DcSplat: Dual-Constraint Human Gaussian Splatting with Latent Multi-View Consistency

Tengfei Xiao[1,2], Yue Wu[1,2,*], Zhigang Gao[1,2], Yongzhe Yuan[1,2]
Can Qin[3], Hao Li[1,2], Mingyang Zhang[1,2]
[1]Xidian University
[2]MoE Key Lab of Collaborative Intelligence Systems, Xidian University
[3]Northeastern University
tfxiao@stu.xidian.edu.cn, ywu@xidian.edu.cn,{zhiganggao,yyz}@stu.xidian.edu.cn
qin.ca@northeastern.edu,{haoli,myzhang}@xidian.edu.cn

Abstract

Human Novel View Synthesis (HNVS) aims to synthesize photorealistic human images from novel viewpoints given observations from known views. Despite significant advances achieved by existing methods such as NeRF, diffusion models, and 3DGS, they still face substantial challenges in achieving stable modeling from a single image. In this paper, we introduce *Dual-Constraint Human Gaussian Splatting (**DcSplat**)*, a novel, simple, and efficient 3D Gaussian-based framework for single-view 3D human reconstruction. To address occlusion-induced texture missing and depth ambiguities, we introduce two key components: a Latent Multi-View Consistency Constraint Mechanism and a Geometric Constraint Module. The former employs a Latent Appearance Transformer (LatentFormer) to learn semantically coherent, view-consistent appearance priors via SMPL-guided pseudo-view fusion. The latter refines noisy SMPL-based depth through a U-Net-like structure conditioned on latent appearance features. These two modules are jointly optimized to generate high-quality Gaussian parameters in a unified latent space. Extensive experiments demonstrate that DcSplat outperforms existing SOTA methods in both geometry and texture quality, while achieving fast inference and lower computational cost.

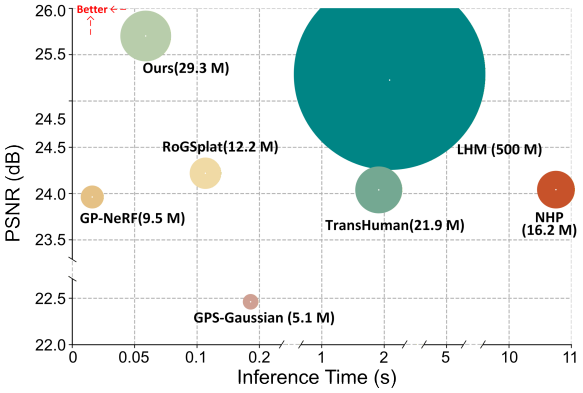

Figure 1: Comparison of different methods in terms of PSNR (dB) on the Real-World test set, inference time (s), and model parameters (M). Our method achieves a favorable trade-off between reconstruction quality and computational efficiency, outperforming existing approaches in both performance and speed.

Introduction

Human Novel View Synthesis (HNVS) aims to generate photorealistic images of human subjects from novel viewpoints, given one or more images captured from known views. This task is essential for a wide range of applications, including the Metaverse (Mystakidis 2022), digital humans (Yichao et al. 2023), and immersive augmented/virtual reality (AR/VR) systems (Carmigniani et al. 2011; Burdea and Coiffet 2003). However, the human body, as a highly non-rigid object, exhibits complex structure, dynamic deformations, diverse clothing, and fine-grained textures, which pose significant challenges for HNVS, including severe self-occlusions, incomplete information, and geometric ambiguities. Especially in the single-view condition, accurate modeling becomes a non-trivial problem due to the limited a priori information of the single view and the lack of geometric constraints between multiple views.

Recent advancements in neural implicit representations (Peng et al. 2021; Liu et al. 2023; Sun et al. 2024) have significantly advanced the field of HNVS. In particular, Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) utilize continuous volumetric representations to enable photorealistic rendering from posed RGB inputs. Despite its impressive performance, NeRF typically requires object-specific overfitting and dense sampling along camera rays, resulting in poor generalization and high computational overhead. Although follow-up works have explored acceleration techniques (Müller et al. 2022; Fridovich-Keil et al. 2022) and sampling reduction strategies (Peng et al. 2021; Shao et al. 2023), the fine-grained querying process remains a major bottleneck, limiting its applicability in real-time or large-scale scenarios.

More recently, diffusion models (Rombach et al. 2022; Zhang et al. 2025) have demonstrated strong generative capabilities and have been adapted to view synthesis tasks by conditioning on human pose or view-specific features (Xiao

*Corresponding author.

et al. 2025b; Pan et al. 2024; Hu 2024; Yang et al. 2024; Liu et al. 2024a; Zhang, Yang, and Yang 2024; Zhang et al. 2026, 2024). While these models offer high flexibility, they often suffer from structural inconsistency, high inference latency, and require subject-specific fine-tuning, which restricts their scalability in practical deployments.

To address these limitations, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has emerged as a promising alternative, offering high-quality rendering with real-time inference. In this paradigm, the scene is represented as a set of learnable anisotropic gaussian primitives, each parameterized by position, color, opacity, and a covariance matrix encoding spatial extent and orientation. Furthermore, the use of a-blending (Kopanas et al. 2022) enables stable gradient propagation and differentiable image synthesis. However, most existing 3DGS-based methods (Zheng et al. 2024; Xiao et al. 2025a; Qiu et al. 2025; Xu et al. 2025) rely heavily on multi-view supervision to construct high-fidelity point cloud initializations. Under single-view conditions, learning stable Gaussian representations remains a significant challenge, often leading to incomplete geometry and unresolved structural ambiguities.

In this paper, we propose **DcSplat**, a simple yet effective framework based on 3D Gaussian representations for synthesizing high-fidelity human novel views from a single input image, without requiring per-instance optimization. The core challenge under single-view conditions lies in the inherent incompleteness of observations: self-occluded regions are unobservable, and the absence of reliable geometric structure and depth cues renders the reconstruction problem highly ill-posed.

To mitigate these challenges, we introduce the *Latent Multi-View Consistency Constraint Mechanism*, which utilizes SMPL (Loper et al. 2023) priors inferred from the input image to generate pseudo views and enforce consistency in the latent appearance space. At the heart of this mechanism is the proposed **Latent Appearance Transformer (LatentFormer)**, which employs cross-attention to fuse pose-aware structural cues from pseudo views with image features. This enables the network to learn semantically aligned and view-consistent latent appearance priors, which guide the regression of 3D gaussian parameters. Furthermore, to address the low reliability of initial point clouds generated from SMPL-based depth maps, we propose a **Geometric Constraint Module (GCM)**. This module employs a U-Net-like architecture to iteratively refine coarse geometry, incorporating latent appearance cues to enhance object boundary localization and handle occlusions effectively. Such refinement leads to more accurate initialization of gaussian primitives and improved final rendering quality.

By jointly modeling geometry and appearance in the latent space, DcSplat enables efficient regression of high-quality 3D Gaussian representations. Extensive experiments demonstrate that DcSplat surpasses prior SOTA approaches in both geometric accuracy and texture fidelity. Furthermore, benefiting from its lightweight design, our method requires only **0.07s** per inference, while reducing model parameters by **16**$\times$ and improving inference speed by **27**$\times$ compared to recent single-view methods LHM (Qiu et al. 2025). A comprehensive comparison is shown in Figure 1. In summary, our contributions are:

- We introduce DcSplat, a novel, simple and effective 3D Gaussian Splatting method for fast synthesis of high-fidelity novel views of people from single-view images without the need to optimize each object individually.
- We propose a Latent Multi-View Consistency Constraint Mechanism, which constructs SMPL-based pseudo-view priors and enforces consistency modeling via a Latent Appearance Transformer.
- We introduce a Geometric Constraint Module that progressively refines coarse depth maps to provide high-quality initialization point clouds for 3DGS.
- Conducting extensive experiments on the THuman2.1 (Yu et al. 2021), RenderPeople (Renderpeople 2026), and Real-World (Zheng et al. 2024) datasets, achieving SOTA results in both quality and fidelity.

Related Work

Accurate 3D modeling from limited 2D observations is critical for human novel view synthesis (HNVS). Early methods (Chen and Williams 2023; Oh et al. 2001; Yuan et al. 2023, 2025) rely on dense multi-view inputs or accurate 3D priors, generating novel views through image-based rendering and blending strategies. However, these approaches degrade significantly under sparse-view or single-view conditions, where occlusions and geometric ambiguities cannot be resolved effectively.

Neural Radiance Fields (NeRF) have become a foundational approach for HNVS due to their continuous volumetric representation. IBRNet (Wang et al. 2021) proposes a ray-based Transformer that estimates radiance and density at continuous 5D coordinates (3D position + 2D view direction), enabling effective cross-view appearance aggregation. GP-NeRF (Chen et al. 2022) integrates depth priors for geometry-guided multi-view fusion and employs progressive sampling to improve rendering efficiency. NHP (Kwon et al. 2021) enhances generalization under sparse-view settings by combining temporal and multi-view Transformers with a parametric human model. TransHuman (Pan et al. 2023) learns canonical-space SMPL textures and part-level dependencies using a Transformer-based architecture, improving robustness across diverse human poses and clothing variations.

Diffusion models have recently been adopted for view synthesis due to their strong generative capacity. DPA-Gen (Xiao et al. 2025b) and AnimateAnyone (Hu 2024) generate new views by conditioning diffusion on target poses, but they operate purely in 2D image space and lack explicit 3D reasoning, limiting their performance on complex poses or loose garments. To improve realism and consistency, Yang *et al.* (2024) and Liu *et al.* (2024a) incorporate 3D modeling into the diffusion pipeline. Nevertheless, their high computational cost and memory footprint hinder scalability, especially in high-resolution settings. Recent methods have begun integrating diffusion with structured 3D representations: HumanGaussian (Liu et al. 2024b) utilizes structure-aware score distillation sampling (SDS) (Poole et al. 2022)

and adaptive density control to generate high-quality 3D humans. HumanSplat (Pan et al. 2024) synthesizes multi-view images via diffusion, followed by a reconstruction Transformer to infer 3D Gaussian parameters, bridging generative modeling and explicit geometry.

To balance rendering quality and inference efficiency, explicit representations based on 3D Gaussian Splatting (3DGS) have gained traction. These methods model scenes using learnable anisotropic Gaussians, enabling structured, real-time rendering. GPS-Gaussian (Zheng et al. 2024) directly regresses Gaussian parameters from input images by predicting a view-conditioned parameter map, eliminating the need for subject-specific optimization. RoGSplat (Xiao et al. 2025a) builds upon this approach by using SMPL mesh vertices as coarse 3D positions and refining the predicted Gaussians through multi-level feature fusion. By incorporating both pixel- and voxel-level cues, it achieves fine-grained, high-resolution reconstructions. While promising, these methods still face challenges under single-view conditions due to limited geometric supervision. In this work, we construct multiple pseudo views in the latent space to compensate for the limited observational information under single-view settings. Building on this, we further design a dual-constrained human Gaussian Splatting framework to achieve accurate and efficient novel view synthesis.

Preliminary

Human Parametric Model. In HNVS tasks, the Skinned Multi-Person Linear (SMPL) model (Loper et al. 2023) parametric human body can provide geometric priors for human representation. Theoretically, the human mesh $\mathcal{M}$ can be defined using shape parameters $\beta \in \mathbb{R}^{10}$, which control body size, and pose parameters $\theta \in \mathbb{R}^{3\times 24}$, which adjust joint positions and orientations:

$$\mathcal{M}(\beta, \theta) : \beta \times \theta \longrightarrow \mathbb{R}^{3\times 6890}. \quad (1)$$

In particular, SMPL-X (Pavlakos et al. 2019a) extends SMPL by adding facial and hand components, enabling more accurate representation of facial expressions and finger movements.

3D Gaussian Splatting. 3D Gaussian splatting (Kerbl et al. 2023) is an explicit point-based 3D representation method that models a 3D scene as a set of Gaussian points with anisotropic covariances. Each Gaussian point is parameterized by its center position in 3D space $\mu_i \in \mathbb{R}^3$, covariance matrix $\sum_i \in \mathbb{R}^{3\times 3}$, color $c_i \in \mathbb{R}^3$, and opacity coefficient $\alpha_i \in [0, 1]$. To enable real-time rendering of 2D images from novel viewpoints, the parameters of each point are encoded into a corresponding Gaussian feature, which is then projected onto the 2D image plane in a depth-aware order.

Dual-Constraint Human Gaussian Splatting

Architecture and Overview

The overall architecture of our proposed framework, **DcSplat**, is illustrated in Figure 2. Given a single human image I_s, we first estimate a coarse 3D body prior from the input image using an SMPL estimator (Feng et al. 2021; Pavlakos et al. 2019b). Leveraging the predicted geometry, we render multiple pseudo-views from side and back perspectives, providing additional structural cues under the single-view setting. To model consistent appearance across views, we propose the *Latent Appearance Transformer (LatentFormer)*, which employs a cross-attention mechanism to fuse semantic geometry and visual features from both the source and pseudo views. This fusion enables the network to learn a semantically aligned and view-consistent appearance embeddings in the latent space. Subsequently, we integrate a GCM to progressively refine the coarse depth maps, ultimately producing an accurate 3D point cloud. Finally, the accurate point cloud is used as a high-quality geometric initialization for 3DGS. Combined with the learned appearance embeddings, DcSplat regresses the parameters of anisotropic 3D Gaussians, producing high-fidelity novel views with real-time rendering performance.

Latent Multi-View Consistency Constraint Mechanism

Most existing HNVS methods rely on explicitly captured multi-view images to achieve structurally consistent and photorealistic 3D human modeling. However, the high acquisition cost and limited flexibility of multi-view setups hinder their practical deployment. In contrast, we focus on single-view 3D human reconstruction, aiming to learn view-consistent appearance embeddings without requiring multi-view input. This task is inherently ill-posed due to the lack of complete geometry and depth information in a single image.

To address this, we propose a Latent Multi-View Consistency Constraint Mechanism that enforces cross-view appearance consistency in the latent space using structural cues from multiple pseudo views. Specifically, we introduce a Latent Appearance Transformer, which jointly models semantic priors and geometric features from both the source view and pseudo views. Guided by 2D pose and 3D depth information of the pseudo-view, the module learns a semantically aligned and view-consistent latent representation.

Latent Appearance Transformer. In recent years, the Transformer architecture has demonstrated significant advantages over convolutional neural networks (CNNs) in modeling long-range dependencies and capturing rich contextual representations. These capabilities are particularly well-suited for vision tasks involving complex geometric structures or high-dimensional spatial transformations. However, due to the inherent ambiguity and lack of depth information in a single-view image, achieving cross-view appearance consistency remains a fundamental challenge in HNVS. To address this issue, we propose constructing pseudo-view features in a latent space to enhance the modeling capacity for cross-view appearance consistency. Toward this goal, we introduce the Latent Appearance Transformer (LatentFormer), a carefully designed module based on the Transformer architecture that enables efficient modeling of pseudo-view appearance features in latent space, as illustrated in Figure 2. For clarity, normalization layers and residual connections are omitted in the figure. The detailed structure of LatentFormer is shown in the Appendix. LatentFormer comprises three main components: the Source View Enhancer (SVE), the Depth Map Enhancer (DME),

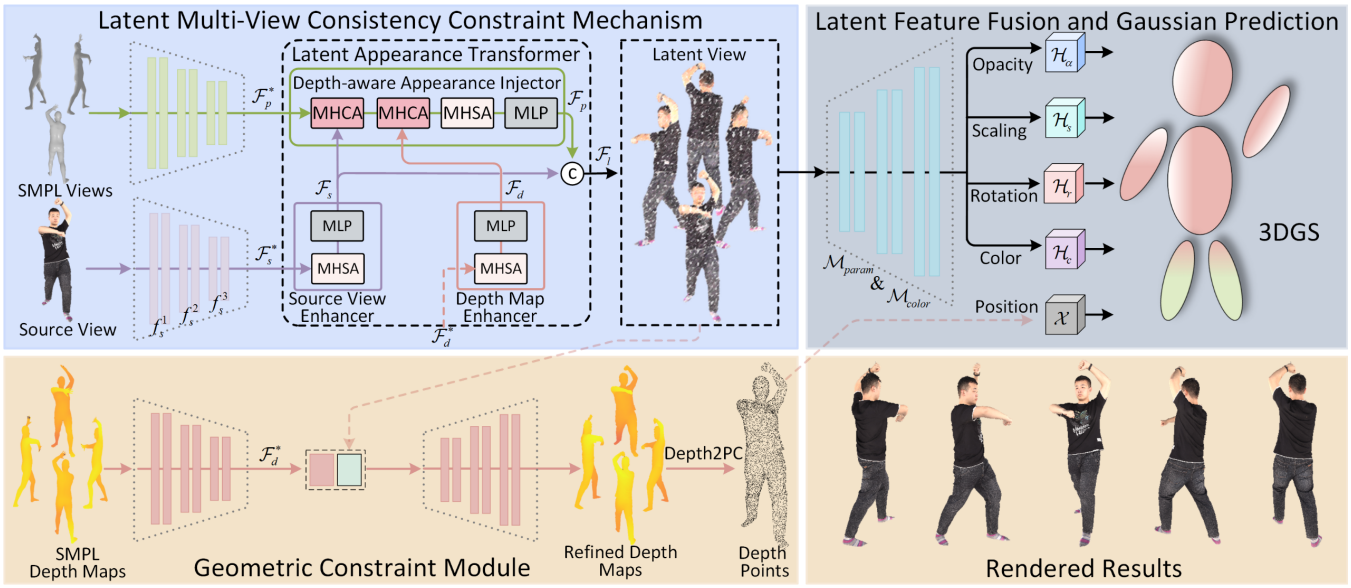

Figure 2: Overview of the proposed **DcSplat** framework. Starting from a single-view human image, we first estimate the corresponding SMPL model to render pseudo-view images and depth maps. The **Latent Appearance Transformer** fuses appearance features and geometric cues via attention mechanisms, producing view-consistent appearance embeddings. A **Geometric Constraint Module (GCM)** refines the initial depth maps to generate a more accurate point cloud, which serves as the geometric basis for 3D Gaussian parameter regression. For clarity, the two modules $\mathcal{M}_{\text{param}}$ and $\mathcal{M}_{\text{color}}$ are illustrated as a single block, and skip connections, residual operations, and normalization layers are omitted.

and the Depth-aware Appearance Injector (DAI). The SVE processes the input image to extract enhanced appearance features $\mathcal{F}_s$ from the initial features $\mathcal{F}_s^*$. The DME refines the depth features $\mathcal{F}_d^*$ rendered from SMPL-based pseudo views, producing reliable geometric priors $\mathcal{F}_d$.

At the core of LatentFormer lies the Depth-aware Appearance Injector (DAI), which integrates appearance and geometry across views. DAI consists of two cascaded Multi-Head Cross Attention (MHCA) blocks, followed by a Multi-Head Self Attention (MHSA) block and a Multi-Layer Perceptron (MLP). The first MHCA fuses source-view appearance $\mathcal{F}_s$ with pseudo-view queries $\mathcal{F}_p^*$ based on camera pose, enabling spatial propagation of features to unseen viewpoints. The second MHCA injects geometric structure by attending to $\mathcal{F}_d$, thereby enriching appearance modeling with depth-aware priors. MHSA and MLP further consolidate appearance and geometry, producing a unified latent embedding $\mathcal{F}_p = DAI(\mathcal{F}_p^*, \mathcal{F}_s, \mathcal{F}_d)$. Finally, the enhanced source-view features $\mathcal{F}_s$ are concatenated with the view-consistent pseudo-view features $\mathcal{F}_p$ along the view dimension to form the final latent representation $\mathcal{F}_l = Concat_v(\mathcal{F}_s, \mathcal{F}_p)$, which is then used to guide 3D Gaussian parameter regression.

Geometric Constraint Module

In order to generate high-quality 3D point clouds that can be used for 3D Gaussian Splatting, we propose a Geometric Constraint Module (GCM) for further refinement of the depth maps obtained from the SMPL model rendering. Inspired by Rogsplat (Xiao et al. 2025a) and GPS-Gaussian (Zheng et al. 2024), the module employs a U-Net-like encoding-decoding structure to enhance the geometric representation layer by layer and recover the spatial details to improve the accuracy and completeness of the depth map. Meanwhile, potential appearance features $\mathcal{F}_l$ from Latent Appearance Transformer are introduced in the bottleneck stage of the U-Net as an appearance prior to bootstrap geometric features. This helps to improve the network's ability to perceive geometric details such as boundary regions and occlusion regions, and alleviate the ambiguity problem caused by single structural information. In addition, through the jump connection mechanism in U-Net, low-level local details and high-level semantic geometric information are fully integrated. Finally, the projection matrix is constructed to map the refined depth map to 3D space by combining the camera parameters. Details are given in the Appendix.

Latent Feature Fusion for Gaussian Parameters Prediction

In this section, we present our approach for predicting the parameters of 3D Gaussians within the latent space. As mentioned in Section , the feature of each Gaussian point in 3D space includes attributes such as 3D position $\mathcal{X}$, color c, rotation r, scale s, and opacity α. Following the findings of GPS-Gaussian (Zheng et al. 2024), we recognize that both global contextual cues and local geometric structures play a vital role in parameter prediction. To this end, we integrate the multi-view latent features $\mathcal{F}_l$ from the Latent Appearance Transformer with the refined 3D positions $\mathcal{X}$ obtained from the Geometric Constraint Module. To effectively fuse multi-view latent features $\mathcal{F}_l$ and obtain a comprehensive

Gaussian representation, we design a CNN-based feature aggregation module, denoted as $\mathcal{M}_{\text{param}}$. This module produces a unified global feature map $\mathcal{F}_g$, which serves as the input to the subsequent parameter prediction heads. The estimation of rotation r, scale s, and opacity α is performed as follows:

$$\mathcal{F}_g = \mathcal{M}_{\text{param}}(\mathcal{F}_l), \tag{2}$$

$$[r,\, s,\, \alpha] = [\mathcal{N},\, \phi,\, \sigma]\left(\mathcal{H}_{r,s,\alpha}\left(\mathcal{F}_g;\, \mathcal{X}\right)\right). \tag{3}$$

Here, $\mathcal{H}r$, $\mathcal{H}s$, and $\mathcal{H}_\alpha$ denote the parameter heads responsible for predicting rotation r, scale s, and opacity α, respectively. $\mathcal{N}(\cdot)$, $\phi(\cdot)$, and $\sigma(\cdot)$ denote the normalization, Softplus, and Sigmoid functions applied to each output branch. Unlike GPS-Gaussian, which leverages explicit multi-view RGB supervision to regress spherical harmonics (SH) color coefficients, our method operates under a single-view setting, where such multi-view color cues are unavailable. To compensate for this, we introduce a color-specific prediction module $\mathcal{M}_{\text{color}}$, reusing the architectural design of $\mathcal{M}_{\text{param}}$ with dedicated input features. As shown in Figure 2, we unify $\mathcal{M}_{\text{param}}$ and $\mathcal{M}_{\text{color}}$ into a single block for simplicity, omitting skip connections. The module predicts RGB information as follows:

$$\begin{aligned} \mathcal{F}_c &= \mathcal{M}_{\text{color}}(\mathcal{F}_l; f_s^1, f_s^2, f_s^3), \\ c &= \sigma(\mathcal{H}_c\left(\mathcal{F}_c; \mathcal{X}\right)), \end{aligned} \tag{4}$$

where f_s^1, f_s^2, and f_s^3 denote multi-scale features from the source view to enhance appearance representation and $\mathcal{H}_c$ is the color prediction head estimating the color c. Finally, the complete set of predicted Gaussian parameters for human representation is defined as:

$$\mathcal{G} = \{\mathcal{X}, c, r, s, \alpha\}. \tag{5}$$

Learning Objectives

To optimize the performance of DcSplat, we design a set of supervision signals that jointly guide both appearance reconstruction and geometric refinement. The overall training objective is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{rec}}^r + \mathcal{L}_{\text{ssim}}^r + \mathcal{L}_{\text{per}}^r + \mathcal{L}_{mae}^d. \tag{6}$$

The term $\mathcal{L}_{\text{rec}}^r$ denotes the pixel-wise MAE loss, adopted as the primary reconstruction objective due to its robustness to outliers and its ability to preserve fine image details by mitigating blurring and over-smoothing artifacts. To further enhance structural consistency, we incorporate the Structural Similarity (SSIM) (Wang et al. 2004) loss $\mathcal{L}_{\text{ssim}}^r$. However, SSIM alone is insufficient for capturing the perceptual realism of the generated images. Therefore, we additionally introduce the perceptual (Johnson, Alahi, and Fei-Fei 2016) loss $\mathcal{L}_{\text{per}}^r$, which measures semantic similarity in a learned high-dimensional feature space, thereby improving perceptual quality. In addition, for supervising the geometric refinement process, we employ a separate MAE loss on the predicted depth map, denoted as $\mathcal{L}_{mae}^d$, to promote accurate 3D position estimation.

Experiments

Setup

Datasets and Metrics. We conduct extensive experiments on THuman2.1 (Yu et al. 2021), RenderPeople (Renderpeople 2026), and Real-world (Zheng et al. 2024) datasets to validate the effectiveness and generalizability of the proposed method. The THuman2.1 dataset consists of 2,500 high-quality 3D human scans captured by a high-density DSLR system. It features a wide variety of clothing appearances and complex body poses, providing a strong benchmark for evaluating the fidelity of appearance and pose reconstruction. RenderPeople provides 482 synthetic 3D human models with accurate geometry and photorealistic textures. Following the setup of Rogsplat (Xiao et al. 2025a), we adopt the version preprocessed by SHERF (Hu et al. 2023), which serves as a standardized benchmark for evaluating reconstruction performance on clean and controlled scans. To further evaluate the cross-domain generalization ability of our approach, we conduct additional experiments on a Real-world Dataset collected by GPS-Gaussian (Zheng et al. 2024). This dataset includes in-the-wild human images with diverse backgrounds and natural variations, posing greater challenges for novel view synthesis. For quantitative evaluation, we employ a set of widely-used metrics, including the Structural Similarity Index (SSIM) (Wang et al. 2004), Peak Signal-to-Noise Ratio (PSNR) (Wang et al. 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), and pixel-wise L1 metric. Additional details are provided in the appendix.

Implementation Details. Our DcSplat framework is implemented based on PyTorch (Paszke et al. 2019) and trained on a single NVIDIA RTX 5880 GPU with a batch size of 8. We adopt the AdamW (Loshchilov and Hutter 2018) optimizer with an initial learning rate of 1×10^{-4}. Following the learning rate warm-up strategy in DPAGen (Xiao et al. 2025b), we gradually increase the learning rate during the first 1000 iterations, and subsequently decay it to 1×10^{-5} over 30 training epochs. For the Latent Appearance Transformer, we build upon the implementation from DPAGen, with key architectural modifications. In particular, to enhance the stylistic consistency of synthesized novel views, we replace the Layer Normalization (Ba, Kiros, and Hinton 2016) in the attention module with Instance Normalization (Ulyanov, Vedaldi, and Lempitsky 2016). Furthermore, both the downsampling and upsampling components throughout the framework are implemented using ResNet-based blocks (He et al. 2016), which offer effective feature extraction and reconstruction capabilities. The overall training objective incorporates four weighted loss terms, with corresponding coefficients set to 0.8, 0.2, 0.01, and 0.1, respectively.

Impact of Pseudo-View Configuration

Single-view images inherently suffer from limitations in geometric structure and appearance detail, making it difficult to provide sufficient prior information for robust 3DGS reconstruction. To address this issue, we propose generating multiple pseudo-views to introduce additional geometric cues and appearance references. As noted by Li *et al.* (2024),

Approaches	Thuman2.1				Real-World data				Overhead	
	SSIM↑	PSNR↑	LPIPS↓	$\mathcal{L}_1$ ↓	SSIM↑	PSNR↑	LPIPS↓	$\mathcal{L}_1$ ↓	#Param↓	Time ↓
NHP	0.931	25.573	0.0766	0.0094	0.901	24.087	0.1007	0.0132	16.2 M	10.9 s
GP-NeRF	0.929	25.660	0.0743	0.0104	0.909	23.983	0.1065	0.0178	9.5 M	**0.02 s**
TranHuman	0.946	25.981	0.0699	0.0082	0.913	24.104	0.0995	0.0129	21.9 M	1.7 s
GPS-Gaussian	0.885	22.398	0.0863	0.0169	0.897	22.464	0.0880	0.0162	**5.1 M**	0.18 s
RoGSplat	0.952	**26.748**	0.0804	**0.0062**	0.917	24.209	0.1465	0.0146	12.2 M	0.11 s
LHM†	-	-	-	-	0.931	25.270	0.0732	0.0091	500 M	2.01 s
Ours	**0.960**	26.639	**0.0452**	0.0068	**0.943**	**25.753**	**0.0646**	**0.0070**	29.3 M	0.07 s

Table 1: Quantitative comparison of various methods trained on the THuman2.1 dataset and evaluated on both the THuman2.1 (in-domain) and Real-World (cross-domain) datasets. Note that the LHM† method is not trained on the THuman2.1 dataset, but on a mixed dataset of 300K video frames and 5K 3D scans.

Model	SSIM(↑)	PSNR(↑)	LPIPS(↓)	L_1 (↓)
GP-NeRF	0.938	27.099	0.0950	0.0084
NHP	0.940	27.054	0.0737	0.0069
TransHuman	0.947	27.502	0.0716	0.0061
RoGSplat	0.952	27.659	0.0650	0.0054
Ours	**0.967**	**27.919**	**0.0500**	**0.0045**

Table 2: Quantitative comparison of the proposed method with other SOTA methods on the RenderPeople dataset.

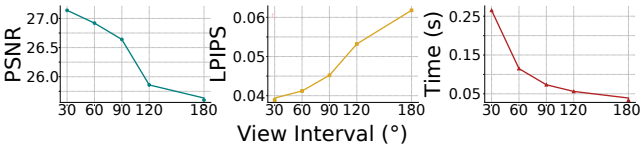

Figure 3: Synthesis performance on THuman2.1 with varying numbers of pseudo-views (90° spacing).

while the incorporation of pseudo-views can enhance the model's ability to synthesize occluded regions and improve fidelity, an excessive number of views may introduce redundant or inconsistent visual information and significantly increase the computational cost during training and inference. Therefore, balancing the number of pseudo-views with their supervision effectiveness is critical. In this subsection, we investigate how the angular spacing and quantity of pseudo-views affect the quality of novel view synthesis. To this end, we design five configurations with decreasing angular intervals—30°, 60°, 90°, 120°,and 180°—corresponding to progressively denser pseudo-view settings. As shown in Figure 3, denser configurations (e.g., 30° and 60°) lead to substantial performance improvements across various metrics, indicating that finer angular resolution facilitates better geometric reasoning and texture reconstruction. However, these performance gains come at the cost of a significantly increased computational burden. For instance, the inference time at 30° is approximately 8× that of the 180° configuration, which severely compromises real-time applicability. In contrast, sparse configurations (e.g., 180°) offer higher efficiency but exhibit suboptimal performance in occluded regions and structural detail synthesis. Medium-density configurations (90° or 60°) achieve the best trade-off between synthesis quality and computational efficiency. In this paper, we use 90° as the default training setting.

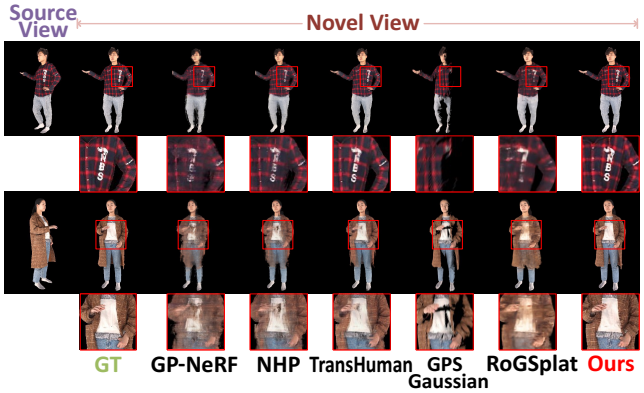

Figure 4: Qualitative comparison of in-domain generalization on the THuman2.1 dataset. Zoom in for details.

Comparisons with State-of-the-art Methods

Tables 1 and 2, along with Figures 4, 5 and 6, present both quantitative and qualitative comparisons between our proposed method and several SOTA approaches for novel human view synthesis. We compare our method against a variety of representative baselines, including NeRF-based methods such as GP-NeRF (Chen et al. 2022), NHP (Kwon et al. 2021), and TransHuman (Pan et al. 2023); 3D Gaussian-based approaches like GPS-Gaussian (Zheng et al. 2024) and RoGSplat (Xiao et al. 2025a); as well as the recently proposed LHM (Qiu et al. 2025) method.

Quantitative Comparisons. Specifically, Table 1 reports the performance of all methods trained on the scanned dataset THuman2.1 (Yu et al. 2021) and tested on both the THuman2.1 test split and a cross-domain Real-World (Zheng et al. 2024) dataset. It is important to note that since LHM does not provide training code, we use its officially released pretrained model, which was trained on a mixed dataset consisting of 300K videos and 5K 3D scans, rather than THuman2.1. Additionally, Table 2 presents evaluation results on the synthetic RenderPeople (Renderpeople 2026) dataset, offering a clean and controlled environment to further verify the method's performance. As shown in Tables 1 and 2, although our method has a relatively larger number of

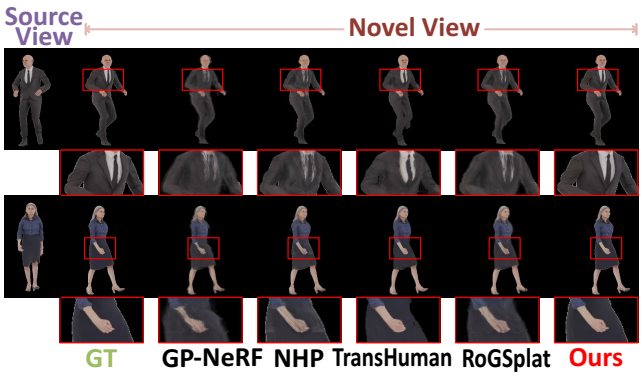

Figure 5: Qualitative comparison of in-domain generalization on the RenderPeople dataset. Zoom in for details.

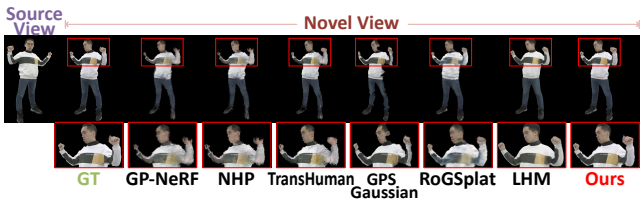

Figure 6: Qualitative cross-domain generalization on Real-World dataset. All results are from models trained on THuman2.1 dataset, except LHM. Zoom in for details.

parameters compared to traditional multi-view methods, it consistently outperforms most existing approaches in terms of inference speed and novel view synthesis quality.

Qualitative Comparisons. Figures 4, 5, and 6 visualize the novel view synthesis results across the THuman2.1, Real-World, and RenderPeople datasets. As illustrated, our method surpasses other leading approaches in terms of pose accuracy, appearance consistency, and visual realism. Notably, NeRF-based multi-view methods (e.g., GP-NeRF, NHP, and TransHuman) tend to suffer from inaccurate geometry due to their limited handling of inter-view misalignment, resulting in artifacts or holes in the rendered outputs. GPS-Gaussian relies heavily on depth cues from the input view for geometry estimation, leading to substantial texture loss and artifacts when the synthesized view significantly deviates from the input. Although the LHM method can produce plausible results based on a single-view image and the SMPL template, it lacks a dedicated mechanism to ensure consistency in occluded regions. As a result, it often borrows textures directly from visible areas, resulting in limited diversity and realism in unseen parts. See Appendix for more qualitative comparisons.

Ablation Studies

To evaluate the effectiveness of each proposed component, we conduct a comprehensive ablation study. Quantitative results and qualitative comparisons are presented in the Table 3 and Figure 7, respectively. Removing the pseudo-view guidance leads to a model that is completely dependent on a single observation, resulting in a severe reduction in the ability to synthesize new views. This suggests that pseudo-views are critical for mitigating view ambiguity and inferring occluded regions. We further analyze the roles of the Depth-aware Appearance Injector and the Geometric Constraint Module. DAI promotes the alignment of cross-view features with source-view features, ensuring that latent appearance representations remain consistent across different viewpoints. It also guides the smooth transition of appearance information from observed to occluded regions, enabling coherent novel view synthesis. GCM refines the coarse geometry and provides reliable 3D priors for Gaussian parameter regression. As illustrated in the Figure 7, removing GCM leads to the most pronounced misalignments and rendering artifacts, highlighting that accurate geometric initialization is crucial for high-fidelity novel view synthesis. These findings are consistent with observations from Zheng *et al.* (2024), which emphasize that accurate global context and fine-grained 3D structures are essential for high-quality Gaussian parameter regression.

Model	SSIM(↑)	PSNR(↑)	LPIPS(↓)	L_1 (↓)
w/o Pseudo Views	0.9308	23.873	0.0728	0.0152
w/o DAI	0.9523	25.951	0.0518	0.0071
w/o GCM	0.9431	25.606	0.0612	0.0081
Full	**0.9595**	**26.639**	**0.0452**	**0.0068**

Table 3: Quantitative ablation study on the THuman2.1 dataset.

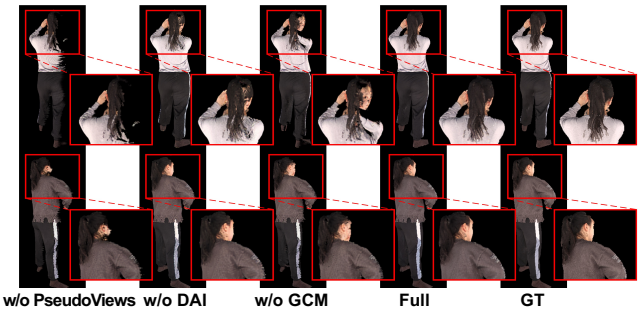

Figure 7: Quantitative Comparison of ablation results on the THuman2.1 dataset.

Conclusion

This paper presents DcSplat, an efficient 3D Gaussian-based framework for novel view synthesis of human subjects from a single image. To address the limitations of single-view inputs in capturing geometry and texture, DcSplat introduces a latent multi-view consistency modeling mechanism and a geometric constraint module, enhancing synthesis quality from both appearance and structural perspectives. DcSplat demonstrates superior generalization and real-time performance across diverse real-world and synthetic datasets. Future work will extend single-view human synthesis to more challenging conditions, including facial details, loose clothing, complex environments, and crowded scenes, enhancing robustness and real-world applicability.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62036006, 62276200) and the Innovation Capability Support Plan of Shaanxi Province (2023KJXX-144).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Burdea, G. C.; and Coiffet, P. 2003. *Virtual reality technology*. John Wiley & Sons.

Carmigniani, J.; Furht, B.; Anisetti, M.; Ceravolo, P.; Damiani, E.; and Ivkovic, M. 2011. Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51: 341–377.

Chen, M.; Zhang, J.; Xu, X.; Liu, L.; Cai, Y.; Feng, J.; and Yan, S. 2022. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *European Conference on Computer Vision*, 222–239. Springer.

Chen, S. E.; and Williams, L. 2023. View interpolation for image synthesis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 423–432.

Feng, Y.; Choutas, V.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2021. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, 792–804. IEEE.

Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5501–5510.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, L. 2024. 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*, 8153–8163.

Hu, S.; Hong, F.; Pan, L.; Mei, H.; Yang, L.; and Liu, Z. 2023. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9352–9364.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kopanas, G.; Leimkühler, T.; Rainer, G.; Jambon, C.; and Drettakis, G. 2022. Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics (TOG)*, 41(6): 1–15.

Kwon, Y.; Kim, D.; Ceylan, D.; and Fuchs, H. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34: 24741–24752.

Li, Y.; Xiao, T.; Geng, L.; and Wang, J. 2024. Direct may not be the best: an incremental evolution view of pose generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3270–3278.

Liu, M.; Shi, R.; Chen, L.; Zhang, Z.; Xu, C.; Wei, X.; Chen, H.; Zeng, C.; Gu, J.; and Su, H. 2024a. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10072–10083.

Liu, P.; Zhang, G.; Zhang, S.; Li, Y.; and Zeng, Z. 2023. Skeleton-aware implicit function for single-view human reconstruction. *CAAI Transactions on Intelligence Technology*, 8(2): 379–389.

Liu, X.; Zhan, X.; Tang, J.; Shan, Y.; Zeng, G.; Lin, D.; Liu, X.; and Liu, Z. 2024b. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6646–6657.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Loshchilov, I.; and Hutter, F. 2018. Decoupled weight decay regularization. *International Conference on Learning Representations*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Mystakidis, S. 2022. Metaverse. *Encyclopedia*, 2(1): 486–497.

Oh, B. M.; Chen, M.; Dorsey, J.; and Durand, F. 2001. Image-based modeling and photo editing. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 433–442.

Pan, P.; Su, Z.; Lin, C.; Fan, Z.; Zhang, Y.; Li, Z.; Shen, T.; Mu, Y.; and Liu, Y. 2024. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *Advances in Neural Information Processing Systems*, 37: 74383–74410.

Pan, X.; Yang, Z.; Ma, J.; Zhou, C.; and Yang, Y. 2023. Transhuman: A transformer-based human representation for generalizable neural human rendering. In *Proceedings of the IEEE/CVF International conference on computer vision*, 3544–3555.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019a. 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*, 10975–10985.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019b. 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*, 10975–10985.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9054–9063.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qiu, L.; Gu, X.; Li, P.; Zuo, Q.; Shen, W.; Zhang, J.; Qiu, K.; Yuan, W.; Chen, G.; Dong, Z.; et al. 2025. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*.

Renderpeople. 2026. Renderpeople. https://renderpeople.com. Accessed: 2026-01-12.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.

Sun, W.; Li, M.; Li, P.; Cao, X.; Meng, X.; and Meng, L. 2024. Sequential selection and calibration of video frames for 3D outdoor scene reconstruction. *CAAI Transactions on Intelligence Technology*, 9(6): 1500–1514.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiao, J.; Zhang, Q.; Nie, Y.; Zhu, L.; and Zheng, W.-S. 2025a. RoGSplat: Learning Robust Generalizable Human Gaussian Splatting from Sparse Multi-View Images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5980–5990.

Xiao, T.; Wu, Y.; Li, Y.; Qin, C.; Gong, M.; Miao, Q.; and Ma, W. 2025b. Disentangled Pose and Appearance Guidance for Multi-Pose Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5646–5655.

Xu, Y.; Ye, K.; Shao, T.; and Weng, Y. 2025. Animatable 3D Gaussians for modeling dynamic humans. *Frontiers of Computer Science*, 19(9): 199704.

Yang, Y.; Huang, Y.; Wu, X.; Guo, Y.-C.; Zhang, S.-H.; Zhao, H.; He, T.; and Liu, X. 2024. Dreamcomposer: Controllable 3d object generation via multi-view conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8111–8120.

Yichao, Y.; Yuhao, C.; Zhuo, C.; Yicong, P.; Sijing, W.; Weitian, Z.; Junjie, L.; Yixuan, L.; Jingnan, G.; Weixia, Z.; et al. 2023. A survey on generative 3d digital humans based on neural networks: representation, rendering, and learning. *SCIENTIA SINICA Informationis*, 1858–2023.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. 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*, 5746–5756.

Yuan, Y.; Wu, Y.; Fan, X.; Gong, M.; Ma, W.; and Miao, Q. 2023. EGST: Enhanced geometric structure transformer for point cloud registration. *IEEE transactions on visualization and computer graphics*, 30(9): 6222–6234.

Yuan, Y.; Wu, Y.; Fan, X.; Gong, M.; Miao, Q.; and Ma, W. 2025. Where Precision Meets Efficiency: Transformation Diffusion Model for Point Cloud Registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9734–9742.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, T.; Liu, P.; Lu, Y.; Cai, M.; Zhang, Z.; Zhang, Z.; and Zhou, Q. 2025. CWNet: Causal wavelet network for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8789–8799.

Zhang, T.; Liu, P.; Zhao, M.; and Lv, H. 2024. DMFourLLIE: dual-stage and multi-branch fourier network for low-light image enhancement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7434–7443.

Zhang, Z.; Yang, Z.; and Yang, Y. 2024. 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*, 9936–9947.

Zhang, Z.-C.; Chen, H.; Deng, J.-S.; Xu, M.; and Pang, Z.-B. 2026. HexaDream: hexaview prior and constraint for text to 3D creation. *Frontiers of Computer Science*, 20(2): 1–14.

Zheng, S.; Zhou, B.; Shao, R.; Liu, B.; Zhang, S.; Nie, L.; and Liu, Y. 2024. 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*, 19680–19690.