broader scenarios. To tackle these issues, we present **HumanSplat**, which predicts the 3D Gaussian Splatting properties of any human from a single input image in a generalizable manner. Specifically, HumanSplat comprises a 2D multi-view diffusion model and a latent reconstruction Transformer with human structure priors that adeptly integrate geometric priors and semantic features within a unified framework. A hierarchical loss that incorporates human semantic information is devised to achieve high-fidelity texture modeling and impose stronger constraints on the estimated multiple views. Comprehensive experiments on standard benchmarks and in-the-wild images demonstrate that HumanSplat surpasses existing state-of-the-art methods in achieving photorealistic novel-view synthesis. Project page: https://humansplat.github.io.

1 Introduction

Realistic 3D human reconstruction is a fundamental task in computer vision with widespread applications in numerous fields, including social media, gaming, e-commerce, telepresence, etc. Previous works for single-image human reconstruction can be broadly categorized into explicit and implicit approaches. Explicit methods [1, 2, 3, 4], such as those based on parametric body models like SMPL [5, 6], estimate human meshes by directly optimizing parameters and clothing offset to fit the observed image. However, these methods often struggle with complex clothing styles and require lengthy optimization. Implicit methods [7, 8, 9, 10, 11] represent humans using continuous functions, such as occupancy [7], SDF [12], and NeRF [13]. While they excel at modeling flexible topology, these methods are limited in scalability and efficiency due to the high computational cost associated with training and inference. Recent advances in 3D Gaussian Splatting (3DGS) [14] have provided a balance between efficiency and rendering quality for reconstructing detailed 3D human models, which however rely on multi-view image [15, 16, 17] or monocular video [18, 19, 20] input. Recent popular human reconstruction studies [21, 22, 23, 24] focus on the score distillation sampling (SDS) technique [25] to lift 2D diffusion priors to 3D, but time-consuming optimization (e.g., 2 hours) is required for each instance. Some generalizable and large-reconstruction-model-based works [10, 26, 27, 28] can directly generalize the regression of 3D representations but either disregard human priors or require multi-view inputs, limiting the stability and feasibility in downstream applications.

In this paper, we propose a novel generalizable approach **HumanSplat** for single-image human reconstruction by introducing a generalizable Gaussian Splatting framework with a fine-tuned 2D multi-view diffusion model and well-conceived 3D human structure priors. Different from existing human 3DGS methods, our approach directly infers Gaussian properties from a single input image,
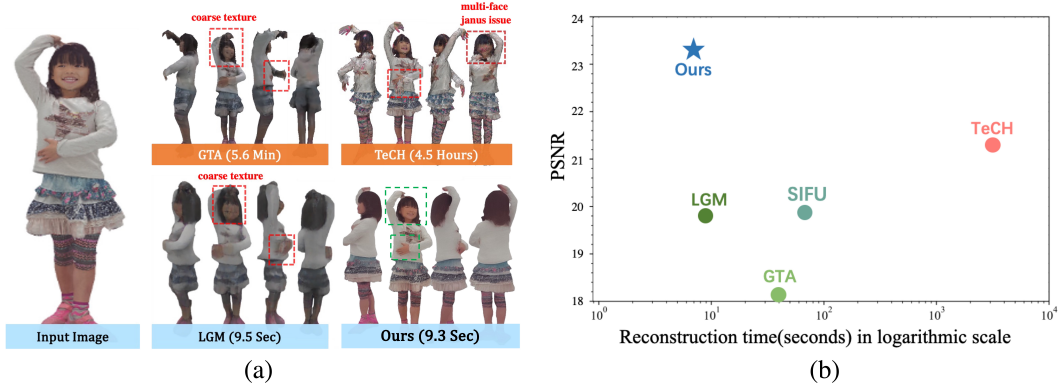
Figure 1: Our method achieves state-of-the-art rendering quality while maintaining the fastest runtime. (a) Qualitative results: LGM [29] and GTA [30] are generalizable but in lower quality, TeCH [21] exhibits issues with multi-face rendering and is time-consuming. In contrast, our method achieves higher fidelity in a much shorter time. (b) Performance and runtime comparison: metrics are evaluated on the challenging Twindom dataset.

eliminating the requirement for per-instance optimization or densely captured images. This empowers our method to generalize effectively in diverse scenarios while delivering high-quality reconstructions.

The key insight behind our method is to reconstruct Gaussian properties from the diffusion latent space in a generalizable end-to-end architecture, integrating a 2D generative diffusion model as appearance prior and a human parametric model as structure prior. Specifically, facing this under-constraint reconstruction problem from single-view input with extensive invisible regions, we first leverage a 2D multi-view diffusion model (denoted as **novel-view synthesizer**) to hallucinate the unseen parts of clothed humans. Then a generalizable **latent reconstruction Transformer** is proposed to enable interaction among the generated diffusion latent embeddings and the geometric information of the structured human model, enhancing the quality of 3DGS reconstruction. During interactions, to mitigate the limitations of inaccurate human priors like the SMPL model, we devise a projection strategy to strike a balance between robustness and flexibility, projecting 3D tokens into the 2D space and conducting searches within adjacent windows using projection-aware attention. Another specific challenge of human reconstruction is to capture fine detail in visually sensitive areas, such as the face and hand. To address this problem, we exploit the semantic cues from the structure priors and propose **semantics-guided objectives** to further promote the fine details reconstruction quality. By synergistically amalgamating these crucial components into a unified framework, our method achieves state-of-the-art performance in striking the right balance between quality and efficiency.

In summary, the main contributions of our work are as follows:

- We propose a novel generalizable human Gaussian Splatting network for high-fidelity human reconstruction from a single image. To the best of our knowledge, it is the first to leverage latent Gaussian reconstruction with a 2D generative diffusion model in an end-to-end framework for efficient and accurate single-image human reconstruction.

- We integrate structure and appearance cues within a universal Transformer framework by leveraging human geometry priors from the SMPL model and human appearance priors from the 2D generative diffusion model. Geometric priors stabilize the generation of high-quality human geometry, while appearance prior helps hallucinate unseen parts of clothed humans.

- We enhance the fidelity of reconstructed human models by introducing semantic cues, hierarchical supervision, and tailored loss functions. Extensive experiments demonstrate that our method achieves state-of-the-art performance, surpassing existing methods.

2 Related Work

Single-Image Human Reconstruction. Single-image human reconstruction methods fall into explicit and implicit approaches. Explicit methods use parametric body models [5, 6] to estimate

<div align="center">2</div>

minimally clothed human meshes [1, 2, 31, 32, 33, 34, 35, 36, 37], adding clothing via 3D offsets [3, 4, 38, 39, 40, 41] or garment templates [42, 43, 44]. However, these methods are constrained by topology, particularly with loose clothing. Implicit methods [7, 8, 9, 10, 11, 22, 27, 45, 46], utilizing representations like occupancy, signed distance fields (SDF) and Neural Radiance Fields (NeRF), offer flexibility with topology, allowing for accurate depiction of 3D clothed humans. Recent methods [30, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61] combine implicit representation with explicit human priors for better robustness. Among these, GTA [30] uses Transformers with fixed embeddings for translating image features into 3D tri-plane features. TeCH [62] employs diffusion-based models for visualizing unseen areas but needs extensive optimization and accurate SMPL-X models. Another approach, HumanSGD [22], uses diffusion models for texture inpainting but struggles with mesh inaccuracies from [8]. NeRF-based methods like SHERF [10] generate human NeRF from single images using hierarchical features, while ELICIT [11] leverages CLIP [63] for contextual understanding. These NeRF-based methods produce high-quality images but often struggle with detailed 3D mesh generation and require extensive optimization time.

Human Gaussian Splatting. While implicit representations like SDF or NeRF struggle with balance optimization efficiency and rendering quality, the high efficiency of 3DGS [14] has advanced 3D human creation. Recent 3DGS methods have targeted multi-view videos, monocular video sequences, and sparse-view input. Multi-view video methods like D3GA [15], HuGS [16], and 3DGS-Avatar [17] exploit dense spatial and temporal information for detailed models. Additionally, HiFi4G [64] combine 3DGS with a dual-graph mechanism to preserve spatial-temporal consistency, ASH [65] utilizes mesh UV parameterization for real-time rendering, and Animatable Gaussians [66] improve garment dynamics via pose projection mechanism and 2D CNNs. The monocular video only provides sufficient observation as humans move, so the monocular human Gaussian Splatting methods [18, 19, 20, 67, 68, 69, 70, 71, 72, 73] often resort to using LBS for mapping to a canonical space and introducing additional regularization terms. For sparse-view images, GPS-Gaussian [28] achieves high rendering speeds without per-subject optimization by using a feed-forward way with efficient 2D CNNs encoding from diverse 3D human scan data. In comparison, our work stands out as the first one-shot generalizable human GS method, using only single image input and achieving high-quality reconstruction without any optimization and fine-tuning.

Generalizable Large Reconstrucion Model. Large reconstruction model (LRM) [74, 75] presents that with sufficient parameters and training datasets, a deterministic feed-forward Transformer [76] is capable of learning a triplane feature for volumetric rendering from a single-view image. Targeting general 3D object, TriplaneGaussian [77] combines 3DGS with triplane as a hybrid 3D representation for fast rendering, where point clouds are first extracted from an image and then projected on triplane features to decode 3DGS attributes. Most recent methods further leverage 2D diffusion models that can generate multi-view images simultaneously [78, 79, 80, 81] to provide plausible and abundant inputs for the following reconstruction, so they focus on the sparse multi-view reconstruction task. Instant3D [81] follows the technique of LRM and produces triplane features from 4 posed images by a Transformer encoder. CRM [82], MeshLRM [83] and InstantMesh [84] replaces triplane NeRF by FlexiCubes [85] to extract meshes directly. LGM [29], GRM [86] and GS-LRM [87] predict 3DGS attributes from each input pixel [88, 89] and combine the Gaussians from multiple viewpoints as the final 3D output. For human task, CharacterGen [90] follows Instant3D [81] to adopt a two-stage approach: first, generating a 2D multi-view canonical images, then using multi-view SDF reconstruction for cartoon digital characters, without direct interaction between stages. A more compact and elegant method HumanLRM [27] replaces training data of LRM [74] from general objects with human data to predict triplane NeRF, showcasing the capacities of the large model. While it relies heavily on the dataset, often leading to issues like missing limbs due to a lack of prior knowledge. In contrast, our 3DGS-based method leverages the latent diffusion and reconstruction model in an end-to-end framework and introduces human priors by addressing their inaccuracy with specific projection strategies, achieving better generalization and stability with even less data.

3 Method

3.1 Preliminary

SMPL [5] is a widely-used parametric human body model that is created by skinning and blend shapes, and learned from thousands of 3D body scans. The SMPL model [5] utilizes shape parameters
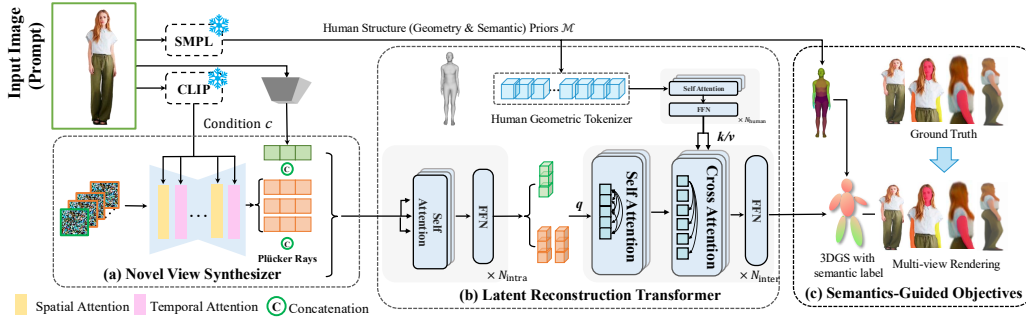
Figure 2: Overview of HumanSplat. (a) Multi-view latent features are first generated by a fine-tuned multi-view diffusion model (Novel View Synthesizer in Sec. 3.3). (b) Then, the Latent Reconstruction Transformer (Sec. 3.4) interacts global latent features (Sec. 3.4.1) and human geometric prior (Sec. 3.4.2). (c) Finally, the semantic-guided objectives (Sec. 3.5) are proposed to reconstruct the final human 3DGS.

$\boldsymbol{\beta} \in \mathbb{R}^{10}$ and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{24 \times 3}$ to parameterize the deformation of the human body mesh $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\beta} \times \boldsymbol{\theta} \mapsto \mathbb{R}^{6890 \times 3}$.

3D Gaussians Splatting [14] is a recently proposed 3D representation that formulates 3D content by a set of $\mathbf{N_p}$ colored Gaussians $\mathcal{G} = \{\mathbf{g}_i\}_{i=1}^{\mathbf{N_p}}$. Each Gaussian $\mathbf{g}_i = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right)$ has opacity $\boldsymbol{\sigma}_i \in \mathbb{R}$ and color $\mathbf{c}_i \in \mathbb{R}^3$ attributes, where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is its location and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$ implies its scaling $\mathbf{s_i} \in \mathbb{R}^3$ and orientation quaternion $\mathbf{q_i} \in \mathbb{R}^4$ in 3D space. These attributes constitute $\mathbf{G} \in \mathbb{R}^{\mathbf{N_p} \times 14}$ can represent a 3D object, and by projecting to the image plane with opacity-based color composition in the depth order, it can be differentiably rendered into 2D images in real-time.

3.2 Overview

Given a single input image of a human body $\mathbf{I_0} \in \mathbb{R}^{H \times W \times 3}$, our task is to reconstruct a 3D representation, i.e., 3DGS in this work, from which novel-view images can be rendered. As illustrated in Fig. 2, our method leverages a 2D diffusion model and a novel latent reconstruction Transformer that adeptly integrates 2D appearance priors, human geometric priors, and semantic cues within a universal framework. First, we use SMPL estimator [91, 92] to predict human structure priors $\mathcal{M}$, and CLIP [63] to generate the image embedding of the input image $\mathbf{c}$. A latent temporal-spatial diffusion model [93], referred to as the **novel-view synthesizer** (Sec. 3.3), is fine-tuned and produces multi-view latent features $\left\{\mathbf{F}_i \in \mathbb{R}^{h \times w \times c}\right\}_{i=1}^{\mathbf{N}}$, where $\mathbf{N}$ is the number of views, and h, w and c are the height, width and channel of the latent features. Then, we utilize a novel **latent reconstruction Transformer** (Sec. 3.4) that adeptly integrates human geometric prior and latent features within a universal framework, which predicts Gaussian attributes $\mathbf{G} := \left\{(\boldsymbol{\mu}_i, \boldsymbol{q}_i, \boldsymbol{s}_i, \boldsymbol{c}_i, \boldsymbol{\sigma}_i) | i = 1, ..., \mathbf{N_p}\right\}$ and rendered into new images, where $\mathbf{N_p}$ represents the number of Gaussian points. Furthermore, to achieve high-fidelity texture modeling and better constrain the estimated multiple views, we designed a hierarchical loss that incorporates human semantic priors (See 3.5). Note that during training, the network is trained using 3D scan data to ensure accurate supervision from various viewpoints, where the registered SMPL is fitted via the multi-view version of SMPLify [92]. During inference, the model employs PIXIE [91] to predict human structure priors $\mathcal{M}$ and synthesize novel views from only a single image input based on the trained model, without any fine-tuning and optimization.

3.3 Video Diffusion Model as Novel-view Synthesizer

We take advantage of the pre-trained video diffusion model SV3D [93] as appearance prior. For the input image $\mathbf{I}_0$, we leverage a CLIP image encoder [63] to obtain the image embedding $\mathbf{c}$ and a pre-trained VAE [94] $\mathcal{E}$ to acquire latent feature $\mathbf{F}_0$ as conditions. Subsequently, we progressively denoises gaussian noises into temporal-continuous $\mathbf{N}$ multi-view latent features $\{\mathbf{F}_i\}_{i=1}^{\mathbf{N}}$ by a spatial-temporal UNet D_θ [95] with the objective:

$$\mathbb{E}_{\epsilon \sim p(\epsilon)} \left[\lambda(\epsilon) \| D_\theta(\{\mathbf{F}_i^\epsilon\}_{i=1}^{\mathbf{N}}; \{e_i, a_i\}_{i=1}^{\mathbf{N}}, \mathbf{c}, \mathbf{F}_0, \epsilon) - \{\mathbf{F}_i\}_{i=1}^{\mathbf{N}} \|_2^2 \right], \tag{1}$$

where $\{\mathbf{F}_i^\epsilon\}_{i=1}^{\mathbf{N}}$ is the noise-corrupted multi-view latent features, $\{e_i, a_i\}_{i=1}^{\mathbf{N}}$ is the corresponding relative elevation and azimuth angles to $\mathbf{I}_0$, $p(\epsilon)$ is the noise probability distribution and $\lambda(\epsilon) \in \mathbb{R}^+$ is the loss weighting term related to the noise level ϵ. Since SV3D is merely pre-trained on the dataset of objects [96], we further fine-tune it using human datasets to improve its effectiveness in the human reconstruction task. More technical details are available in Appendix A.1.

3.4 Latent Reconstruction Transformer

Latent reconstruction Transformer contains two parts: **latent embedding interaction** (Sec. 3.4.1) and **geometry-aware interaction** (Sec. 3.4.2), which seamlessly incorporate latent features from novel-view synthesizer and human structure priors into the reconstruction process. The two components together form a cohesive framework for accurately recovering intricate human details.

3.4.1 Latent Embedding Interaction

We utilize a pretrained VAE [94] encoder $\mathcal{E}$ to encode the input image $\mathbf{I}_0$ into a latent representation $\mathbf{F}_0 = \mathcal{E}(\mathbf{I}_0) \in \mathbb{R}^{h \times w \times c}$ and combine it with the generated $\{\mathbf{F}_i\}_{i=1}^{\mathbf{N}}$ (Sec. 3.3). Following previous works [97, 98], latent representations are concatenated with their corresponding Plücker embeddings [99] along the channel dimension, resulting in a dense pose-conditioned feature map. These latents are then divided into non-overlapping patches [100] and mapped to d-dimensional latent tokens by a linear layer. An intra-attention module is followed to thoroughly extract spatial correlations of latent tokens, as shown in Fig. 3, by a standard Transformer [76] with N_{intra} blocks of multi-head self-attention and feed-forward network (FFN):

$$\bar{\mathbf{F}}_i = [\text{FFN}(\text{SelfAttention}(\mathbf{F}_i))]_{\times N_{\text{intra}}}. \tag{2}$$

3.4.2 Geometry-aware Interaction

Human Geometric Tokenizer. Given a human geometric model $\mathcal{M} \in \mathbb{R}^{6890 \times 3}$, to obtain the feature representation, we project $\mathcal{M}$ onto the input view and compute the feature vector through bilinear interpolation on the feature grids by

$$\mathbf{\Pi}_0(\mathcal{M}) = \mathbf{K}(\mathbf{R}_0 \mathcal{M} + \mathbf{t}_0), \tag{3}$$

where $\mathbf{R}_0$, $\mathbf{t}_0$ are camera extrinsic parameters of the input view, $\mathbf{K}$ is the intrinsic parameters. After the projection operation, we obtain geometric human prior tokens by concatenating $\mathcal{M}$ and $\mathbf{F}_0[\mathbf{\Pi}_0(\mathcal{M})]$, and mapping them to d dimension. $\mathbf{F}_0[\mathbf{\Pi}_0(\mathcal{M})]$ stands for querying the input latent feature $\mathbf{F}_0$ from projected locations. Similar to Latent Embedding Interaction (Sec. 3.4.1), an intra-attention module is utilized to interact among these human prior tokens and get human geometric features $\bar{\mathbf{H}} \in \mathbb{R}^{6890 \times d}$.

Human Geometry-aware Attention. Empirically, we find that incorporating human priors [5, 6, 101] (e.g., SMPL model) into human reconstruction faces a significant dilemma of the trade-off between robustness and flexibility [30, 102]. On the one hand, the SMPL model provides geometric priors that alleviate common issues (e.g., broken body parts and various artifacts). On the other hand, it struggles to accurately capture and represent a broad spectrum of complex clothing styles, particularly more fluid garments like dresses and skirts. This limitation highlights a fundamental shortcoming in the model's capacity to accurately depict diverse apparel. As shown in Fig. 3, HumanSplat addresses inaccuracies in human priors by projecting 3D tokens into 2D space and conducting local searches within adjacent windows, efficiently using priors while minimizing redundancy. Specifically, we introduce projection-aware attention within the inter-attention module, utilizing a window $\mathbf{W}(K_{\text{win}} \times K_{\text{win}})$ and employing masked multi-head cross-attention. Here, the latent features $\{\bar{\mathbf{F}}_i\}_{i=1}^{\mathbf{N}}$ serve as queries, and the human prior tokens $\bar{\mathbf{H}}$ serve as keys and values. The process can be formulated as follows:

$$\{\tilde{\mathbf{F}}_i\}_{i=1}^{\mathbf{N}} = [\text{FFN}(\text{CrossAttention}_{\text{mask}}(\{\bar{\mathbf{F}}_i\}_{i=1}^{\mathbf{N}}, \bar{\mathbf{H}}))]_{\times N_{\text{inter}}}, \tag{4}$$

where feature interactions occur when $\mathbf{\Pi}(\bar{\mathbf{H}}, \mathbf{K}, \mathbf{R}_i, \mathbf{t}_i)$ is within the window of $\bar{\mathbf{F}}_i$. It is noteworthy that the complexity of the cross-attention operation is reduced from $\mathcal{O}(\mathbf{L}_F \times \mathbf{L}_H)$ to $\mathcal{O}(\mathbf{L}_F \times K_{\text{win}}^2)$, compared to the vanilla multi-head cross-attention setting, where $\mathbf{L}_F$ and $\mathbf{L}_H$ represent the lengths of the latent and human geometric tokens respectively.
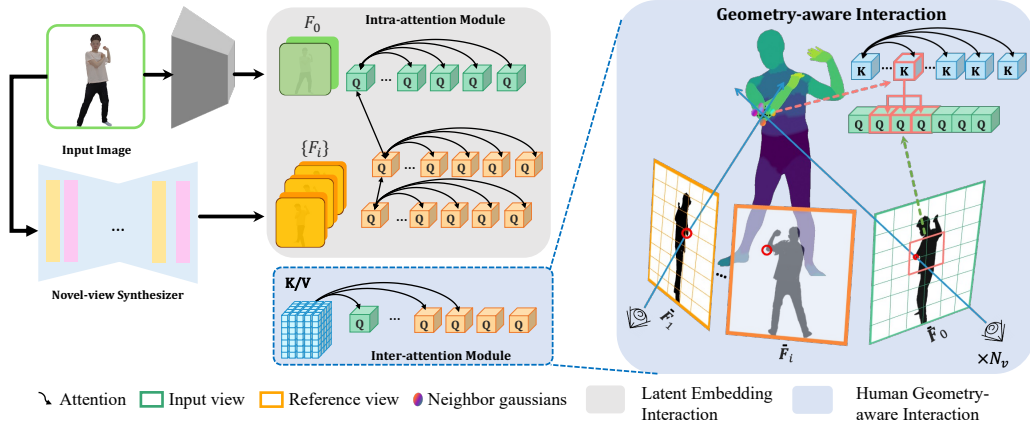
Figure 3: Illustration of latent reconstruction Transformer. It first divides $\mathbf{F}_0$ and $\mathbf{F}_i$ into non-overlapping patches, which are then processed through an intra-attention module (Sec. 3.4.1). Within the iter-attention module (Sec. 3.4.2), we introduce the projection-aware attention with a window $\mathbf{W}(K_{\text{win}} \times K_{\text{win}})$, and the attributes of 3D Gaussians are decoded with a Conv 1×1 layer.

3.5 Semantics-guided Objectives

From each output token $\tilde{\mathbf{F}}_i$, we decode the attributes of pixel-aligned Gaussians $\mathbf{G}$ in the corresponding patch by a convolutional layer with a 1×1 kernel. These attributes are used to render novel-view images through 3D Gaussian Splatting rasterization. An ideal training objective should ensure that the rendered outputs closely match the supervised images, striving for overall consistency between the reconstructed renderings and the ground truth. However, human attention tends to be particularly focused on facial regions, where intricate details play a pivotal role in perception. Therefore, to enhance the fidelity of reconstructed human models, we optimize our objective functions to preferentially focus on the facial regions.

Hierarchical Loss. Traditional approaches often overlook the semantic richness inherent in human anatomy, leading to reconstructions that lack crucial details and accuracy. To address this, we propose a novel framework that harnesses semantic cues, hierarchical supervision, and tailored loss functions to guide the training process. This ensures that not only are the overall structures of the human body accurately represented, but also that critical areas such as the face are reconstructed with exceptional detail and fidelity. Specifically, by incorporating a human prior model equipped with semantic information, which adheres to the widely-used definition of 24 human body parts, such as "right hand", "left hand", "head", etc. The ground-truth body part segmentation is obtained from SMPL meshes with predefined semantic vertices [103]. Therefore, we can utilize different attention weights for different body parts and render different levels of real images from different viewpoints to provide hierarchical supervision. This approach significantly facilitates the precise localization of essential body parts, including the head, hands and arms. Please refer to Appendix A.2 for more detailed procedural insights. The overall loss is defined as a weighted sum of part-specific losses:

$$\mathcal{L}_H = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i\lambda_j(\mathcal{L}_{\text{MSE}}(\mathbf{I}_{i,j}, \hat{\mathbf{I}}_{i,j}) + \lambda_p\mathcal{L}_{\text{p}}(\mathbf{I}_{i,j}, \hat{\mathbf{I}}_{i,j})), \tag{5}$$

where $i \in \{1, ..., n\}$ and $j \in \{1, ..., m\}$ denote different resolution levels and human parts. $\hat{\mathbf{I}}_{i,j}$ and $\mathbf{I}_{i,j}$ are the predicted and ground-truth data for part j, λ_i and λ_j are weighting factors that reflect the relative importance of each resolution level and each body part, respectively. $\mathcal{L}_{\text{MSE}}$ measures the MSE loss and $\mathcal{L}_p$ measures the perceptual loss [104].

Reconstruction Loss. Reconstruction loss $\mathcal{L}_{\text{Rec}}$ based on the $\mathbf{N}_{\text{render}}$ rendered multi-view images is defined as

$$\mathcal{L}_{\text{Rec}} = \sum_{i=1}^{\mathbf{N}_{\text{render}}} \mathcal{L}_{\text{MSE}}(\mathbf{I}_i, \hat{\mathbf{I}}_i) + \lambda_m\mathcal{L}_{\text{MSE}}(\mathbf{M}, \hat{\mathbf{M}}_i) + \lambda_p\mathcal{L}_{\text{p}}(\mathbf{I}_i, \hat{\mathbf{I}}_i), \tag{6}$$

Table 1: Quantitative comparison on texture against other methods. * indicates that this method is fine-tuned on our training dataset. † indicates that this method requires per-instance optimization.

Method	Category		THuman2.0 [105]			Twindom [106]			Time
	Diffusion	Human Prior	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
PIFu [107]	✗	✗	18.093	0.911	0.137	-	-	-	30.0s
LGM* [29]	✓	✗	20.013	0.893	0.116	19.840	0.851	0.292	9.5s
GTA [30]	✗	✓	18.050	-	-	17.669	0.741	0.418	43s
SIFU [102]	✗	✓	22.025	0.921	0.084	19.714	0.832	0.312	44s
SIFU† [102]	✓	✓	22.102	0.923	0.079	-	-	-	6min
Magic123† [108]	✓	✗	14.501	0.874	0.145	-	-	-	1h
HumanSGD† [22]	✓	✓	17.365	0.895	0.130	-	-	-	7min
TeCH† [21]	✓	✓	**25.211**	**0.936**	0.083	21.192	0.884	0.188	4.5h
HumanSplat (Ours)	✓	✓	24.033	0.918	**0.055**	**23.346**	**0.913**	**0.125**	**9.3s**

where $\mathbf{I}_i$ and $\hat{\mathbf{I}}_i$ are the ground-truth images and render images via Gaussian Splatting, $\mathbf{M}_i$ and $\hat{\mathbf{M}}_i$ are original and rendered foreground mask. λ_m and λ_p are hyperparameters for balancing three kinds of losses. The final training objective is a combination of the two loss terms: $\mathcal{L} = \mathcal{L}_H + \mathcal{L}_{\text{Rec}}$.

4 Experiments

4.1 Implementation Details

Training Details. HumanSplat is trained on 500 THuman2.0 [105], 1500 2K2K [109], and 1500 Twindom [106] high-fidelity human scans. We evenly position 36 cameras across each of three hierarchical cycles to capture the full body, half body, and face, with rendering resolution set to 512×512. We conduct 200 epochs of 256-res training with a learning rate of 1e-5 and a batch size of 32 over 2 days on 8 A100 (40G VRAM) GPUs, while 512-res finetuning costs 2 additional days. We train our model with AdamW [110] optimizer, whose β_1, β_2 are set to 0.9 and 0.95 respectively. A weight decay of 0.05 is used on all parameters except those of the LayerNorm layers. We use a cosine learning rate decay scheduler with a 2000-step linear warm-up and the peak learning rate is set to 4e-4. For parts related to the head, hands, and arms, λ_j are set to 2, while the rest human part are set to 1. The parameters λ_i, λ_p and λ_m are set to 1. The model is trained for 80K iterations on 256-res and then fine-tuned on 512-res for another 20K iterations. To enable efficient training and inference, we adopt Flash-Attention-v2 [111] in the xFormers [112] library, gradient checkpointing [113], and FP16 mixed-precision [114]. Fine-tuning the novel view synthesizer takes about 18 hours. Please refer to Appendix A.1 for more details about novel view synthesizer fine-tuning.

Inference Time. The video diffusion model takes about 9s to generate multi-view latent features while the subsequent latent reconstruction for 3DGS takes only about 0.3s. Additionally, it can render novel views at a rate exceeding 150 FPS on a NVIDIA A100 GPU.

4.2 Comparison

Quantitative Comparison. We conduct a quantitative comparison on 21 THuman2.0 [105] and 51 Twindom [106] scans, rendering textured meshes from multiple angles and using PSNR, SSIM [115], and VGG-LPIPS [104] as evaluation metrics. To ensure fairness, we fine-tune LGM [29] on our datasets and fully optimize each TeCH [21] instance for comparison. As shown in Tab. 1, HumanSplat consistently outperforms previous generalizable methods across all datasets. Specifically, HumanSplat surpasses SIFU [102] by +1.92 (8.72%) in PSNR and reduces LPIPS from 0.079 to 0.055. Notably, HumanSplat also excels on the more challenging Twindom dataset, outperforming TeCH [21] by +2.15 (10.16%) in PSNR and reducing LPIPS from 0.188 to 0.125. Although the values can further decrease with model training, we control the convergence to the current extent in the dataset for better generalization under more general in-the-wild scenarios. Besides, we also report the reconstruction times for 512×512 input images on a NVIDIA A100 GPU. HumanSplat achieves a remarkably fast reconstruction time, approximately 9.3 seconds on a single NVIDIA A100 GPU, making it much more practical than TeCH's 4.5-hour reconstruction time.

Qualitative Comparison. As illustrated in Fig. 4 and Fig. 5, by incorporating semantics-guided objectives, Humansplat achieves more detailed and higher fidelity results against GTA [30] and

Figure 4: Qualitative comparison of ours against TeCH [21], GTA [30] and LGM [29] on Thuman2.0 [105], Twindom [106] and in-the-wild images. Our method achieves the highest quality. Note that TeCH achieves clearer results but fails to preserve the face identity.

LGM [29]. Although TeCH [21] also generates high-quality images, its SDS optimization often results in overly saturated multi-face outcomes [116, 117]. Moreover, while TeCH delivers the highest PSNR on the Thuman2.0 dataset, the LPIPS metric in Tab. 1, confirms that our method provides more realistic results. As shown in Fig. 6, Humansplat outperforms HumanSGD on in-the-wild images from Adobe Stock, effectively predicting 3D Gaussian Splatting properties without per-instance optimization. Notably, the results on in-the-wild images further demonstrate our strong generalization capability and our model provides superior textures on both the front and invisible regions, highlighting the effectiveness of incorporating appearance and geometric priors.

4.3 Ablation Study

Latent Reconstruction Transformer. HumanSplat leverages the Stable Diffusion latent space for feature extraction. Experimental results demonstrate that utilizing the compact latent space for reconstructions leads to better rendering quality compared to using the pixel space, as detailed in
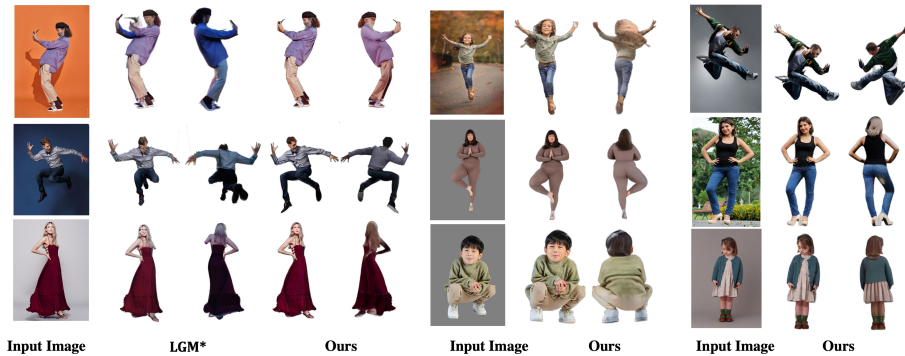
Figure 5: Qualitative results showcasing reconstructions of humans in challenging poses, diverse identities, and varying camera viewpoints from in-the-wild images.

Figure 6: Qualitative Comparison of ours against HumanSGD [22] on in-the-wild images.

Tab. 2 (a). We also find that patching in the time dimension for latent decoding by temporal VAE [95] adversely affects the quality, so we employ the vanilla VAE [94] decoder in this work.

Human Geometric Prior. Tab. 2 (b) illustrates the significance of human geometric prior in our approach. By utilizing SMPL as the geometric prior for human representation, we assess the robustness of HumanSplat with respect to different SMPL parameters. We report both the estimated [2] and Ground Truth SMPL parameters [92] and compare the model that does not incorporate a human structure prior. These experiments provide insight into the benefit of incorporating a human structure prior, as significant performance degradation is observed without it. Furthermore, thanks to treat human priors as keys and values, HumanSplat exhibits robustness in testing against challenging poses with imperfect estimation, with only minor declines in metrics. The effectiveness of the human geometric prior is also demonstrated qualitatively in Fig. 7 (a).

Window Size. As analyzed in Sec. 3.4.2, human priors are crucial for the robustness of the human reconstruction model. As evidenced in Tab. 2 (c), when the window size K_{win} is set to 1 (dubbed "Projection Only"), employing human priors significantly enhances performance, improving the PSNR metric from 22.635 to 23.333. We further explore setting window size to $K_{\text{win}} = 2$ in a latent space with a resolution of 64×64, which further improves the model's PSNR metric from 23.333 to 24.294. This improvement is attributed to Humansplat with $K_{\text{win}} = 2$ is capable of handling loose clothing and tolerating imprecise SMPL parameters. In Tab. 3, we present results for the 2K2K dataset across various window sizes. Further refinement with a window size of $K_{\text{win}} = 3$ continues to enhance the results, whereas using $K_{\text{win}} = 4$ does not yield additional improvements.

Semantics-guided Training Objectives. Thanks to the incorporation of semantic-guided hierarchical losses, HumanSplat can generate higher-quality textures for both the face and hand, while also promoting the convergence of the training process, as depicted in Fig. 7 (a) and (b). Furthermore, Fig. 7 (c) demonstrates that HumanSplat further produces 3D coherent segmentation results, which have potential applications in editing and specific generation of individual parts of 3D humans.
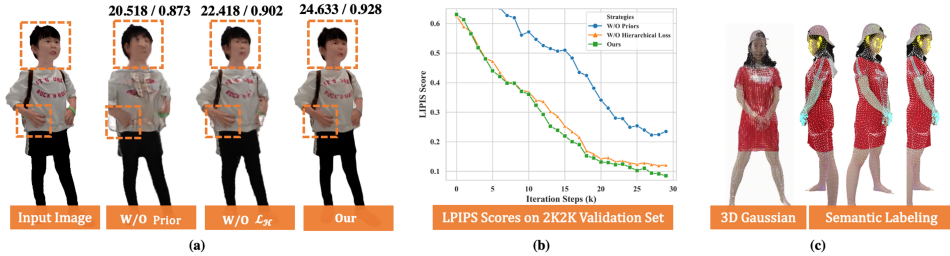
9

Figure 7: (a) Qualitative evaluation of human prior and hierarchical loss $\mathcal{L}_{\mathcal{H}}$. PSNR/SSIM values are shown at the top of each image, presenting the effectiveness of the two strategies in the full pipeline. (b) LPIPS scores during training on the 2K2K validation set. (c) HumanSplat can provide 3D coherent segmentation results for potential applications.

Table 2: Ablation study of several proposed designs on 2K2K [109] evaluation dataset.

(a) Reconstruction Space.

Space	PSNR↑	SSIM↑	LPIPS↓
Pixel Space	23.125	0.892	0.066
Temporal VAE [95]	24.133	0.902	0.047
Ours [94]	**24.293**	**0.915**	**0.040**

(b) Geometric Prior Initialization.

Method	PSNR↑	SSIM↑	LPIPS↓
GT SMPL	**24.633**	**0.935**	**0.025**
w/o SMPL	22.635	0.893	0.182
Ours [2]	24.293	0.915	0.040

(c) Different projection methods.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o Human Prior	22.635	0.893	0.182
Projection Only	23.333	0.916	0.057
Ours ($K_{\text{win}} = 2$)	**24.293**	**0.915**	**0.040**

Table 3: Evaluation of different window sizes K_{win} on the 2K2K [109] evaluation dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Human Prior	22.635 (-1.658)	0.893 (-0.022)	0.182 (+0.142)
$K_{\text{win}} = 1$	23.333 (-0.960)	0.916 (-0.001)	0.057 (+0.017)
$K_{\text{win}} = 3$	24.383 (+0.089)	0.931 (+0.016)	0.041 (+0.001)
$K_{\text{win}} = 4$	24.013 (-0.280)	0.922 (+0.007)	0.048 (+0.008)
$K_{\text{win}} = 2$ **(Ours)**	24.293	0.915	0.040

5 Conclusion

We present HumanSplat, a pioneering generalizable human reconstruction network that derives 3D Gaussian Splatting properties from a single image. This model integrates generative diffusion and latent reconstruction Transformer models with human structure priors, enhanced by a tailored semantic-aware hierarchical loss. These innovations achieve high-fidelity reconstruction results in a feed-forward manner without any optimization or fine-tuning, especially in the important focal areas such as the face and hands. Extensive experiments demonstrate that HumanSplat surpasses existing state-of-the-art methods in both quality and efficiency, including robust performance in handling challenging poses and loose clothing. This capability opens up a broad spectrum of potential applications.

Limitations and Future Works. Our method has some limitations that can be addressed in future works: (a). While effective with most attire, the model may struggle with intricate garments and accessories. Future enhancements should incorporate 2D data and innovative training strategies, which is a promising direction for improving the handling of diverse and unusual clothing styles. (b). Although already efficient, increasing the computational speed by constraining Gaussian's number or combining compact 3DGS with texture representation could further facilitate real-time applications, particularly on mobile devices. (c). Animating the reconstructed human models currently requires post-processing. Future work could introduce canonical space to reconstruct animatable avatars directly, streamlining the process and improving efficiency.

References

[1] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] 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.

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[4] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision (3DV)*, 2018.

[5] 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.

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[8] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] 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.

[11] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *IEEE Conference on Computer Vision (ICCV)*, 2023.

[12] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019.

[13] 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.

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023.

[15] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023.

[16] Arthur Moreau, Jifei Song, Helisa Dhamo, 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[23] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] Xihe Yang, Xingyu Chen, Daiheng Gao, Shaohui Wang, Xiaoguang Han, and Baoyuan Wang. Have-fun: Human avatar reconstruction from few-shot unconstrained images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

[26] Jiteng Mu, Shen Sang, Nuno Vasconcelos, and Xiaolong Wang. Actorsnerf: Animatable few-shot human rendering with generalizable nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[27] 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.

[28] 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.

[29] 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.

[30] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[33] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2023.

[34] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. NIKI: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[35] Xin Chen, Zhuo Su, Lingbo Yang, Pei Cheng, Lan Xu, Bin Fu, and Gang Yu. Learning variational motion prior for video-based motion capture. *arXiv preprint arXiv:2210.15134*, 2022.

[36] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[37] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[38] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Thiemo Alldieck, Marcus A. Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-Degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, 2019.

[42] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[43] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2020.

[44] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023.

[45] 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.

[46] Muxin Zhang, Qiao Feng, Zhuo Su, Chao Wen, Zhou Xue, and Kun Li. Joint2human: High-quality 3d human generation via compact spherical embedding of 3d joints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[47] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[48] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[49] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2021.

[51] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[52] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[53] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K. Wong. JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[54] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[55] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[56] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[57] 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 *European Conference on Computer Vision (ECCV)*, 2020.

[58] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[59] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[60] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[61] Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. Ohta: One-shot hand avatar via data-driven implicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[77] 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.

[78] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations (ICLR)*, 2024.

[79] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

[80] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

[81] 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.

[82] 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.

[83] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024.

[84] 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.

[85] 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.

[86] 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.

[87] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024.

[88] 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.

[89] 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.

[90] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)*, 2024.

[91] 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.

[92] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[93] 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.

[94] 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 (CVPR)*, 2022.

[95] 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.

[96] 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.

[97] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *International Conference on Learning Representations (ICLR)*, 2024.

[98] Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, and Abe Davis. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[99] Julius Plücker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, 1865.

[100] 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.

[101] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[102] 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.

[103] Smpl segmentation. https://meshcapade.wiki/assets/SMPL_body_segmentation/ smpl/smpl_vert_segmentation.json.

[104] 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.

[105] 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.

[106] Twindom. https://web.twindom.com.

[107] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[108] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[109] 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.

[110] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[111] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[112] 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.

[113] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[114] 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.

[115] 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.

[116] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

[117] 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 (CVPR)*, 2023.

[118] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[119] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[120] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[121] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024.

[122] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024.

[123] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

A Implementation Details

A.1 Additional Novel View Synthesizer Details

Conditions of Diffusion Model. SVD and SV3D [93, 95] replace CLIP text embeddings [63] with the image embedding of the conditioning. Compared to text prompts, image prompts provide information more accurately and directly, and also save time that would otherwise be spent on text inversion or image caption [21, 102]. SV3D takes an image and the relative camera orbit elevation and azimuth angles as inputs to generate several corresponding novel-view images, which are continuous in 3D space and can be regarded as a video with the camera moving. In our work, we finetuned SV3D on human datasets with relative camera elevation and azimuth angles to generate 4 orthogonal views, which we found to be sufficient for reconstruction [102].

Transformer Block. The CLIP embeddings obtained from the conditioning image is fed into the cross-attention layers of each Transformer block, acting as keys and values, while the feature at that layer serves as the queries. Moreover, the camera trajectory and the diffusion noise timestep are incorporated into the residual blocks.

Fine-tuning Details. We involves the widely used EDM [118, 119] framework, incorporating a simplified diffusion loss for finetuning. We fine-tune the novel-view synthesizer at 512-res uses 1000-step warmup with the peak learning rate 1e-4 in the cosine learning rate decay schedule. We use a per-GPU batch size of 8 objects during 256-res training, and a per-GPU batch size of 4 during 512-res finetuning stage. For each instance, we use 4 input views and 4 novel supervision views at each iteration of 256-res training and 512-res finetuning.

Triangular CFG Scaling. **C**lassifier-**f**ree **G**uidance (**CFG**) [120] is a widely used technique to trade off controllability with diversity. However, this scaling causes the last few frames in our generated orbits to be over-sharpened. To tackle this issue, we use a triangle wave CFG [93, 121] scaling during inference: linearly increase CFG from 1 at the front view to 2.5 at the back view, then linearly decrease it back to 1 .

3D Consistency Evaluation. We also conduct an evaluation of the 3D consistency of our novel-view synthesizer using the advanced local correspondence matching algorithm, MASt3R [122]. This approach involves one-to-one image matching between input views and their generated novel views, with the average number of matching correspondences serving as our metric. Our synthesizer exhibits a notable improvement in 3D consistency for human novel-view generation after fine-tuning with human datasets. Specifically, the original SV3D achieves an average of 723.25 matching points, while our novel-view synthesizer increases this number to 930.13. In comparison, the ground truth demonstrates a matching point number of 1106.33. This evaluation highlights that fine-tuning on human datasets significantly enhances multi-view consistency metrics.

A.2 Additional Latent Reconstruction Transformer Details

Network Architecture. The network architecture of the proposed latent reconstruction Transformer is shown in Fig. 8. The framework is composed of a latent encoder with intra- and inter-attention and a Gaussian parameter prediction module. It takes initial multi-view latent embeddings as input and performs projection-aware attention to aggregate human geometric information. From each output token, we decode the attributes of pixel-aligned Gaussians in the corresponding patch with a Conv 1×1 layer.

Hierarchical Loss. Our approach is efficient and can be utilized online without pre-computation. We calculate the visible face triangles given the mesh and camera parameters. Then each visible triangular face is assigned the corresponding semantic label [103]. Additionally, it provides labels with superior 3D consistency and avoids issues associated with training instability. First, the hierarchical loss is specifically focused on the head and hands, excluding hair and clothing, which can be further supervised using reconstruction loss $\mathcal{L}_{\text{Rec}}$. Second, instead of using segmentation to reconstruct and then concatenate different components, we design our loss function based on part segmentation. Thank to this design, HumanSplat is tolerant of inaccuracies in segmentation.
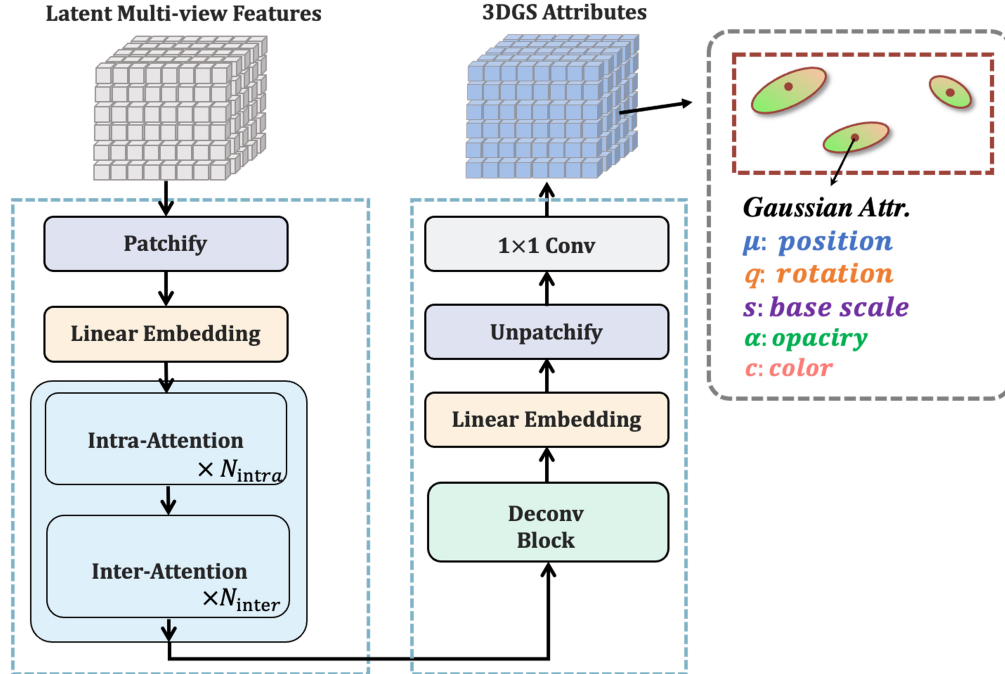
Figure 8: Detailed network architecture of latent reconstruction Transformer.

B   More Results

For practical applications demanding high input view reconstruction quality, which could potentially be directly derived from the input image. We improve the weights of the reconstruction loss for input views (referred to as **reweighting loss**), demonstrating that the current model with reweighting loss exhibits sufficient capacity to render input views with higher fidelity, as illustrated in Fig.9. We also conduct an ablation experiment for the reweighting loss function on the 2K2K Dataset [109]. There are slight declines in quantitative metrics, specifically, HumanSplat with reweighting loss achieves a numerical change of -0.141, -0.015, and +0.006 in PNSR, SSIM, and LPIPS. This suggests that reweighting loss is not "a bag of freebies" and HumanSplat without reweighting loss strikes a balance between the quality of novel views and the input views.

As for novel view and novel pose rendering, more results are shown in Fig. 10 and Fig. 11. For a single input image, we use Open-AnimateAnyone [123] for novel pose synthesis and HumanSplat for novel view synthesis. This functionality enables dynamic timestamps and real-time rendering of novel views, thereby enhancing immersive virtual exploration. Please also kindly refer to our supplementary video for more results.

C   Broader Impacts

While our model demonstrates the capability to reconstruct photo-realistic 3D avatars, it also introduces potential risks such as privacy violations. To mitigate these concerns, it is imperative to establish robust ethical guidelines and legal frameworks. This necessitates a collaborative effort among researchers, developers, and policymakers to promote the responsible use of this technology and safeguard against its potential misuse.
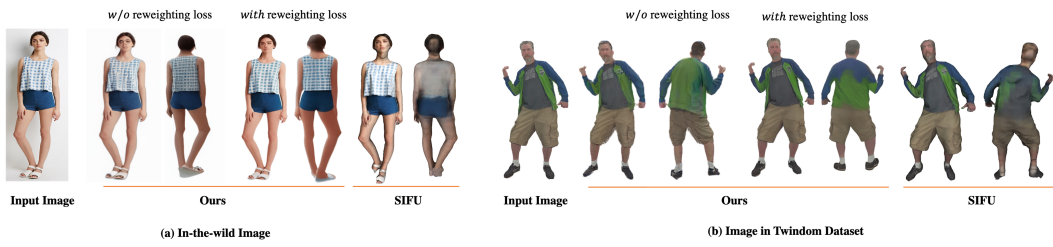
w/o reweighting loss   with reweighting loss   w/o reweighting loss   with reweighting loss

Input Image   Ours   SIFU   Input Image   Ours   SIFU

(a) In-the-wild Image   (b) Image in Twindom Dataset

Figure 9: **Ablation Study on Reweighting Loss.** Comparison of the original HumanSplat, trained with reweighting loss, and SIFU featuring texture.
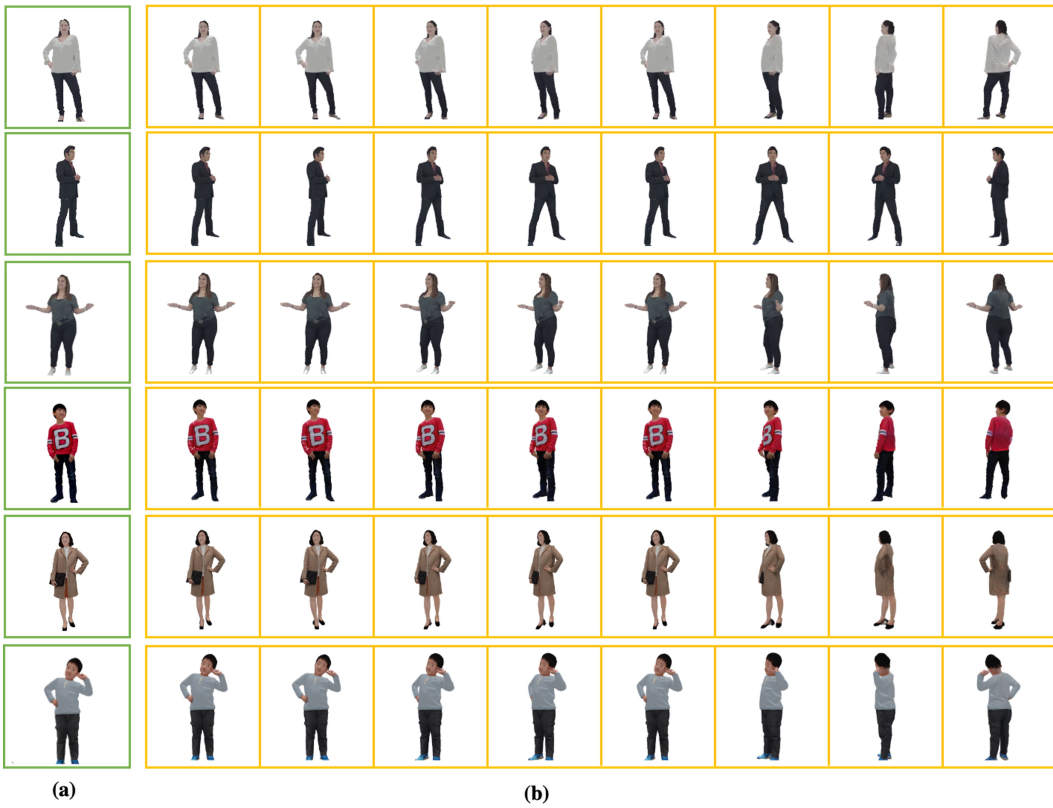

(a)   (b)

Figure 10: Qualitative 3D Gaussian Splatting results of diversified evaluation dataset. (a) Input Image. (b) Novel view Rendering Results.
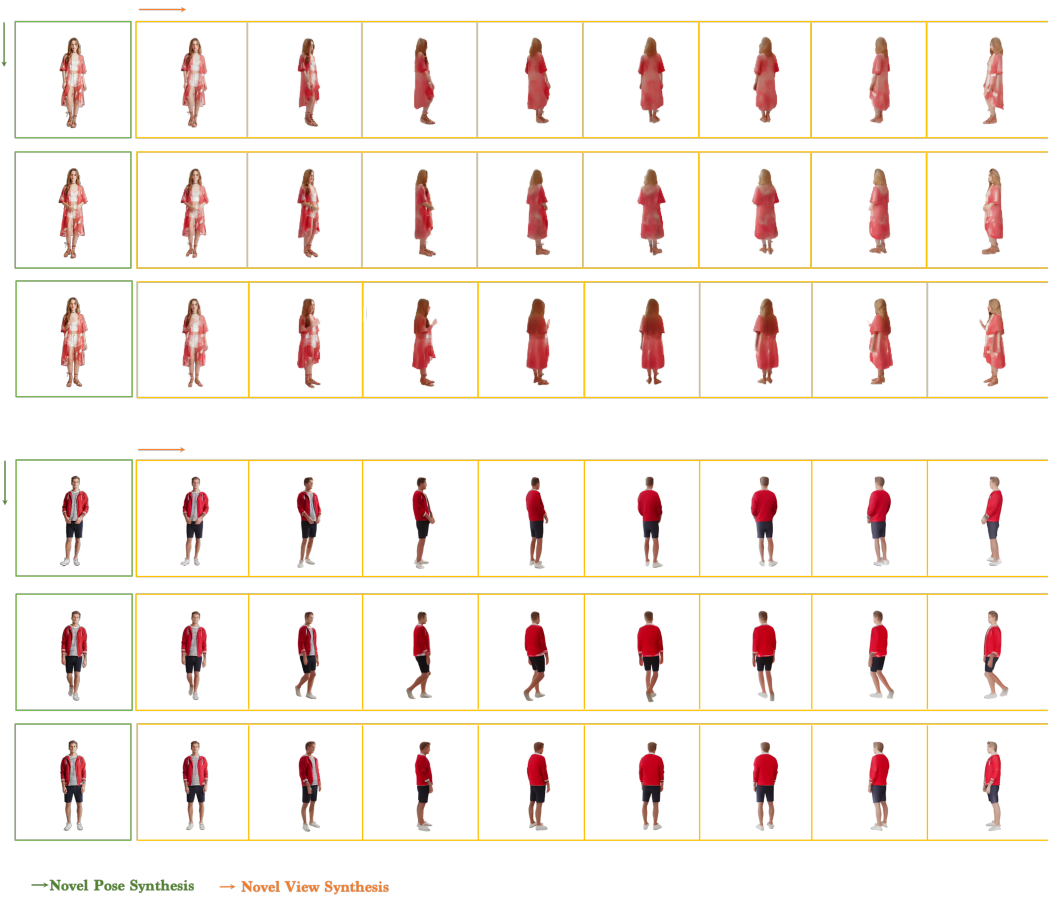
→ Novel Pose Synthesis   → Novel View Synthesis

Figure 11: Qualitative 4D Gaussian Splatting results on In-the-wild images, including novel view and pose rendering images. (Please zoom in for a detailed view)

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: See Sec. 5.

   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: [Yes]

Justification: Code will be available at: `https://humansplat.github.io`.

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: See Sec. 4.

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: See Sec. 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: See Sec. 4.1.

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: See Appendix C.

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 (e.g., code, data, models), used in the paper, are properly credited and are the license and terms of use 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 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: [Yes]

Justification: See Appendix C.

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.