# HiReFF: High-Resolution Feedforward Human Reconstruction from Uncalibrated Sparse-View Video

Yiming Jiang[1⋆], Hanzhang Tu[2], Wenfeng Song[3], Siyou Lin[2], Liang An[2], Shuai Li[1,4], Aimin Hao[1(✉)], and Yebin Liu[2]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
{jiangyimingjym, lishuai, ham}@buaa.edu.cn
[2] Tsinghua University, Beijing, China
{thz22, linsy21}@mails.tsinghua.edu.cn
{anliang, liuyebin}@mail.tsinghua.edu.cn
[3] College of Computer Science, Beijing Information Science and Technology University, Beijing, China
songwenfenga@gmail.com
[4] Zhongguancun Laboratory, Beijing, China

**Abstract.** Uncalibrated volumetric video streaming for human reconstruction is essential for holographic communication and AR/VR, yet remains challenging due to the need for temporal consistency and computational efficiency from sparse-view inputs. Existing methods rely on per-scene optimization or calibrated cameras, while recent feed-forward models are limited to low-resolution (0.5K) single-frame synthesis. We present HiReFF, a feed-forward method for 2K-resolution 360° human video reconstruction from uncalibrated sparse-view videos. Our framework decomposes the problem into two key tasks: foreground 3D Gaussian reconstruction from sparse-view videos (four views separated by 90°) and computationally efficient high-resolution synthesis. To enable the former, we propose Scale-synchronized Camera Calibration to resolve scale ambiguity for multi-view supervision, and Gaussian-wise Foreground Masking to reconstruct clean foregrounds by modulating Gaussian parameters. For efficient high-resolution synthesis, our High-resolution Side-tuning achieves 2K rendering by augmenting the Gaussian head with supplementary features while keeping the backbone at 0.5K, drastically reducing computational overhead. Experiments demonstrate that HiReFF significantly outperforms existing methods in high-resolution streaming volumetric video reconstruction. https://iridescentjiang.github.io/HiReFF

**Keywords:** 3D reconstruction · Feedforward 3D gaussian · Digital human

⋆ Y.Jiang—Work done during an internship at Tsinghua University.

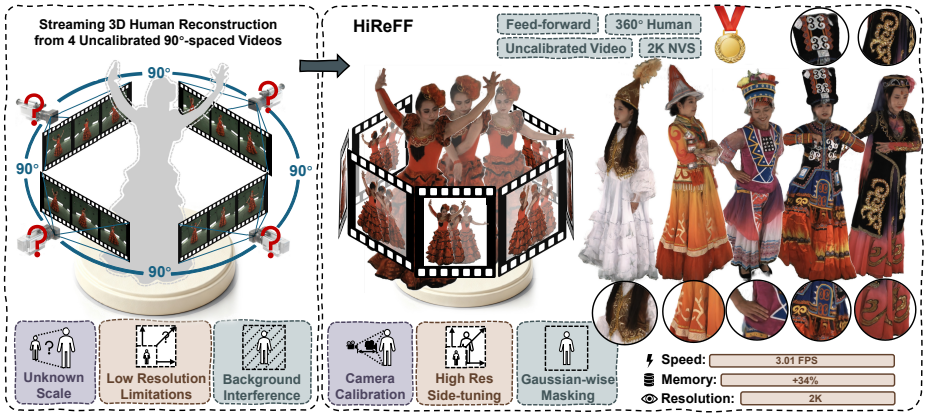


**Fig. 1:** We present HiReFF, a feed-forward method for 2K-resolution 360° human video reconstruction from uncalibrated sparse-view videos. With four-view uncalibrated videos as input, HiReFF reconstructs a 360° human in a streaming fashion at 3.01 FPS on a single RTX 4090 GPU and achieves 2K resolution with only 34% additional VRAM during training compared to 0.5K.

## 1 Introduction

Uncalibrated volumetric video streaming for human reconstruction holds diverse applications, including holographic communication [50], augmented/virtual reality (AR/VR) [23, 39, 45, 46, 48, 73], and sports broadcasting [2]. Despite recent breakthroughs in static human reconstruction [13, 15, 18, 19, 25, 33, 47, 49], uncalibrated volumetric video streaming remains exceptionally challenging, as existing methods are either reliant on per-scene iterative optimization over video sequences [1,26,58,64] or necessitate calibrated camera setups [20,40,54,75]. The advent of uncalibrated 3D feed-forward reconstruction models [6,22,42,52,55,78] has laid a foundation for addressing this task, demonstrating the ability to predict depth, point clouds, and camera parameters from sparse-view uncalibrated imagery. Concurrently, AnySplat [17] has extended these capabilities to achieve photorealistic novel view synthesis from such inputs. Nevertheless, these methods are primarily designed for single-frame inference, present challenges in sparse-view scenarios with wide baselines (e.g., 90° spacing), and are computationally constrained to $518 \times 518$ resolution outputs. To overcome these limitations, as shown in Fig. 1, we present HiReFF, a feed-forward method for high-resolution 360° human video reconstruction that enables streaming reconstruction of human models and photorealistic rendering from uncalibrated sparse-view videos.

Our approach enables **(1) the reconstruction of foreground human 3D Gaussians from uncalibrated, streaming sparse-view video inputs** (four views separated by 90°), and **(2) efficient high-resolution novel view synthesis without incurring substantial computational overhead.** To achieve the first objective, we leverage pretrained parameters from the feed-forward 3D reconstruction model VGGT [52] to obtain predicted camera pa-

rameters and depth priors, and augment it with Gaussian and mask prediction heads to reconstruct human 3D Gaussians. In doing so, we address two key challenges: 1) The scale mismatch hindering additional view supervision. Our setup involves four 90°-spaced input views, rendering supervision from only these views is insufficient. We therefore require additional novel viewpoint supervision, which necessitates ground-truth camera parameters. This presents a fundamental challenge: VGGT's predictions lack metric scale, preventing direct alignment with the ground-truth camera. To address this, we propose Scale-synchronized Camera Calibration (Sec. 3.2), which dynamically adjusts camera parameters for extra supervision views during training and employs indirect supervision of the camera head. 2) Foreground reconstruction compromises camera estimation. Directly masking input images for foreground-only reconstruction severely degrades camera parameter accuracy in widely-spaced (90°) sparse-view settings, creating a conflict between foreground masking and camera prediction. Therefore, we propose Gaussian-wise Foreground Masking (Sec. 3.2), which introduces a mask head to modulate Gaussian parameters. Approximately orthogonal input views may leave a small number of extraneous Gaussians after masking; however, our experiments show that they vanish during training, yielding a clean foreground reconstruction.

To achieve computationally efficient high-resolution reconstruction, we propose High-resolution Side-tuning (Sec. 3.3), inspired by side-tuning [70], which augments the Gaussian head with supplementary features to enable high-resolution output. Constrained by computational overhead, prior models [17] are limited to $518 \times 518$ novel-view rendering, which proves inadequate for practical high-resolution applications. Critically, however, high-resolution rendering does not necessitate higher precision in VGGT features, but rather demands enhanced appearance attributes. Consequently, we only increase the number of depth and Gaussian positions through interpolation and employ a supplementary architecture that injects additional image-derived features into the mid-level of the Gaussian head, thereby maintaining the backbone's input resolution at $518 \times 518$. During training, we render novel-view images at high resolution and supervise against high-resolution ground-truth imagery, while employing patch-wise perceptual loss to further mitigate computational burden. Our high-resolution rendering results, together with measurements of memory footprint and computation time, substantiate the effectiveness of this approach.

The contributions of this work are summarized as follows:

- We introduce HiReFF, a feed-forward framework for high-resolution 360° human video reconstruction, achieving 2K-resolution streaming 3D Gaussian reconstruction from four uncalibrated 90°-spaced views.
- We propose Scale-synchronized Camera Calibration to resolve metric scale ambiguity for effective multi-view supervision, and Gaussian-wise Foreground Masking to reconstruct clean foregrounds by modulating Gaussian parameters.
- We design High-resolution Side-tuning that achieves efficient 2K rendering by augmenting the Gaussian head with supplementary image features while

maintaining the backbone at 518×518 resolution, significantly reducing computational overhead.
– Extensive experiments on benchmark datasets demonstrate that our method significantly outperforms existing approaches in high-resolution streaming volumetric video reconstruction, both quantitatively and qualitatively.

## 2 Related Work

**Feed Forward Reconstruction.** Significant progress has been made in feed-forward 3D reconstruction from multi-view images. To achieve photorealistic novel view synthesis, a line of research [3,5,28,49,56,62,72,76] focuses on directly predicting static 3D Gaussians from calibrated multi-view images. Meanwhile, another line of methods, such as DUSt3R [53], FLARE [71], VGGT [52], and MapAnything [22] can now directly regress camera poses and 3D point clouds within the coordinate frame of the first image, while $\pi$3 [55] extends this capability to relative coordinate systems. Furthermore, to reduce reliance on camera calibration, methods such as NoPosplat [65] and AnySplat [17] aim to reconstruct 3D Gaussians from uncalibrated images. Recently, leveraging the capabilities of uncalibrated 3D reconstruction models, FastAvatar [59] and Human3R [4] have extended 3D Gaussian reconstruction from uncalibrated images to human reconstruction. However, the high computational cost of geometry transformers limits feed-forward 3D reconstruction networks to input resolutions around 0.5K, preventing direct high-resolution novel view synthesis. To address this, our method is designed for efficient high-resolution rendering. It can quickly reconstruct high-quality 3D Gaussian and render 2K images without a significant computational penalty.

**Sparse-view Human Reconstruction.** Reconstructing dynamic humans from sparse camera arrays is crucial for practical deployment, yet severely ill-posed due to limited geometric cues. While some approaches attempt reconstruction via parametric avatars [27, 37, 44, 57]. IDOL [77] and LHM [36] can reconstruct an avatar in several seconds using human-specific priors such as SMPL [30]. However, they encounter fundamental difficulties in modeling loose clothing and complex deformations. Other methods [35,74] rely on per-scene optimization, demanding calibrated multi-view rigs and prohibitive training times. To relax these constraints, feed-forward methods [20, 40, 41, 60, 75] learn generalizable priors from large-scale datasets, enabling faster reconstruction but still requiring precise camera calibration. Recently, Forge4D [14] first achieved frontal human novel-view synthesis from uncalibrated videos. Consequently, no existing method simultaneously addresses the core challenges of uncalibrated sparse-view inputs, streaming video reconstruction, and high-resolution 360° rendering—all critical for real-world applications in holographic communication and immersive broadcasting. HiReFF closes this gap by introducing a feed-forward framework that achieves 3D Gaussian reconstruction at 2K resolution without camera calibration.

**High-Resolution Novel View Synthesis.** High-resolution Novel View Synthesis (HRNVS) targets high-resolution novel view generation from low-

resolution (LR) inputs. Early NeRF-based methods [12,16,51] pioneered optimizing high-resolution neural radiance fields under sub-pixel constraints. Recently, 3D Gaussian Splatting (3DGS) capitalizes on its advantage of faster rendering speeds to produce high-quality imagery [8,9]. Subsequent render methods [68,69] have been proposed to address the rendering clarity of 3D Gaussian across resolutions. Super-resolution for 3D Gaussian Splatting has been addressed in several recent works. SRGS [11] achieves high-quality HRNVS using only existing low-resolution views, while GaussianSR [67] incorporates 2D diffusion priors for super-resolution. However, these methods are not readily compatible with modern feed-forward 3D reconstruction backbones. To bridge this gap, we introduce a sidetune adaptation strategy that enables fast and computationally efficient HRNVS within a feed-forward 3DGS framework.

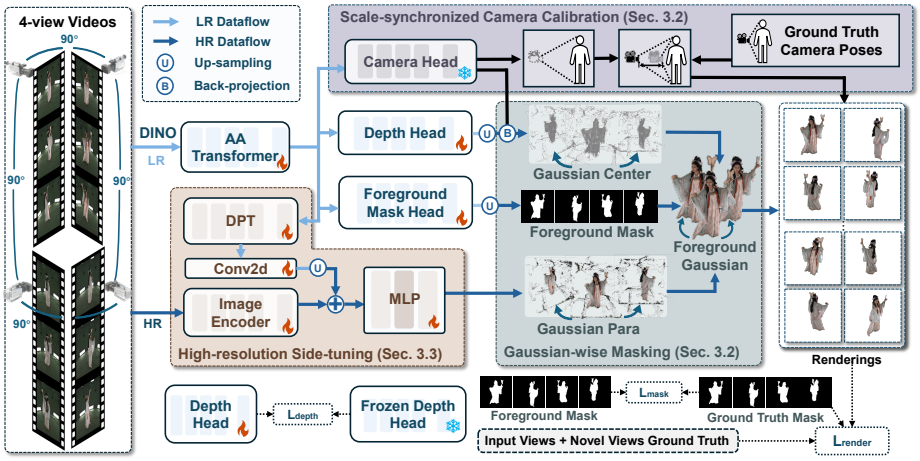


**Fig. 2: Method Overview** (§3). Taking four-view uncalibrated videos as input, we first extract features using an Alternating-Attention (AA) Transformer, then decode to obtain Gaussian parameters, supervising through rendered multi-view images. Specifically, HiReFF employs Scale-synchronized Camera Calibration (Sec. 3.2) to introduce supervision from additional viewpoints while indirectly supervising the Camera Head, uses Gaussian-wise Masking (Sec. 3.2) to remove background while preserving camera parameter accuracy, and leverages High-resolution Side-tuning (Sec. 3.3) to ensure computational efficiency at high resolution.

## 3 Method

As shown in Fig. 2, we present a method that achieves 2K-resolution feed-forward human volumetric video reconstruction: from just four uncalibrated videos captured 90° apart, our network rapidly reconstructs the 3D Gaussian capable of being rendered at 2K resolution from any novel 360° viewpoint.

The section is structured as follows: the problem is formulated in Sec 3.1, the overall architecture and pipeline are introduced in Sec 3.2, and our technique for computationally efficient high-resolution reconstruction is detailed in Sec 3.3.

### 3.1 Problem Setup

Consider four uncalibrated high-resolution (HR) RGB videos capturing a dynamic scene from viewpoints spaced 90° apart, denoted as $\{V_i\}_{i=1}^4$, where each video $V_i$ consists of frames $I_i^t \in \mathbb{R}^{H_{\text{HR}} \times W_{\text{HR}} \times 3}$. Our goal is to reconstruct a HR 3D Gaussian Splatting (3DGS) volumetric video by estimating, for each time step $t$: 1. A set of $G$ anisotropic 3D Gaussians $\{(\mu_g^t, \sigma_g^t, r_g^t, s_g^t, c_g^t)\}_{g=1}^G$, where $G = 4 \times H_{\text{HR}} \times W_{\text{HR}}$ corresponds to the total number of pixels across all four input HR views, parameterized by position $\mu \in \mathbb{R}^3$, opacity $\sigma \in \mathbb{R}^+$, rotation $r \in \mathbb{R}^4$, scale $s \in \mathbb{R}^3$, and spherical harmonic coefficients $c \in \mathbb{R}^{3\times(k+1)^2}$ of degree $k$; 2. Temporally smooth per-view camera parameters $\{p_i^t \in \mathbb{R}^9\}_{i=1}^4$ for each view, estimated with inter-frame continuity constraints to ensure stable trajectory reconstruction. Formally, our model performs the mapping from high-resolution multi-view videos to a high-resolution dynamic 3DGS representation: $f_\theta : \{V_i^{\text{HR}}\}_{i=1}^4 \longmapsto \left\{(\mu_g^t, \Theta_g^t)_{g=1}^G \cup (p_i^t)_{i=1}^4\right\}_{t=1}^T$, enabling high-resolution novel view synthesis at $H_{\text{HR}} \times W_{\text{HR}}$ resolution and arbitrary temporal intervals. $\Theta_g^t = (\sigma_g^t, r_g^t, s_g^t, c_g^t)$ represents the 3DGS parameters.

The model is evaluated on temporally consistent 3D reconstruction and high-resolution novel view synthesis at $H_{\text{HR}} \times W_{\text{HR}}$, and also produces auxiliary outputs such as per-frame depth maps and smooth camera trajectories usable in volumetric video applications.

### 3.2 Inter-frame Scale-stabilized Feedforward 3D Gaussian Reconstruction

Recent foundational models for uncalibrated feedforward 3D reconstruction [22, 52] employ an Alternating-Attention (AA) Transformer to extract features, which are subsequently decoded into basic 3D information (camera parameters, depth, point clouds), yet they fail to produce photorealistic imagery.

**Side-tuning Gaussian Prediction.** Recently, AnySplat [17] demonstrated the feasibility of feedforward photorealistic synthesis by utilizing a DPT [38] as a Gaussian head $F_G$ to decode AA Transformer features into 3D Gaussian Splatting (3DGS) parameters. Inspired by AnySplat, we introduce a side-tuning [70] approach where features extracted from images via an MLP $F_a$ are combined with intermediate $F_G$ features to predict per-frame 3DGS parameters for videos. Specifically, let $I^t \in \mathbb{R}^{H\times W\times 3}$ denote the video frame at time $t$. The 3DGS parameters $\Theta^t$ are obtained through:

$$\begin{aligned} f_A^t &= F_{AA}(I^t), \\ \Theta^t &= F_D\big(\, F_a(I^t) \oplus F_G^{\text{mid}}(f_A^t)\,\big), \end{aligned} \tag{1}$$

where $F_{AA}$ represents AA Transformer, $f_{AA}$ represents the feature encoded by $F_{AA}$, $F_G^{\text{mid}}(\cdot)$ represents intermediate features from the Gaussian head, $\oplus$ indicates adding operation, $F_D(\cdot)$ is another MLP generates the final 3DGS parameters including Gaussian attributes $\{\sigma, r, s, c\}$. Position $\mu$ is obtained via back-projection using depth predictions from the depth head and camera parameters from the camera head, thereby establishing per-pixel alignment with the input image. Specifically, for a pixel with homogeneous coordinates $\mathbf{p} = [u, v, 1]^T$, predicted depth $d$, camera intrinsics $\mathbf{K}$, and extrinsics $[\mathbf{R} \,|\, \mathbf{t}]$, the corresponding 3D point in world coordinates is: $\mu = \mathbf{R}^T \left( d\,\mathbf{K}^{-1}\mathbf{p} - \mathbf{t} \right)$, where $R$ is the $3 \times 3$ rotation matrix and $\mathbf{t}$ is the 3-dimensional translation vector of the camera pose.

The proposed sidetune module augments the Gaussian DPT head by providing crucial low-level, fine-grained details from the image, serving as a crucial complement to the high-level features encoded by the AA Transformer.

**Gaussian-wise Foreground Masking.** For human reconstruction tasks, the objective is to reconstruct only the foreground human region while masking out the surrounding background. However, with four input views spaced 90° apart exhibiting minimal overlap, using only the foreground human regions presents significant challenges for camera parameter estimation. We empirically observe that larger input regions yield more accurate camera predictions. To address this, we employ the full original images as input and introduce a mask head $F_m$ to selectively filter the predicted Gaussians, ensuring that only the foreground human region is reconstructed. Formally, this process is defined as:

$$(\Theta_g^t, \mu_g^t)_{g=1}^{\text{H}} = F_m((\Theta_g^t, \mu_g^t)_{g=1}^{G}) \odot (\Theta_g^t, \mu_g^t)_{g=1}^{G}, \tag{2}$$

where $\odot$ denotes Gaussian-wise masking based on the predicted foreground probabilities. $(\Theta_g^t, \mu_g^t)_{g=1}^{\text{H}}$ denotes the Gaussian parameters constituting the human subject to be reconstructed.

**Scale-Synchronized Camera Calibration.** Previous methods [17] for supervising the camera head typically employ direct supervision using ground-truth camera parameters, while simultaneously rendering 3D Gaussians based on the camera head's predictions and supervising these renderings with real images. However, we observe that for inputs with large viewpoint intervals (e.g., 90°), supervision from only the input views proves insufficient, necessitating the introduction of additional supervision viewpoints. Moreover, such large viewpoint intervals cause the camera head to exhibit significant fluctuations during training. These fluctuations substantially impact the final rendering outcome, causing a large variance in the rendering loss, which hinders convergence. To address this, we freeze the camera head during training and instead use the ground-truth camera parameters from both the input and supervision viewpoints to render the Gaussian parameters $\mu^t$ and $\Theta^t$, while activating the AA Transformer to modify the input features to the camera head, thereby indirectly optimizing it. This approach introduces novel viewpoints for supervision and avoids large fluctuations in the camera head, leading to improved convergence. Since the base model lacks a true scale—the initialized positions $\mu^t$ of the Gaussian points only align with the predicted camera parameters and do not possess a realistic scale (e.g., the

predicted human body might be only 0.8m tall)—it is infeasible to directly use the true camera parameters to render $\mu^t$. Therefore, we scale the true camera parameters by adjusting the translation $\hat{\mathbf{t}}$ to match the scale of the predicted translation $\mathbf{t}$. Furthermore, we observe that this indirect supervision approach maintains the stability of the camera head's predictions across frames, effectively mitigating scale fluctuation in inter-frame reconstruction results.

Specifically, for each non-reference view $i$ (where $i = 2, 3, \ldots, V$, with $V$ being the number of views, and $i = 1$ as the reference view), the scale adjustment is computed as:

$$\begin{aligned} \bar{s} &= \frac{1}{V-1} \sum_{i=2}^{V} \frac{\hat{\mathbf{t}}_i}{\mathbf{t}_i}, \\ \hat{\mathbf{t}}'_i &= \frac{\hat{\mathbf{t}}_i}{\bar{s}} \quad \text{for } i = 2, \ldots, V. \end{aligned} \tag{3}$$

Subsequently, we use the true intrinsic parameters $\hat{\mathbf{K}}^{\mathbf{t}}$, rotation matrices $\hat{\mathbf{R}}^{\mathbf{t}}$, and the adjusted translation $\hat{\mathbf{t}}^{\mathbf{t}'}$ to render the Gaussian parameters. While the position $\mu$ is back-projection using the predicted camera parameter $R^t$, $K^t$, and $t^t$. The rendering process can be formulated as:

$$\begin{aligned} \mu_g^t &= \mathbf{R}^{\mathbf{t}^T} \left( d^t \, \mathbf{K}^{\mathbf{t}^{-1}} \mathbf{p} - \mathbf{t}^{\mathbf{t}} \right), \\ I^t &= \mathcal{R}(\{\mu_g^t, \Theta_g^t\}_{g=1}^H, \hat{\mathbf{K}}^{\mathbf{t}}, \hat{\mathbf{R}}^{\mathbf{t}}, \hat{\mathbf{t}}^{\mathbf{t}'}), \end{aligned} \tag{4}$$

where $\mathcal{R}(\cdot)$ denotes the rendering function that projects the 3D Gaussians onto the image plane using the given camera parameters, producing the rendered image $I^t$. $I^t$ is then used for supervision with the ground truth image $\hat{I}^t$.

### 3.3 Computationally Efficient High-resolution Reconstruction

Volumetric video reconstruction demands high-resolution rendering. However, feeding high-resolution images (e.g., 2K) directly into the AA Transformer of feed-forward 3D foundation models incurs excessive computational cost—existing models [17, 52] typically operate at $518 \times 518$ resolution. Thus, we need to seek an approach that enhances reconstruction resolution while preserving the AA Transformer input resolution.

**High-resolution Side-tuning.** Importantly, high-resolution rendering quality depends primarily on the high-dimensional Gaussian attributes $\Theta_g$ output by the Gaussian head $F_G$, rather than on position $\mu_g$ or camera position $p_i$. Therefore, we maintain the input of AA Transformer at $518 \times 518$ while providing complementary high-resolution information to $F_G$ through a supplementary pathway. Specifically, we employ a side-tuning strategy: high-resolution images $I_{HR}$ are fed to the supplementary network $F_a$ while low-resolution images $I_{LR}$ are processed by the AA Transformer. Intermediate features $F_G^{\text{mid}}(I_{LR})$ are upsampled to high resolution before fusion with $F_a(I_{HR})$ and subsequent feeding

into $F_D$, formulated as:

$$\Theta^t = F_D\Big(F_a\big(I^t_{HR}\big) \oplus \mathcal{U}\big(F^{\text{mid}}_G(I^t_{LR})\big)\Big), \tag{5}$$

where $\mathcal{U}(\cdot)$ denotes upsampling to high resolution and $F_a$ is an MLP modified by EdgeNeXt [31]. Subsequently, the mask head $F_m$ and depth head $F_d$ outputs are upsampled to high resolution, and rendering is performed at high resolution for supervision against ground-truth high-resolution images.

**Training Loss.** We employ a rendering loss to supervise image synthesis. During training, in addition to the four input views, we sample $V_a$ additional novel viewpoints from frontal, top, and bottom positions outside the input viewing sphere, yielding $4+V_a$ supervision viewpoints in total. We render the network outputs of all $4+V_a$ views and supervise them against ground-truth images. The rendering loss combines L1 loss and perceptual loss, formulated as:

$$\mathcal{L}_{\text{render}} = \sum_{i=1}^{4+V_a} \|\hat{I}_i - I_i\|_1 + \lambda_{\text{P}} \sum_{i=1}^{4+V_a} \text{Perceptual}(\hat{I}_i, I_i), \tag{6}$$

where $\hat{I}_i$ and $I_i$ denote the rendered and ground-truth images for view $i$, respectively. The perceptual loss [21] measures the distance in the VGG [43]'s feature space, as Animatable Gaussians [27]. We use computation graph surgery [32] to lower the memory consumption of perceptual loss.

For mask supervision, we employ L1 loss between the predicted and ground-truth masks in 4 input views:

$$\mathcal{L}_{\text{mask}} = \sum_{i=1}^{4} \|\hat{M}_i - M_i\|_1, \tag{7}$$

where $\hat{M}_i$ and $M_i$ are the predicted and ground-truth masks for view $i$.

To mitigate overfitting of the depth head to input viewpoints, we utilize a depth distillation loss. Specifically, we maintain a frozen reference depth head initialized with VGGT [52] parameters and enforce consistency with our active depth head via MSE loss:

$$\mathcal{L}_{\text{depth}} = \sum_{i=1}^{4} \|\hat{D}^{\text{a}}_i - \hat{D}^{\text{f}}_i\|^2_2, \tag{8}$$

where $\hat{D}^{\text{a}}_i$ and $\hat{D}^{\text{f}}_i$ denote depth predictions from the active and frozen heads for view $i$, respectively. The overall training objective combines these losses with weighted coefficients:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{render}}\mathcal{L}_{\text{render}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}. \tag{9}$$

# 4 Experiment

## 4.1 Experimental Settings

**Datasets.** We conduct training mainly on the DNA-Rendering [7] dataset, which encompasses 153 actors and 439 distinct motion sequences. Each sequence comprises 48-view synchronized RGB video streams at $2448 \times 2048$ resolution, together with corresponding viewpoint camera parameters. To improve generalization, we also utilize the ZJU-MoCap [10,34] and MVHumanNet [24,61] datasets. For validation, we select 20 motion sequences with distinct identities from within the DNA-Rendering dataset as the test set. Detailed dataset configurations and additional results on more datasets are provided in the supplementary materials.

**Evaluation Metrics.** We assess novel-view rendering quality using PSNR, SSIM, and LPIPS metrics at $2072 \times 2072$ resolution. For models rendering at alternative resolutions (e.g., GPS-Gaussian [75], AnySplat [17]), we bilinearly upsample their outputs to 2K. All methods are evaluated under a 4-view input setting with approximately 90° angular separation, providing comprehensive 360° coverage of the human subject. For models requiring camera parameters as input, we supply ground-truth parameters, with this condition clearly indicated in our experimental results.

**Baselines.** We compare our approach against existing novel view synthesis methods. Specifically, for uncalibrated camera methods, we benchmark against AnySplat [17] and NoPoSplat [65]. We also include comparisons with 4DGT [63], a monocular volumetric video reconstruction method. For calibrated camera approaches, we evaluate against GPS-Gaussian [75]. Since our task involves foreground human segmentation and the comparison methods lack built-in foreground masks, we employ an open-source video portrait segmentation algorithm [29] to generate masks. GPS-Gaussian receives segmented foreground images as input, whereas other methods (including ours) take full images with background intact. To ensure fair comparison, we multiply all methods' predicted results by a consistent mask before computing evaluation metrics. Please refer to the supplementary materials for experimental setup details.

**Implementation Details.** The AA Transformer, camera head, and both active and frozen depth heads are initialized with pretrained weights from VGGT [52], whereas the 3D Gaussian prediction head, side-tuning supplementary network, and mask head are zero-initialized. Our Gaussian splatting renderer is based on gsplat [66]. We define HR resolution as $2072 \times 2072$ and LR resolution as $518 \times 518$. The network receives four HR images from viewpoints separated by 90°. For supervision, we render at HR resolution ($2072 \times 2072$) and provide ground-truth images at the same scale. We set $V_a = 4$, supervising across $4 + V_a = 8$ total viewpoints. Loss weights are configured as $\lambda_{\mathrm{P}} = 0.1$, $\lambda_{\mathrm{render}} = 1.0$, $\lambda_{\mathrm{mask}} = 5 \times 10^{-2}$, and $\lambda_{\mathrm{depth}} = 10^{1}$. All training stages are performed on 8 A800 GPUs. Automatic mixed precision is employed to accelerate training.


**Fig. 3: Qualitative results of novel-view synthesis of reconstruction** (§4.2). Our method surpasses all others in global shape, garment details, and facial fidelity.

### 4.2 Evaluation

**Qualitative Evaluation.** As shown in Fig. 3, our method surpasses all others in global shape, garment details, and facial fidelity. AnySplat [17] and NoPoSplat [65] struggle to estimate accurate camera parameters under the large 90° baseline, causing misaligned reconstructions across views. GPS-Gaussian [75], supplied with eight calibrated views, ground-truth cameras, and pre-segmented foreground images, yields a comparatively complete model, yet still exhibits leg-artifacts. 4DGT [63] fails to produce plausible results when only a single-frame video is provided. As shown in Fig. 4, we present test results from additional novel viewpoints. For both top-down and bottom-up perspectives, our method successfully reconstructs the correct geometry and accurately reproduces surface coloration. In the bottom two rows showcasing complex clothing, our method correctly reconstructs garment patterns. As shown in Fig. 5, our method maintains excellent temporal consistency for both human body and clothing details when reconstructing garments with complex patterns. Additional dynamic results are available in our supplementary video.

**Quantitative Evaluation.** As demonstrated in the Table. 1, our method outperforms all uncalibrated camera input approaches across all metrics. For NoPoSplat [65] and AnySplat [17], we optimized their camera parameters for 200 steps to better evaluate reconstruction quality, whereas our method achieves alignment with ground-truth cameras without any optimization. For AnySplat, the most directly comparable method, our approach improves PSNR by 2.7294, SSIM by 0.0124, and reduces LPIPS by 0.0460 compared to its unoptimized ver-


**Fig. 4: More visualization results.** For both top-down and bottom-up perspectives, our method successfully reconstructs the correct geometry and accurately reproduces surface coloration. Please **zoom in** to see details.

**Table 1: Quantitative results of novel-view synthesis of reconstruction (§4.2).** Ours outperforms all uncalibrated camera input methods across all metrics.

| Method | Cam. Pose | Views Num. | Aligned | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|---|---|
| GPS-Gaussian | ✓ | 8 | - | 26.1039 | 0.9172 | 0.1384 |
| 4DGT | × | 1 | × | 17.1689 | 0.8395 | 0.2719 |
| NoPoSplat | × | 2 | × | 22.6296 | 0.8876 | 0.1736 |
| | × | 2 | ✓ | 23.4321 | 0.8939 | 0.1588 |
| AnySplat | × | 4 | × | 23.7844 | 0.9040 | 0.1737 |
| | × | 4 | ✓ | 25.5875 | 0.9140 | 0.1598 |
| Ours | × | 4 | × | **26.5138** | **0.9164** | **0.1277** |

sion, while still maintaining clear advantages over its optimized variant. For GPS-Gaussian [75], despite achieving comparable SSIM when provided with ground-

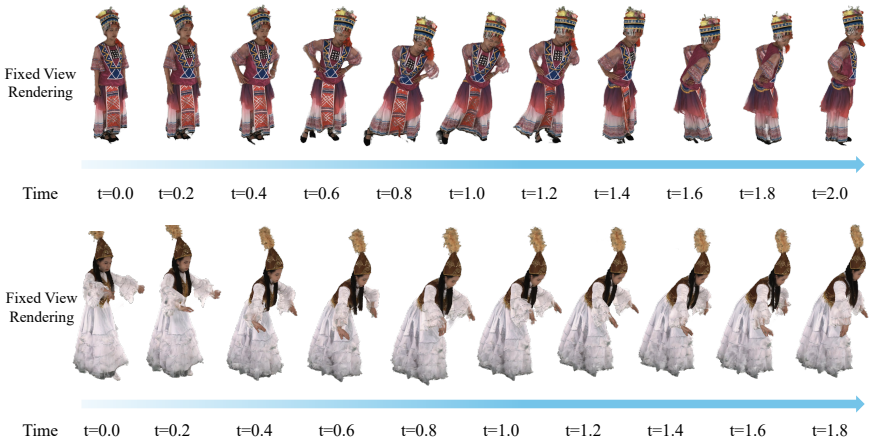


**Fig. 5: Qualitative results on timing consistency.** Our method maintains excellent temporal consistency for both human body and clothing details when reconstructing garments with complex patterns. Please **zoom in** to see details.

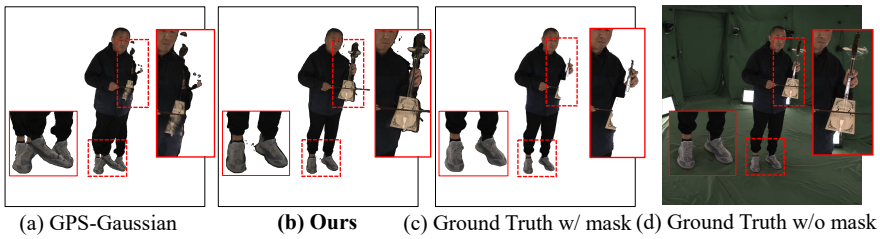


**Fig. 6:** Pre-trained prediction masks [29] occasionally exhibit imperfections, introducing challenges in metric evaluation. (§4.2). Please **zoom in** to see details.

truth cameras and eight input views, it cannot reconstruct from as few views as our method. Moreover, we observe that GPS-Gaussian's metric advantage stems from its use of masked input images, making it favorable when metrics are computed against masked ground truth. As illustrated in Fig. 6, our method inputs full images with background (d) and thus reconstructs the instrument region, while GPS-Gaussian's masked input (c) omits it. However, since all results are masked before evaluation against (c), GPS-Gaussian's artifact (extra feet in (a)) is suppressed, and our complete instrument reconstruction does not confer metric benefit.

### 4.3 Ablation Study

**Ablation on Computational Consumption.** As shown in the Table. 2a, we report input resolution, supervision resolution, and VRAM consumption across

**Table 2: Overall computational consumption**(§4.3).

**(a)** Training consumption

| In. Res. | Sup. Res. | Side-tuning | VRAM (MiB) |
|---|---|---|---|
| 518 | 518 | ✓ | 40503 |
| 518 | 2072 | ✓ | 44125 |
| 2072 | 2072 | ✓ | 59095 |
| 2072 | 2072 | × | OOM |

**(b)** Inference consumption

| In. Res. | Ren. Res. | VRAM (MiB) | Frame Rate (RTX 4090) |
|---|---|---|---|
| 518 | 2072 | 10852 | 4.40 FPS |
| 1036 | 2072 | 10886 | 4.02 FPS |
| 2072 | 2072 | 14052 | 3.01 FPS |

training stages. In the third row, with 2K input and 2K supervision, High-resolution Side-tuning increases VRAM usage by only 33.9% compared to 0.5K input and supervision. Without side-tuning, directly increasing the AA Transformer input resolution exceeds 80GB VRAM, resulting in out-of-memory errors. This validates the computational efficiency of our side-tuning approach. During inference, as shown in the Table. 2b, we demonstrate the inference computational consumption using a single RTX 4090 GPU. Our method achieves 3.01 FPS when rendering at 2K resolution, only 24% lower than the 4.40 FPS achieved at 0.5K. This performance reduction primarily stems from the increased resolution of the Gaussian renderer. Notably, when both configurations render at 2K, using 1K image input reduces speed by merely 8.6% compared to 0.5K input.

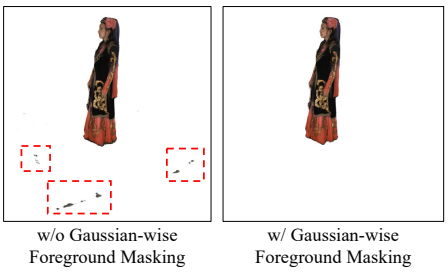


**Fig. 7: Ablation on Gaussian-wise Foreground Masking.** Our method effectively removes the background.

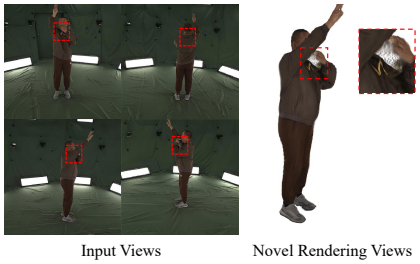


**Fig. 8: Limitation.** Our method exhibits isolated points in regions occluded from multiple viewpoints.

**Ablation on Gaussian-wise Foreground Masking.** As shown in Fig. 7, our masking method effectively removes the background. The minimal redundant elements in w/o Masking images demonstrate that our mask has trained the Gaussian Decoder to avoid reconstructing regions beyond the mask boundary. With Gaussian-wise Foreground Masking, the decoder focuses exclusively on

the foreground human subject rather than background regions during training, thereby enhancing reconstruction quality.

## 5 Limitation and Future Work

As shown in Fig. 8, our method exhibits isolated points in regions occluded from multiple viewpoints. Among the four input views, only the frontal view captures the region circled in red; consequently, insufficient side-view information leads to a few discrete points appearing in this area. In the future, we plan to incorporate human priors for hands and faces to enhance geometric and topological accuracy in these regions.

## 6 Conclusion

We present HiReFF, a feed-forward framework achieving 2K-resolution 360° human video reconstruction from uncalibrated sparse-view inputs. By decomposing the problem into temporally consistent 3D Gaussian reconstruction and efficient high-resolution synthesis, we bridge the gap between feed-forward efficiency and photorealistic volumetric video. Our Scale-synchronized Camera Calibration and Gaussian-wise Foreground Masking resolve scale ambiguity and foreground reconstruction, while the High-resolution Side-tuning enables 2K rendering with only modest computational overhead. Extensive experiments demonstrate state-of-the-art performance, unlocking practical applications in holographic communication, AR/VR, and live sports broadcasting without requiring calibrated rigs or per-scene optimization.

## Acknowledgements

Supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0124000). This paper is also supported by National Natural Science Foundation of China (62125107, 62572062, 62525204).

## References

1. Bahmani, S., Shen, T., Ren, J., Huang, J., Jiang, Y., Turki, H., Tagliasacchi, A., Lindell, D.B., Gojcic, Z., Fidler, S., Ling, H., Gao, J., Ren, X.: Lyra: Generative 3d scene reconstruction via self-distillation with video diffusion models. arXiv preprint arXiv:2509.19296 (2025)
2. Baumgartner, T., Klatt, S.: Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5109–5118 (2023)
3. Chen, J., Li, C., Zhang, J., Zhu, L., Huang, B., Chen, H., Lee, G.H.: Generalizable human gaussians from single-view image. In: The Thirteenth International Conference on Learning Representations, ICLR (2025)

4. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Gerard, P.M.: Human3r: Everyone everywhere all at once. arXiv preprint arXiv:2510.06219 (2025)
5. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference. vol. 15079, pp. 370–386 (2024)
6. Chen, Z., Liu, T., Zhuo, L., Ren, J., Tao, Z., Zhu, H., Hong, F., Pan, L., Liu, Z.: 4dnex: Feed-forward 4d generative modeling made easy. arXiv preprint arXiv:2508.13154 (2025)
7. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., Ren, D., Yang, L., Liu, Z., Loy, C.C., Qian, C., Wu, W., Lin, D., Dai, B., Lin, K.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 19925–19936 (2023)
8. Deng, T., Chen, X., Chen, Y., Chen, Q., Xu, Y., Yang, L., Xu, L., Zhang, Y., Zhang, B., Huang, W., Wang, H.: Gaussiandwm: 3d gaussian driving world model for unified scene understanding and multi-modal generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10656–10667 (June 2026)
9. Deng, T., Chen, Y., Yang, J., Yuan, S., Liu, J., Wang, D., Chen, W.: Cgs-slam: Compact 3d gaussian splatting for dense visual slam. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1606–1613 (2025)
10. Fang, Q., Shuai, Q., Dong, J., Bao, H., Zhou, X.: Reconstructing 3d human pose by watching humans in the mirror. In: CVPR (2021)
11. Feng, X., He, Y., Wang, Y., Yang, Y., Kuang, Z., Yu, J., Fan, J., Ding, J.: SRGS: super-resolution 3d gaussian splatting. CoRR **abs/2404.10318** (2024)
12. Han, Y., Yu, T., Yu, X., Xu, D., Zheng, B., Dai, Z., Yang, C., Wang, Y., Dai, Q.: Super-nerf: View-consistent detail generation for nerf super-resolution. IEEE Trans. Vis. Comput. Graph. **31**(9), 6053–6066 (2025)
13. He, X., Wu, Z., Li, X., Kang, D., Zhang, C., Ye, J., Chen, L., Gao, X., Zhang, H., Zhuang, H.: Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3437–3445 (2025)
14. Hu, Y., He, Y., Chen, J., Yuan, W., Qiu, K., Lin, Z., Zhu, S., Dong, Z., Zhang, J.: Forge4d: Feed-forward 4d human reconstruction and interpolation from uncalibrated sparse-view videos. CoRR **abs/2509.24209** (2025)
15. Hu, Y., Liu, Z., Shao, J., Lin, Z., Zhang, J.: Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse multi-view camera settings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2613–2622 (2025)
16. Huang, X., Li, W., Hu, J., Chen, H., Wang, Y.: Refsr-nerf: Towards high fidelity and super resolution view synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8244–8253 (2023)
17. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
18. Jiang, Y., Song, W., Li, S., Hao, A.: Decon: Reconstruction of clothed-geometric multiple humans from a single image via geometry-guided decoupling. Proceedings of the AAAI Conference on Artificial Intelligence **40**(7), 5459–5467 (2026)

19. Jiang, Y., Song, W., Li, S., Hao, A.: Hfhuman: High-fidelity human reconstruction from single image with multi-modality fusion. IEEE Transactions on Visualization and Computer Graphics **32**(2), 2152–2164 (2026)
20. Jin, Y., Peng, S., Wang, X., Xie, T., Xu, Z., Yang, Y., Shen, Y., Bao, H., Zhou, X.: Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11047–11057 (2025)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference. vol. 9906, pp. 694–711 (2016)
22. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In: 2026 International Conference on 3D Vision (3DV). pp. 499–509. IEEE (2026)
23. Kirschstein, T., Romero, J., Sevastopolsky, A., Nießner, M., Saito, S.: Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12089–12100 (2025)
24. Li, C., Liao, H., Zhi, Y., Yang, X., Sun, Z., Chang, J., Cui, S., Han, X.: Mvhumannet++: A large-scale dataset of multi-view daily dressing human captures with richer annotations for 3d human digitization. arXiv preprint arXiv:2505.01838 (2025)
25. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y., Xue, W., Luo, W., Guo, Y.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16008–16018 (2025)
26. Li, X., Wang, T., Gu, Z., Zhang, S., Guo, C., Cao, L.: Flashworld: High-quality 3d scene generation within seconds (2025)
27. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19711–19722 (2024)
28. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
29. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 238–247 (2022)
30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (2015)
31. Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S.W., Anwer, R.M., Khan, F.S.: Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In: International Workshop on Computational Aspects of Deep Learning at 17th European Conference on Computer Vision (CADL2022). Springer (2022)
32. Mescheder, L., Dong, W., Li, S., Bai, X., Santos, M., Hu, P., Lecouat, B., Zhen, M., Delaunoy, A., Fang, T., et al.: Sharp monocular view synthesis in less than a second. arXiv preprint arXiv:2512.10685 (2025)
33. Pan, P., Su, Z., Lin, C., Fan, Z., Zhang, Y., Li, Z., Shen, T., Mu, Y., Liu, Y.: Humansplat: Generalizable single-image human gaussian splatting with structure priors. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak,

J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems (2024)
34. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
35. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5020–5030 (2024)
36. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14184–14194 (2025)
37. Qiu, L., Zhu, S., Zuo, Q., Gu, X., Dong, Y., Zhang, J., Xu, C., Li, Z., Yuan, W., Bo, L., Chen, G., Dong, Z.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21148–21158 (2025)
38. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1623–1637 (2022)
39. Shao, R., Pang, Y., Zheng, Z., Sun, J., Liu, Y.: Human4dit: 360-degree human video generation with 4d diffusion transformer. ACM Transactions on Graphics (TOG) **43**(6) (2024)
40. Shao, R., Zhang, H., Zhang, H., Chen, M., Cao, Y., Yu, T., Liu, Y.: Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15851–15861 (2022)
41. Shao, R., Zheng, Z., Zhang, H., Sun, J., Liu, Y.: Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference. vol. 13692, pp. 702–720 (2022)
42. Shen, Y., Zhang, Z., Qu, Y., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer (2025)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, Conference Track Proceedings (2015)
44. Song, W., Ding, Y., Hou, F., Li, S., Hao, A., Hou, X.: Ctrlavatar: Controllable avatars generation via disentangled invertible networks. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence. pp. 6959–6967 (2025)
45. Song, W., Wang, X., Jiang, Y., Li, S., Hao, A., Hou, X., Qin, H.: Expressive 3d facial animation generation based on local-to-global latent diffusion. IEEE Trans. Vis. Comput. Graph. **30**(11), 7397–7407 (2024)
46. Song, W., Ye, Z., Wu, Z., Li, S., Hou, X., Hao, A.: Dynavatar: Dynamic 3d head avatar deformation with expression guided gaussian splatting. IEEE Transactions on Visualization and Computer Graphics **32**(3), 2454–2466 (2026)
47. Sun, J., Luo, F., Fan, W., Jiang, Y., Xiao, C.: Humanpro: Single-view 3d clothed human reconstruction with progressive normal guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9180–9188 (2026)
48. Tian, H., Liu, R., Shen, W., Hu, Y., Zheng, Z., Qin, X.: Efficienthuman: Efficient training and reconstruction of moving human using articulated 2d gaussian. In:

2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2025)
49. Tu, H., Liao, Z., Zhou, B., Zheng, S., Zhou, X., Zhang, L., Wang, Q., Liu, Y.: Gbc-splat: Generalizable gaussian-based clothed human digitalization under sparse RGB cameras. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26377–26387 (2025)
50. Tu, H., Shao, R., Dong, X., Zheng, S., Zhang, H., Chen, L., Wang, M., Li, W., Ma, S., Zhang, S., Zhou, B., Liu, Y.: Tele-aloha: A telepresence system with low-budget and high-authenticity using sparse RGB cameras. In: Burbano, A., Zorin, D., Jarosz, W. (eds.) ACM SIGGRAPH 2024 Conference Papers. p. 116 (2024)
51. Wang, C., Wu, X., Guo, Y., Zhang, S., Tai, Y., Hu, S.: Nerf-sr: High quality neural radiance fields using supersampling. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022. pp. 6445–6454 (2022)
52. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: VGGT: visual geometry grounded transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5294–5306 (2025)
53. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
54. Wang, W., Chen, Y., Zhang, Z., Liu, H., Wang, H., Feng, Z., Qin, W., Zhu, Z., Chen, D.Y., Zhuang, B.: Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. arXiv preprint arXiv:2509.19297 (2025)
55. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: $\pi^3$: Scalable permutation-equivariant visual geometry learning. CoRR **abs/2507.13347** (2025)
56. Wang, Y., Huang, T., Chen, H., Lee, G.H.: Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems (2024)
57. Weng, C., Curless, B., Kemelmacher-Shlizerman, I.: Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. CoRR **abs/2012.12884** (2020)
58. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: CAT4D: create anything in 4d with multi-view video diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26057–26068 (2025)
59. Wu, Y., Chen, X., Wu, Y., Li, W., Lu, Y., Feng, K.: Fastavatar: Towards unified and fast 3d avatar reconstruction with large gaussian reconstruction transformers. In: The Fourteenth International Conference on Learning Representations (2026)
60. Xiao, J., Zhang, Q., Nie, Y., Zhu, L., Zheng, W.S.: Rogsplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5980–5990 (2025)
61. Xiong, Z., Li, C., Liu, K., Liao, H., Hu, J., Zhu, J., Ning, S., Qiu, L., Wang, C., Wang, S., et al.: Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19801–19811 (2024)
62. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16453–16463 (2025)

63. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
64. Xu, Z., Xu, Y., Yu, Z., Peng, S., Sun, J., Bao, H., Zhou, X.: Representing long volumetric video with temporal gaussian hierarchy. ACM Trans. Graph. **43**(6), 171:1–171:18 (2024)
65. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In: The Thirteenth International Conference on Learning Representations, ICLR (2025)
66. Ye, V., Li, R., Kerr, J., Turkulainen, M., Yi, B., Pan, Z., Seiskari, O., Ye, J., Hu, J., Tancik, M., Kanazawa, A.: gsplat: An open-source library for gaussian splatting. Journal of Machine Learning Research **26**(34), 1–17 (2025)
67. Yu, X., Zhu, H., He, T., Chen, Z.: Gaussiansr: 3d gaussian super-resolution with 2d diffusion priors. CoRR **abs/2406.10111** (2024)
68. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19447–19456 (2024)
69. Zeng, H., Bai, Y., Fu, Y.: Arbitrary-scale 3d gaussian super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12304–12312 (2026)
70. Zhang, J.O., Sax, A., Zamir, A., Guibas, L.J., Malik, J.: Side-tuning: A baseline for network adaptation via additive side networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference. vol. 12348, pp. 698–714 (2020)
71. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: FLARE: feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21936–21947 (2025)
72. Zhang, S., Fei, X., Liu, F., Song, H., Duan, Y.: Gaussian graph network: Learning efficient and generalizable gaussian representations from multi-view images. Advances in Neural Information Processing Systems **37**, 50361–50380 (2024)
73. Zhang, Z., Yang, Z., Yang, Y.: SIFU: side-view conditioned implicit function for real-world usable clothed human reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9936–9947 (2024)
74. Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., Xu, L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7743–7753 (June 2022)
75. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19680–19690 (2024)
76. Zhou, B., Zheng, S., Liao, Z., Ma, Z., Tu, H., Liu, B., Liu, Y.: Splat-sap: Feed-forward gaussian splatting for human-centered scene with scale-aware point map reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
77. Zhuang, Y., Lv, J., Wen, H., Shuai, Q., Zeng, A., Zhu, H., Chen, S., Yang, Y., Cao, X., Liu, W.: Idol: Instant photorealistic 3d human creation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26308–26319 (2025)

78. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)