PoseGaussian: Pose-Driven Novel View Synthesis for Human Representation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we present PoseGaussian, a pose-guided Gaussian Splatting framework for high-fidelity human novel view synthesis. Body pose serves a dual role: first, as a structural prior fused with a color encoder to refine depth estimation; second, as a temporal cue in a dedicated pose encoder to enhance the consistency across frames. Both pathways are unified in a fully differentiable pipeline for end-to-end training. Our design specifically targets the challenges of handling articulated body motion and severe self-occlusion, common in dynamic human scenes. In terms of runtime efficiency, PoseGaussian achieves real-time rendering at 100 FPS, matching typical Gaussian Splatting pipelines while offering superior robustness to human motion. Extensive experiments across diverse datasets—including ZJU-MoCap, THuman2.0, and a custom real-world capture—demonstrate the effectiveness of our method, achieving state-of-the-art performance with PSNR of 30.86, SSIM of 0.979 for structural consistency, and LPIPS as low as 0.028, reflecting strong perceptual quality.

1 Introduction

Novel View Synthesis (NVS) is a core problem in computer vision and graphics and has evolved substantially—from traditional image-based rendering (IBR) techniques [64, 103, 16] to modern neural representations, such as Neural Radiance Fields (NeRF) [54] and its various extensions [68, 81, 95], which have achieved remarkable results in synthesizing realistic views. This progress has also fueled interest in free-viewpoint human rendering, a key sub-area of NVS that is increasingly driving innovation in immersive technologies, powering applications such as telepresence and embodied interaction. However, synthesizing high-quality human views remains a challenging task due to dynamic motion, identity-specific appearance, real-time performance demands, and the inherent ambiguity of reconstructing 3D geometry from sparse observations. Recent methods have extended NeRF-based frameworks to model human appearance and motion [62, 61, 78]. While demonstrating the ability to extend NeRF to model dynamic and person-specific variations, the implicit volumetric rendering paradigm, which relies on dense spatial queries, poses a bottleneck for real-time operation.

Recent advances in 3D Gaussian Splatting (3D-GS) have demonstrated impressive performance in novel view synthesis by representing scenes with explicit 3D Gaussians [38]. Unlike NeRF, which relies on implicit volumetric queries over a continuous field, 3D-GS employs an explicit function representation that enables direct projection of Gaussian primitives. This shift significantly improves rendering efficiency while maintaining high visual fidelity—making real-time synthesis practical. The method has shown promise across various domains, including free-viewpoint human rendering [102, 25, 39]. These methods have broadened the applicability of 3D-GS to dynamic human rendering, yet how to effectively integrate structured human priors into the pipeline for joint Gaussian estimation remains an open question. For instance, while GPS-Gaussian [102] demonstrates real-time human

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

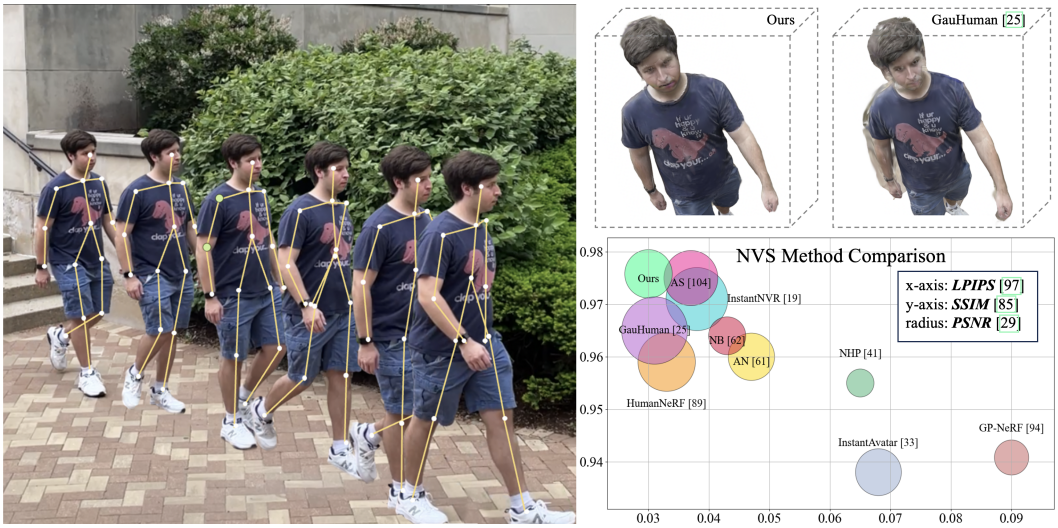

Figure 1: **PoseGaussian Visualization and Comparison.** Left: reconstructed pose-guided motion sequence reprojected into the original scene. Top right: visual comparison with GauHuman [25]. Bottom-right: performance chart highlighting selected methods.

view synthesis, it does not leverage explicit human semantics—such as pose or body structure—in the prediction of Gaussian parameters. GauHuman [25] achieves efficient training and real-time rendering by leveraging SMPL fits to guide Gaussian generation. However, its strong reliance on precise SMPL alignment hinders generalization when pose estimates are noisy or body shapes fall outside the training distribution. Additionally, the method struggles with fast motions, where reconstructed results often exhibit motion blur and background leakage. As shown in Fig. 1, walking sequences reveal unclear boundaries and ghosting artifacts, indicating challenges in preserving temporal and structural consistency under dynamic conditions.

To advance human-centric novel view synthesis, we observe that dynamic human motion introduces unique challenges not typically present in general scene modeling—such as frequent self-occlusion, non-rigid articulation, and rapid temporal changes. Unlike static objects, humans exhibit structured, temporally coherent behaviors naturally described by their skeletal pose [51]. This structured motion prior provides a powerful cue for guiding scene representation, especially under sparse or ambiguous observations [35]. By treating pose as a central organizing signal, we can build representations that are semantically grounded and temporally stable, enabling more accurate and robust synthesis in dynamic settings [1]. As shown in Fig. 1, we visualize a few representative frames with reconstructed 3D poses reprojected into the original scene, illustrating the core idea of our approach and its strong performance relative to state-of-the-art methods. Our main contributions are as follows (code and data are available anonymously at `https://anonymous.4open.science/r/PoseGaussian/`):

- **Pose-aware initialization scheme** that leverages human body priors to guide the estimation of Gaussian primitives, enabling semantically informed and structurally coherent representations.
- **Temporal regularization strategy** that promotes inter-frame consistency while preserving fine-grained details and accommodating natural motion variations.
- **End-to-end differentiable framework** that integrates pose-informed feature encoding and motion-consistent regularization into a unified architecture for dynamic human reconstruction, providing a principled and extensible foundation for future research and applications.

2 Related Works

Human view synthesis has progressed from early 3D reconstruction methods using structured light and laser scanning systems [80, 12, 21], to model-driven approaches like SCAPE [4], SMPL [51], and STAR [56], which represent human bodies via mesh-based and articulated models. While influential, these approaches often lacked photorealism and required strong priors or manual tuning.

More recently, learning-based methods such as Neural Radiance Fields (NeRF) [54] have enabled photorealistic novel view synthesis directly from images. Building on this, 3D-GS [38] introduces an explicit, point-based representation using 3D Gaussians to improve rendering efficiency. Below, we focus on this modern paradigm and its potential for data-driven, photorealistic reconstruction.

Radiance Fields and Human Representation: Building on the core NeRF architecture [54] and its extensions [13, 53], recent research has focused on adapting NeRF to more complex geometries and constrained scenarios, such as static human subjects in fixed poses with minimal motion [59, 60, 44]. These approaches allow NeRF to render humans without explicit motion modeling. Several works integrate human representations like SMPL [51] or 3D skeletons [95, 48, 62, 61], enabling pose-conditioned synthesis. Models like Neural Body [62] and Neural Actor [48] leverage sparse convolutions and mesh-based warping, respectively, but often require per-subject optimization and struggle with generalization to unseen poses. Other approaches integrating human features, such as HumanNeRF [89] and Vid2Actor [88], focus on single-video human rendering. These methods also face challenges in capturing fine details without dense supervision, especially in dynamic scenarios where errors in pose estimation or mesh warping can lead to distortions like flickering or jittering [48, 61]. In general, NeRF-style models based on implicit volumetric representations are computationally intensive and memory-consuming, often requiring long training and rendering times to produce high-resolution outputs. Although various acceleration techniques have been proposed to address these issues [17, 22], the implicit nature of these models continues to limit scalability and hinders real-time or interactive applications.

Gaussian Splatting for Human Novel View Synthesis: 3D-GS [38] has recently emerged as a highly active and rapidly advancing direction for human-centric representations, with a surge of representative works drawing broad interest in the field [39, 55, 58, 25, 105, 43, 69, 75]. One emerging branch within this trend explores sparse-camera setups for improving efficiency and training speed [102, 14]. These approaches are promising for real-time applications but may face challenges in capturing fine-grained geometry or detailed motion. For instance, GPS-Gaussian [102] omits explicit human-specific modeling, while HFGaussian [14] incorporates human-aware cues from depth maps predicted from RGB, which can introduce temporal artifacts. A second prominent branch of work focuses on dense multi-view setups and combines Gaussian Splatting with parametric body models such as SMPL, typically deformed using Linear Blend Skinning (LBS) [36], to achieve more accurate and controllable human representations [39, 55, 58, 75, 67, 50, 69, 43, 25]. This integration enables enhanced realism in modeling dynamic human poses and detailed body shapes. However, these methods often rely on fixed body templates and complex pipelines, which can limit their flexibility and generalization across diverse identities or motion types [39, 75]. Our method bridges these two branches by adopting a lightweight, model-free approach that leverages sparse 3D pose cues to introduce human-specific structure without relying on explicit body templates. This design retains the efficiency benefits of sparse-camera setups while improving adaptability to diverse poses and appearances, enabling temporally stable and geometrically consistent synthesis across challenging motion scenarios.

3 The PoseGaussian Method

Fig. 2 illustrates the overall architecture, which comprises three primary encoders: an image encoder (*IE*) for extracting dense visual features, a pose encoder (*PE*) for interpreting skeletal keypoint distributions, and a depth encoder (*DE*) for estimating scene geometry. The pipeline takes input images processed by the IE and fuses pose heatmaps from a pre-trained estimator with image features, providing the DE with human-centric cues to improve scene geometry (§ 3.1). In parallel, the keypoint heatmaps are further processed through a Temporal Pose Stabilizer (TPS) module and subsequently encoded by the PE. The resulting pose features are decoded with the help of joint skip connections to reinforce temporal consistency and preserve semantic details (§ 3.2). The final output consists of 2D Gaussian parameter maps $\mathbf{G}(x) = \{\mathcal{M}_\tau(x)\}$, where $\tau \in \{\mathrm{p}, \mathrm{c}, \mathrm{r}, \mathrm{s}, \alpha\}$ denotes the projected position, color, rotation, scale, and opacity at each pixel location x.

3.1 Pose-Aware Feature Fusion for Depth Map Inference

To improve geometric accuracy under motion, we leverage pose priors that inform the visual representation by introducing semantically aligned structural guidance. Simple pose heatmaps—such as those

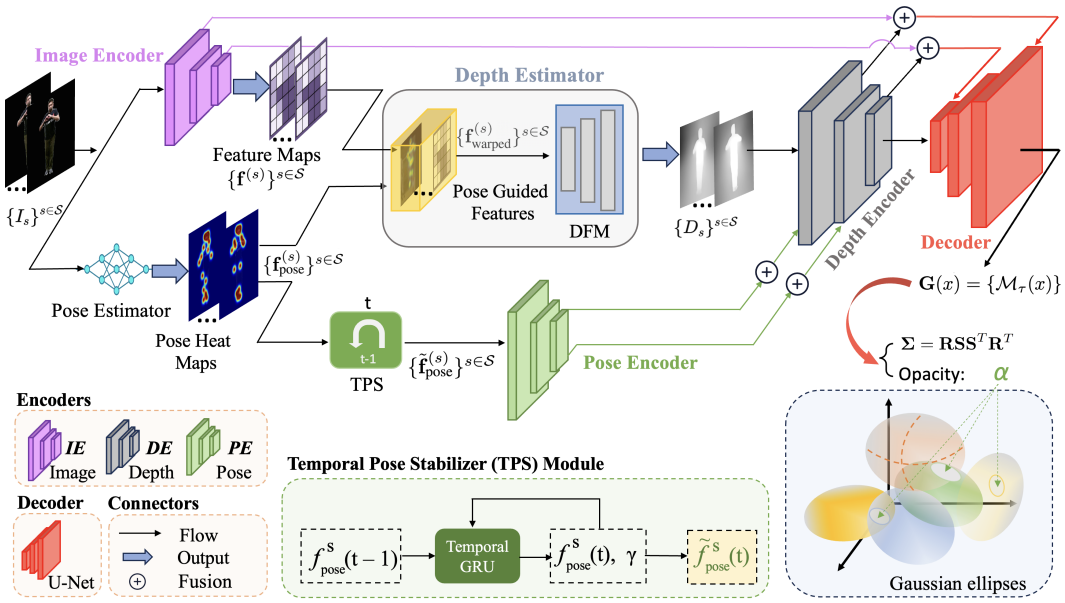

Figure 2: The PoseGaussian pipeline. *Top:* The overall workflow, illustrating the process from input color images to the predicted Gaussian parameter maps, specifically the rotation $\mathcal{M}_r(x)$, scale $\mathcal{M}_s(x)$, and opacity $\mathcal{M}_\alpha(x)$. *Bottom:* A detailed view of the Temporal Pose Stabilizer (TPS) module, along with visual annotations clarifying the roles of various modules and connections.

from BlazePose [6]—are lightweight, resolution-aligned, and readily compatible with the *PE*, making them well-suited for this purpose. Our method remains agnostic to the pose estimation backbone: when pose data is available in different formats (e.g., from previous works on robust pose estimation [49, 101]), joints can be projected and encoded as 3D Gaussian heatmaps via joint-to-heatmap encoding, enabling support for both monocular and multi-view sources [70].

From each source view $s \in \mathcal{S}$, two types of features are obtained: image features are extracted using a convolutional encoder, while pose features are generated directly by a pose estimation network [6].

$$\mathbf{f}_{\text{img}}^{(s)} = \mathcal{E}_{\text{img}}(I_s) \in \mathbb{R}^{H/2^k \times W/2^k \times D_i}, \quad \mathbf{f}_{\text{pose}}^{(s)} = P_s \in \mathbb{R}^{H/2^k \times W/2^k \times D_p} \tag{1}$$

where P_s and I_s are the 2D pose and RGB image from the s-th source view, respectively. The spatial resolution $H/2^k \times W/2^k$ reflects the output size after k stages of downsampling in the encoder. Here, D_i and D_p denote the channel dimensions for the image and pose features, respectively. The two streams are fused along the channel dimension to form the unified feature map $\mathbf{f}^{(s)}$, where simple concatenation or alternative fusion strategies (e.g., [23, 77]), as discussed in the ablation study. The fused feature set $\{\mathbf{f}^{(s)}\}^{s\in\mathcal{S}}$ is first warped to the target view using the corresponding extrinsics, producing the warped feature representation $\{\mathbf{f}_{\text{warped}}^{(s)}\}^{s\in\mathcal{S}}$. Following the design of RAFT-Stereo [47], we construct a correlation-based feature volume by aggregating view-wise similarities between the warped source features $\mathbf{f}_{\text{warped}}^{(s)}$ and the target view $\mathbf{f}_{\text{warped}}^{(t)}$:

$$\mathbf{C}_{ijk} = \sum_{s\in\mathcal{S}} \sum_{h} \mathbf{f}_{\text{warped}}^{(t)}(i,j,h) \cdot \mathbf{f}_{\text{warped}}^{(s)}(i,k,h) \tag{2}$$

Here, (i,j,h) and (i,k,h) represent spatial locations and feature channels, with h indexing the concatenated color and depth features, and j and k indexing different spatial positions. The resulting correlation features form a 3D volume $\mathbf{C}$, which is then fed into a lightweight GRU-based update module, referred to as the *DFM* (Depth Refinement Module) in Fig. 2, that iteratively refines the depth prediction [47].

$$D_s^{(t)} = \Phi_{\text{depth}}(\mathbf{C}, \{K_s\}_{s\in\mathcal{S}}; D_s^{(t-1)}), \quad t = 1, \ldots, T \tag{3}$$

where $D_s^{(t)}$ denotes the estimated depth map for source view s at iteration t, Φ_{depth} is a lightweight GRU-based update module, and $\{K_s\}_{s\in\mathcal{S}}$ are the camera intrinsics of the source views. At each

iteration, the module refines the depth estimate using the previous prediction $D_s^{(t-1)}$ and local correlation slices extracted from the cost volume $\mathbf{C}$. Unlike one-shot predictors, this recurrent formulation enables the system to capture fine-grained spatial cues and gradually resolve ambiguities in challenging regions.

3.2 Temporal Pose Stabilization for Robust Feature Guidance

Beyond depth generation, pose features play a crucial role in enhancing the decoding stage for Gaussian parameter prediction. As illustrated in Fig. 2, the heatmap output $\{\mathbf{f}_{\text{pose}}^{(s)}\}^{s\in\mathcal{S}}$ branches into a secondary pathway directed to the pose decoder, whose feature maps are then fused with the corresponding outputs from the depth and image encoders. Together, these fused features form the skip connections (indicated by red arrows) that feed into the decoder, which adopts a U-Net-style architecture to predict per-pixel Gaussian attributes, including rotation map $\mathcal{M}_r$, scale map $\mathcal{M}_s$, and opacity map $\mathcal{M}_\alpha$. Pose heatmaps serve as a dynamic auxiliary stream, injecting semantic human structural cues into skip connections to reinforce encoder-decoder alignment and preserve structural consistency during occlusion and fine-grained motion. To ensure reliable pose-based guidance in dynamic scenes, we introduce a *Temporal Pose Stabilizer (TPS)* module that processes heatmaps across adjacent frames. Inspired by temporal filtering techniques in video-based pose estimation [7, 15], TPS employs a lightweight recurrent mechanism to smooth pose signals and suppress jitter caused by motion or occlusion. This is particularly important for dynamic, non-rigid subjects like humans, where even minor inconsistencies in pose estimates can degrade reconstruction quality. Specifically, TPS takes the pose heatmaps from two consecutive frames, denoted as $\mathbf{f}_{\text{pose}}^{(s)}(t)$ and $\mathbf{f}_{\text{pose}}^{(s)}(t-1)$, and applies a temporal filter to obtain a smoothed pose signal $\tilde{\mathbf{f}}_{\text{pose}}^{(s)}(t)$:

$$\tilde{\mathbf{f}}_{\text{pose}}^{(s)}(t) = \gamma \cdot \mathbf{f}_{\text{pose}}^{(s)}(t) + (1-\gamma) \cdot \mathbf{f}_{\text{pose}}^{(s)}(t-1) \tag{4}$$

where $\gamma \in [0,1]$ is a blending factor that controls the contribution of the current frame and the previous frame's pose signal. We use the past frame $(t-1)$ instead of the future frame $(t+1)$ to ensure causality, which is essential for real-time inference and streaming scenarios. This recurrent mechanism ensures that the pose representation is temporally stable and smooth across frames. Importantly, TPS operates as a pre-processing step before pose encoding, preserving the overall structure of the feature fusion and depth estimation modules.

3.3 Pose-Conditioned Loss Objective

Our training objective integrates pose-aware supervision into the standard differentiable rendering framework, enhancing both geometric fidelity and human-centric feature guidance. The overall loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{render}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pose-fusion}} \tag{5}$$

While prior works [58, 43, 102] typically optimize only for image reconstruction and depth consistency, our formulation explicitly introduces a pose-conditioned feature alignment loss, enabling stronger structural supervision during training. Specifically, the photometric rendering loss $\mathcal{L}_{\text{render}}$ combines pixel-level fidelity and structural similarity between the rendered view $\hat{I}$ and ground-truth I:

$$\mathcal{L}_{\text{render}} = \beta\, \mathcal{L}_{\text{MAE}}(\hat{I}, I) + \gamma\, \mathcal{L}_{\text{SSIM}}(\hat{I}, I) \tag{6}$$

where β and γ balance the reconstruction terms. The depth supervision term $\mathcal{L}_{\text{depth}}$ follows an exponentially weighted scheme to encourage consistent depth predictions across multiple decoding stages:

$$\mathcal{L}_{\text{depth}} = \sum_{t=1}^{T} \mu^{T-t}\, \left\| d_t - d_{\text{gt}} \right\|_1 \tag{7}$$

where d_t denotes the predicted depth at stage t, d_{gt} is the ground-truth depth, and μ controls the decay rate. Finally, the pose-fusion loss $\mathcal{L}_{\text{pose-fusion}}$ supervises the intermediate feature decoding by aligning fused features f_{joint} with pose-encoded features f_{pose}:

$$\mathcal{L}_{\text{pose-fusion}} = \lambda\, \left\| f_{\text{joint}} - f_{\text{pose}} \right\|_1 \tag{8}$$

This term ensures that human pose structures are faithfully retained during feature decoding, providing richer spatial guidance for the 3D reconstruction.

ZJU-MoCap					Twindom			
Method	PSNR↑	SSIM↑	LPIPS↓	Train/FPS	Method	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [CVPR'21] [92]	24.71	0.892	0.120	1h / 1.20	KeypointNeRF [ECCV'22] [52]	19.68	0.890	-
NHP [ANIPS'21] [41]	28.25	0.955	0.065	1h / 0.15	PixelHuman [arXiv'23] [76]	24.20	0.948	-
NB [CVPR'21] [62]	29.03	0.964	0.043	10h / 1.48	3D-GS [ACMTOG'23] [38]	22.77	0.785	0.153
AN [CVPR'21] [61]	29.77	0.965	0.470	10h / 1.11	FloRen [SIGGRAPH'22] [73]	22.96	0.838	0.165
AS [TPAMI'24] [104]	30.38	0.975	0.037	10h / 0.40	IBRNet [CVPR'21] [83]	22.92	0.803	0.238
ENeRF [SIGGRAPH'22] [45]	28.90	0.967	-	-	Ours	**24.28**	**0.959**	**0.101**
					HuMMan			
					Method	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF [arXiv'22] [89]	30.66	0.969	0.033	10h / 0.30	NHP[NIPS'21] [40]	18.99	0.845	0.182
DVA [SIGGRAPH'22] [71]	29.45	0.956	0.038	1.5h / 16.5	MPS-NERF [SIGGRAPH'22] [45]	17.44	0.824	0.193
InstantNVR [CVPR'23] [19]	31.01	0.971	0.038	5m / 2.20	SHERF [ICCV'23] [27]	20.83	0.891	0.125
InstantAvatar [CVPR'23] [33]	29.73	0.938	0.068	3m / 4.15	GHG [arXiv'24] [10]	**23.86**	0.952	**0.0591**
KeypointNeRF [ECCV'22] [52]	25.03	0.898	0.104	20h / 1.05	GST [arXiv'24] [66]	18.40	0.87	0.14
Humannerf [CVPR'22] [100]	30.24	0.9679	-	-	Ours	22.18	**0.977**	0.060
					DNA-Rendering			
					Method	PSNR↑	SSIM↑	LPIPS↓
FlexNeRF [CVPR'23] [30]	31.73	0.9767	0.29	-	HuGS [CVPR'24] [55]	31.5	0.98	0.022
MonoHuman [CVPR'23] [94]	30.05	0.9684	0.031	-	DVA [SIGGRAPH'22] [71]	29.8	0.97	0.025
Sem-Human [arXiv'23] [96]	30.80	0.967	0.033	-	ENeRF [SIGGRAPH'22] [45]	28.108	0.972	0.056
SMPLPix [WACV'21] [65]	27.00	0.91	0.090	-	IBRNet [SIGGRAPH'23] [2]	27.844	0.967	0.081
GP-NeRF [CVPR'23] [94]	28.80	0.9408	0.090	20h / 1.05	Im4D [SIGGRAPH'23] [46]	28.991	0.973	0.062
AniNeRF [CVPR'21] [61]	24.56	0.89	0.12	-	4K4D [CVPR'24] [91]	**31.173**	0.976	0.055
INR [TPAMI'23] [63]	30.54	0.970	-	-	Ours	30.18	**0.989**	**0.012**
					People-Snapshot			
					Method	PSNR↑	SSIM↑	LPIPS↓
HumanSplat [NIPS'24] [57]	29.82	0.9396	0.1048	-	HumanNeRF [arXiv'21] [89]	26.90	0.9605	0.018
GPS-Gaussian [CVPR'24] [102]	29.68	0.95	-	-	Neural Dressing [CVPR'21] [20]	-	0.91	0.07
Deform3GS [SIGGRAPH'22] [45]	24.10	0.869	0.126	-	InstantAvatar [CVPR'23] [19]	29.53	0.9716	0.016
GoMAvatar [CVPR'24] [86]	30.37	0.9689	0.032	23m / 4.15	Anim-NeRF [arXiv'21][9]	29.37	0.970	0.017
HuGS [CVPR'24] [55]	26.58	0.934	0.022	-	Neural Body [CVPR'21] [62]	25.49	0.928	-
SplatArmor [arXiv'23] [31]	30.24	-	0.032	-	GoMAvatar [CVPR'24] [86]	30.68	0.9767	0.0213
3DGS-Avatar [CVPR'24] [69]	30.61	0.9703	-	30m / 20	SelfRecon [CVPR'22] [32]	24.91	-	0.061
Deform 3D [arXiv'23] [34]	29.28	0.964	0.040	-	SMPLPix [WACV'21] [65]	17.90	-	0.165
HUGS [CVPR'24] [39]	30.54	0.970	0.030	-	FlexNeRF [CVPR'23] [30]	28.77	0.904	0.035
GauHuman [CVPR'24] [25]	31.34	0.965	0.031	**1m / 189**	SplatArmor [arXiv'23] [31]	27.08	-	0.43
GART [CVPR'24] [42]	**32.22**	0.977	0.29	-	GART [CVPR'24] [42]	30.40	0.976	0.037
Ours	30.86	**0.979**	**0.028**	30m / 100	Ours	**32.86**	**0.98**	**0.014**

Table 1: Comparison of methods on ZJU-MoCap [62], Twindom[82], HuMMan[8], DNA-Rendering[11] and People-Snapshot [3] datasets.

4 Experiment Results

Brief Model Description: Our model adopts a U-Net architecture with three parallel encoders, each beginning with a 3×3 convolution (32 channels), followed by six residual units with progressively increasing widths (32, 64, 96, 128). Each residual block incorporates a Squeeze-and-Excitation (SE) module for channel-wise attention [24]. Skip connections align encoder and decoder features to preserve multi-scale information. The decoder upsamples the features and outputs three heads: (1) an SR-Opacity head predicting Gaussian parameters—scale (3 channels, Softplus), rotation (3 channels, normalized), and opacity (2 channels, Sigmoid); (2) a depth head for auxiliary supervision; and (3) a confidence head for per-pixel uncertainty. The final prediction $\hat{\mathbf{G}}(x)$ is a blend of the predicted output $\mathbf{G}_{\text{dec}}(x)$ and prior $\mathbf{G}_{\text{prior}}(x)$, modulated by the confidence maps $\mathbf{c}(x)$:

$$\hat{\mathbf{G}}(x) = \sigma(\mathbf{c}(x)) \cdot \mathbf{G}_{\text{dec}}(x) + (1 - \sigma(\mathbf{c}(x))) \cdot \mathbf{G}_{\text{prior}}(x) \tag{9}$$

where $\mathbf{c}(x)$ is a learned confidence map. A broader discussion of uncertainty-aware blending strategies can be found in [37]. Additional implementation details are provided in the Appendix.

4.1 Comparison to State-of-the-arts

Dataset and Evaluation Protocols: To ensure a fair and comprehensive comparison, we evaluate PoseGaussian on a diverse set of widely used benchmarks for dynamic human reconstruction, spanning both controlled multi-view studio setups (e.g., ZJU-MoCap) and in-the-wild scenarios with varied poses, clothing, and motion complexity (e.g., THuman2.0, HuMMan, DNA-Rendering). To qualitatively assess the role of pose under diverse motion conditions, we further collect a custom dataset featuring 28 subjects performing natural and varied motions in everyday environments using a simple multi-camera setup. This configuration emphasizes generalization to fast and unconstrained human movements. For quantitative evaluation, we adopt the widely used metrics PSNR [29],

Method	THuman2.0 Dataset											
	Training						Testing					
	6-camera setup			8-camera setup			Real World Data			Rendered Data		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3D-GS [[ACMTOG'23]] [38]†	-	-	-	-	-	-	22.97	0.839	0.125	24.18	0.821	0.144
FloRen [SIGGRAPH'22] [73]	18.72	0.770	0.267	23.26	0.812	0.184	22.80	0.872	0.136	23.26	0.812	0.184
IBRNet [CVPR'21] [83]	21.08	0.790	0.263	23.38	0.836	0.212	22.63	0.852	0.177	23.38	0.836	0.212
NARF [SIGGRAPH'22] [45]	-	-	-	-	-	-	21.80	0.8088	-	-	-	-
PIFu [ICCV'19] [72]	-	-	-	-	-	-	-	-	-	20.40	0.921	0.079
ENeRF [SIGGRAPH'22] [45]	21.78	0.831	0.181	24.10	0.869	0.126	23.26	0.893	0.118	24.10	0.869	0.126
SHERF [ICCV'23] [26]	-	-	-	-	-	-	24.26	0.91	0.11	-	-	-
Contex-Human [CVPR'24] [18]	-	-	-	-	-	-	-	-	-	21.40	0.923	0.063
DoubleField [CVPR'22] [74]	-	-	-	-	-	-	25.10	0.905	–	-	-	-
GPS-Gaussian [CVPR'24] [102]	23.03	0.884	0.168	25.57	0.898	0.112	24.64	0.917	0.088	25.57	0.898	0.112
FreeSplat [NIPS'24] [84]	23.35	0.843	0.184	–	–	–	25.90	0.808	0.252	-	-	-
RoGSplat*[arXiv'25] [90]	23.12	0.8980	0.1661	–	–	–	**25.99**	0.9452	0.057	-	-	-
HumanSplat [NIPS'24] [57]	-	-	-	-	-	-	-	-	-	24.033	0.918	0.055
LIFe-GoM [arXiv'25] [87]	-	-	-	24.65	-	0.110	25.32	–	0.099	-	-	-
SIFU [CVPR'24] [99]	-	-	-	-	-	-	-	-	-	22.102	0.923	0.079
HumanSGD [SIGGRAPH'23]† [2]	-	-	-	-	-	-	-	-	-	17.365	0.895	0.130
TeCH† [3DV'24] [28]	-	-	-	-	-	-	-	-	-	25.211	0.936	0.083
Ours	**24.9**	**0.94**	**0.09**	**26.07**	**0.927**	**0.081**	25.47	**0.966**	**0.05**	**25.78**	**0.957**	**0.031**

Table 2: Comparison of methods on the THuman2.0 test set using 6-camera and 8-camera setups, covering both Real World and Rendered data.

SSIM [85], and LPIPS [97], which respectively measure pixel-level accuracy, structural similarity, and perceptual fidelity.

Quantitative Benchmark Comparison: Table 1 presents a comprehensive quantitative evaluation of our method against state-of-the-art approaches across a diverse set of public benchmark datasets. By leveraging pose as guidance, our method consistently achieves superior structural consistency, attaining the highest SSIM scores across all datasets—including a peak score of 0.979 on ZJU-MoCap [62]. The integration of structural cues and temporal consistency further enhances perceptual quality, as reflected by favorible LPIPS scores ranging from 0.014 to 0.101, underscoring the method's effectiveness in preserving fine appearance details and visual coherence. While our method did not achieve the highest PSNR, it remains competitive, which is reasonable given that PSNR primarily measures pixel-wise differences and may not fully capture perceptual or temporal quality improvements. In addition, Table 2 focuses exclusively on the THuman 2.0 dataset, where *6-view* and *8-view* settings on selected sequences are employed to construct controlled benchmarks. This targeted analysis enables a more rigorous evaluation under varying view configurations, which is essential for assessing pose-guided novel view synthesis. Across all tested conditions on THuman 2.0, our method consistently outperforms competing methods, achieving an SSIM score of 0.957 and an LPIPS score of 0.031, demonstrating strong generalization and rendering quality.

Demonstrations

We qualitatively compare our approach under two challenging scenarios: occlusion and fast motion. As shown in Fig. 3, the virtual view is rendered from a significantly different viewpoint than the input cameras, revealing heavily self-occluded regions such as the back and rear arm in scene 1, and the inner arms and buttocks in scene 2. Our method better preserves contours and structure in these regions by enforcing pose constraints that keep Gaussian primitives aligned with the body. Fig. 4 evaluates reconstruction quality across motions of varying speeds. *Row 1* (slow motion) shows only subtle differences between methods, while faster motions in *Row 2* and *Row 3* reveal more prominent artifacts, including "snowflake" noise and surface flickering. Our method delivers more stable and coherent results overall, though minor artifacts still appear under rapid motion. In our testing, for some scenes, methods like GauHuman [25] and AS [104] show reduced artifacts,

Ours GauHuman[25]

Figure 3: Occlusion Scenario on challenging views revealing occluded regions (e.g., back, inner arms).

Figure 4: Fast Motion Scenario. Comparison with recent NeRF- and Gaussian-based methods on sequences with increasing motion speed (top: slow, middle: medium, bottom: fast).

likely due to their use of explicit body representations to guide reconstruction.

Temporal Coherence and Motion Stability: To demonstrate the role of pose in enhancing temporal coherence, we conduct a comprehensive evaluation using SSIM, PSNR, and LPIPS. Here, we present SSIM as a representative example. Our evaluation covers 32 motion sequences featuring challenging dynamics such as rapid limb movement. For statistical robustness, we analyze the average and standard deviation of per-frame ΔSSIM, as well as ΔSSIM values averaged across all sequences. As shown in Table 3, our method achieves the lowest $\mu(\Delta\text{SSIM})$ of 0.023 and $\sigma(\Delta\text{SSIM})$ of 0.015, almost half of the next best method. Due to space constraints, we report only a subset of methods here; a comprehensive evaluation of representative state-of-the-art methods, including PSNR and LPIPS, is provided in the supplementary material.

Method	$\mu(\Delta\text{SSIM})$	$\sigma(\Delta\text{SSIM})$
3DGS [38]	0.052	0.039
IBRNet [83]	0.0451	0.035
GPS-Gaussian [102]	0.040	0.028
Ours	**0.023**	**0.015**

Table 3: Frame-to-frame ΔSSIM consistency.

4.2 Ablation Study

Fusion Strategies: We explored several fusion strategies to combine the outputs $\{\mathbf{f}^{(s)}\}^{s\in\mathcal{S}}$ and $\{\mathbf{f}^{(s)}_{pose}\}^{s\in\mathcal{S}}$ from Eq. 1, including concatenation, Feature-wise Attention [98], and Gated Fusion [5]. Feature-wise Attention achieved the best performance, with the lowest EPE (0.75) and 1px accuracy of 78%, at a moderate model size (1.2M parameters). Concatenation offered a favorable trade-off between accuracy and complexity (EPE 0.80, 1px 80%, 0.8M parameters), while Element-wise Addition provided a simpler, more lightweight option (0.5M parameters) but with higher EPE (1.00). Weighted Average Fusion [77] and Element-wise Multiplication also delivered reasonable compromises (EPE $\sim$1.1 and 0.98; 0.9–1.0M parameters). More complex methods such as Outer Product

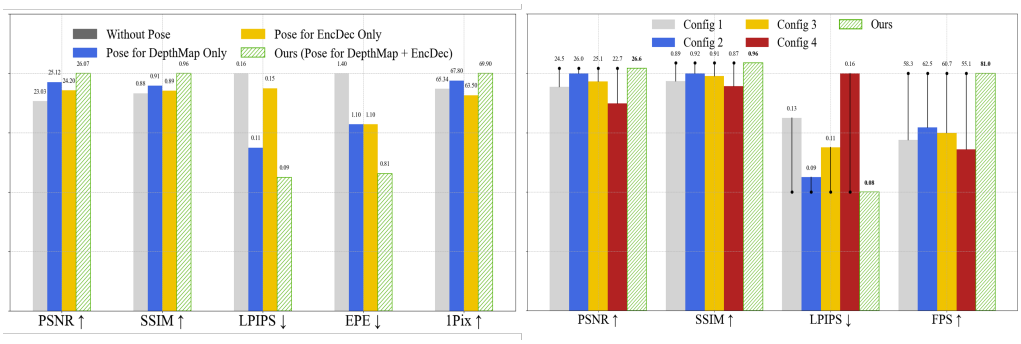

Figure 5: (Left) Impact of pose information. (Right) Impact of pose encoder configurations.

Fusion [79] increased parameter counts substantially (1.3–1.8M) with only marginal improvements. Based on these results, we select concatenation as the best balance between accuracy and efficiency.

Loss Function: To determine the hyperparameters β, γ, and λ introduced in the loss terms (Eq. 6 and Eq. 8), we evaluate performance on the THuman2.0 dataset [93]. Without pose fusion ($\lambda = 0$), a balanced rendering loss ($\beta = 0.5, \gamma = 0.5$) achieves PSNR of 19.9, SSIM of 0.879, and LPIPS of 0.205. SSIM-heavy settings (e.g., $\beta = 0.3, \gamma = 0.7$) slightly improve SSIM (0.871) but worsen LPIPS (0.212), while MAE-heavy settings (e.g., $\beta = 0.8, \gamma = 0.2$) increase PSNR at the cost of perceptual quality. Introducing pose fusion ($\lambda = 0.1$) improves all metrics, and moderate fusion ($\lambda = 0.5$) further enhances SSIM to 0.905 and reduces LPIPS to 0.148. The best results are achieved with full pose fusion ($\lambda = 1.0$) and balanced rendering loss ($\beta = 0.5, \gamma = 0.5$), yielding PSNR of 28.5, SSIM of 0.940, and LPIPS of 0.090.

Pose and Architecture Effects: Fig. 5 illustrates the effects of pose feature integration (left) and network architecture variations (right) on performance. Integrating pose inputs into both the depth map and encoder–decoder pathways delivers the best results, significantly outperforming configurations that incorporate pose in only one component or omit it altogether. This integration notably enhances perceptual quality, improving the LPIPS score by 45% (0.09 vs. 0.16) compared to the model without pose information. For network design, we evaluated four alternative architecture variants—detailed in the Appendix—that explore the use of Batch Normalization, Residual Blocks, and downsampling layers. Their performance is summarized in Fig. 5 (right).

5 Broader Impacts and Limitations

Our work contributes to improved human rendering quality, which can benefit applications such as virtual try-on, telepresence, and animation, enhancing accessibility and user experiences in digital environments. However, as with any technology involving human modeling, it may raise concerns related to privacy and potential misuse, such as in surveillance or deepfake creation. While we do not directly address these risks, we emphasize the importance of ethical deployment. Although our method achieves competitive scores and improves virtual view quality with motion and a certain level of occlusion, it still cannot completely remove artifacts in challenging scenarios such as very fast body movement or severe occlusion. Future work will address these limitations.

6 Conclusion

We demonstrate that leveraging pose as a structural prior combined with temporal constraints enables our method to achieve state-of-the-art performance on standard benchmarks (e.g., SSIM and LPIPS) across multiple public datasets. PoseGaussian effectively reduces motion-induced artifacts in dynamic human scenes, resulting in superior temporal coherence and visual stability in novel view synthesis. While our approach significantly improves reconstruction quality under challenging fast motions, some artifacts persist, especially when the virtual viewpoint shifts considerably from the original capture. In future work, we plan to explore pose-guided post-processing strategies that refine depth or geometry estimates after initial rendering, aiming to further reduce residual artifacts and improve reconstruction fidelity in extreme motion scenarios.

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62, 2020.

[2] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. ACM, 2005. doi: 10.1145/1073204.1073209.

[5] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[6] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking, 2020.

[7] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely labeled videos. In *Advances in Neural Information Processing Systems 33*, 2019.

[8] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022.

[9] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021.

[10] Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. Generalizable human gaussians from single-view image. *arXiv preprint arXiv:2406.06050*, 2024.

[11] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023.

[12] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69:1–69:13, 2015.

[13] Frank Dellaert and Lin Yen-Chen. Neural volume rendering: Nerf and beyond, 2021.

[14] Arnab Dey, Cheng-You Lu, Andrew I. Comport, Srinath Sridhar, Chin-Teng Lin, and Jean Martinet. Hfgaussian: Learning generalizable gaussian human with integrated human features, 2024.

[15] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17131–17141, 2023.

[16] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2016.

[17] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510. IEEE, 2022.

[18] Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, and Long Quan. Contex-human: Free-view rendering of human from a single image with texture-consistent synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10084–10094, 2024.

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

[20] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5151–5160, 2021.

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

[22] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5875–5884. IEEE, 2021.

[23] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[25] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20418–20431. IEEE, 2024.

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

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

[28] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542. IEEE, 2024.

[29] Quang Huynh-Thu and Mahdad Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008.

[30] Vinoj Jayasundara, Amit Agrawal, Nicolas Heron, Abhinav Shrivastava, and Larry S. Davis. Flexnerf: Photorealistic free-viewpoint rendering of moving humans from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21118–21127, 2023.

[31] Rohit Jena, Ganesh Subramanian Iyer, Siddharth Choudhary, Brandon Smith, Pratik Chaudhari, and James Gee. Splatarmor: Articulated gaussian splatting for animatable humans from monocular rgb videos. *arXiv preprint arXiv:2311.10812*, 2023.

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

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

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

[35] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018.

[36] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games (I3D)*, pages 39–46. ACM, 2007.

[37] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 2017.

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

[39] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 505–515. IEEE, 2024.

[40] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24741–24752. Curran Associates, Inc., 2021.

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

[42] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19876–19887, 2024.

[43] Mengtian Li, Shengxiang Yao, Zhifeng Xie, and Keyu Chen. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024.

[44] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6498–6508, 2021.

[45] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[46] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–9, 2023.

[47] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 218–227, 2021.

[48] Lingjie Liu, Marc Habermann, Vladislav Rudnev, Kiran Sarkar, Jiakai Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control, 2021.

[49] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020.

[50] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models, 2024.

[51] Matthew Loper, Nanyang Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6): 1–16, 2015.

[52] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, pages 179–197. Springer, 2022.

[53] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020.

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

[55] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024.

[56] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020.

[57] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 74383–74410. Curran Associates, Inc., 2024.

[58] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024.

[59] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5865–5874, 2021.

[60] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields, 2021.

[61] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14314–14323. IEEE, 2021.

[62] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9054–9063, 2021.

[63] Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Implicit neural representations with structured latent codes for human body modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9895–9907, 2023. doi: 10.1109/TPAMI.2023.3245815.

[64] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72:179–193, 2007.

[65] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the '21'*, pages 1810–1819, 2021.

[66] Lorenza Prospero, Abdullah Hamdi, Joao F Henriques, and Christian Rupprecht. Gst: Precise 3d human body from a single image with gaussian splatting transformers. *arXiv preprint arXiv:2409.04196*, 2024.

[67] Lorenza Prospero, Abdullah Hamdi, Joao F. Henriques, and Christian Rupprecht. Gst: Precise 3d human body from a single image with gaussian splatting transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

[68] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327. IEEE, 2021.

[69] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024.

[70] Haoxuan Qu, Li Xu, Yujun Cai, Lin Geng Foo, and Jun Liu. Heatmap distribution matching for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24327–24339. Curran Associates, Inc., 2022.

[71] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022.

[72] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019.

[73] Ruizhi Shao, Liliang Chen, Zerong Zheng, Hongwen Zhang, Yuxiang Zhang, Han Huang, Yandong Guo, and Yebin Liu. Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–10, 2022.

[74] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15872–15882, 2022.

[75] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024.

[76] Gyumin Shim, Jaeseong Lee, Junha Hyung, and Jaegul Choo. Pixelhuman: Animatable neural radiance fields from few images. *arXiv preprint arXiv:2307.09070*, 2023.

[77] Le Song, Yuchi Lin, Weichang Feng, and Meirong Zhao. A novel automatic weighted image fusion algorithm. In *2009 International Workshop on Intelligent Systems and Applications*, pages 1–4. IEEE, 2009.

[78] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021.

[79] Fabricio S Terra, Raphael A Viscarra Rossel, and Jose AM Dematte. Spectral fusion by outer product analysis (opa) to improve predictions of soil organic c. *Geoderma*, 335:35–46, 2019.

[80] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. Scanning 3d full human bodies using kinects. *IEEE Transactions on Visualization and Computer Graphics*, 18(4):643–650, 2012.

[81] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Nonrigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12959–12970. IEEE, 2021.

[82] Twindom. Twindom: Full body 3d scanners for 3d printed figurines, 3d portraits, 3d selfies, and avatar products, 2025.

[83] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2021.

[84] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 107326–107349. Curran Associates, Inc., 2024.

[85] Z Wang. Image quality assessment: Form error visibility to structural similarity. volume 13, pages 604–606, 2004.

[86] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G. Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2059–2069, 2024.

[87] Jing Wen, Alexander G Schwing, and Shenlong Wang. Life-gom: Generalizable human rendering with learned iterative feedback over multi-resolution gaussians-on-mesh. *arXiv preprint arXiv:2502.09617*, 2025.

[88] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild, 2020.

[89] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video, 2022.

[90] Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, and Wei-Shi Zheng. Rogsplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2503.14198*, 2025.

[91] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20029–20040, 2024.

[92] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021.

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

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

[95] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4), 2021.

[96] Jie Zhang, Pengcheng Shi, Zaiwang Gu, Yiyang Zhou, and Zhi Wang. Semantic-human: Neural rendering of humans from monocular video with human parsing. *arXiv preprint arXiv:2308.09894*, 2023.

[97] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[98] Sheng Zhang, Min Chen, Jincai Chen, Fuhao Zou, Yuan-Fang Li, and Ping Lu. Multimodal feature-wise co-attention method for visual question answering. *Information Fusion*, 73:1–10, 2021.

[99] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024.

[100] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, 2022.

[101] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. pages 11656–11665, 2021.

[102] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19680–19690, 2024.

[103] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

[104] Xiaowei Zhou, Sida Peng, Zhen Xu, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, and Hujun Bao. Animatable implicit neural representations for creating realistic avatars from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4147–4159, 2024.

[105] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars, 2025.

Appendix

3D Gaussian Splatting Preliminary The proposed *PoseGaussian* approach builds upon the core principles of 3D Gaussian Splatting (3D-GS) [38]. In 3D-GS, a static scene is modeled using a collection of spatial primitives, where each primitive is parameterized as an anisotropic Gaussian distribution centered at a 3D location. The density at a spatial point $\mathbf{x} \in \mathbb{R}^3$ due to the i-th Gaussian is given by:

$$g_i(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right),$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the center of the Gaussian, and $\Sigma_i \in \mathbb{R}^{3\times3}$ is the covariance matrix that encodes both the shape and orientation. This covariance can be decomposed as $\Sigma_i = R_i S_i R_i^T$, where R_i is a rotation matrix and S_i is a diagonal scaling matrix representing variances along the local principal axes.

When the 3D Gaussians are projected into 2D image space, their covariances are transformed via a view transformation matrix W and a Jacobian matrix J from an affine approximation of the camera projection. The resulting 2D covariance becomes:

$$\Sigma_i' = JW\Sigma_i W^T J^T,$$

preserving the geometric effects of both rotation and scale during projection.

In our approach, we predict a dense 2D map of Gaussian parameters per pixel location x, denoted as:

$$\mathbf{G}(x) = \{\mathcal{M}_\tau(x)\}, \quad \tau \in \{\mathrm{p}, \mathrm{c}, \mathrm{r}, \mathrm{s}, \alpha\},$$

where $\mathcal{M}_\mathrm{p}(x)$ gives the projected center position, $\mathcal{M}_\mathrm{c}(x)$ the color, $\mathcal{M}_\mathrm{r}(x)$ the rotation, $\mathcal{M}_\mathrm{s}(x)$ the scale, and $\mathcal{M}_\alpha(x)$ the opacity. These parameters are designed to mirror the structure of 3D-GS after projection:

- $\mathcal{M}_c$ represents the pixel color, which is directly retrieved as the mean color value at each pixel location.
- $\mathcal{M}_p$ denotes the projected pixel position, computed by transforming the estimated depth map through camera intrinsics and projecting it into the virtual view.
- $\mathcal{M}_\mathrm{r}(x)$ corresponds to the local rotation matrix R_i used in the covariance decomposition;
- $\mathcal{M}_\mathrm{s}(x)$ encodes per-axis scale values analogous to the diagonal entries in S_i;
- $\mathcal{M}_\alpha(x)$ defines the pixel-level opacity, which is used in Gaussian alpha blending as described in the following color blending equation.

The final rendered color at each pixel x is then computed using front-to-back compositing across multiple overlapping Gaussians, sorted by depth:

$$C(x) = \sum_i \alpha_i(x)\, c_i(x) \prod_{j<i} (1-\alpha_j(x)),$$

where $\alpha_i(x) = \mathcal{M}_\alpha^{(i)}(x)$ represents the pixel-level opacity, and $c_i(x) = \mathcal{M}_c^{(i)}(x)$ denotes the color directly obtained from the pixel color map $\mathcal{M}_c$. This blending formulation enables smooth, differentiable rendering and effectively captures both appearance and temporal continuity.

Therefore these five output maps—$\mathcal{M}_\mathrm{p}(x)$, $\mathcal{M}_\mathrm{c}(x)$, $\mathcal{M}_\mathrm{r}(x)$, $\mathcal{M}_\mathrm{s}(x)$, and $\mathcal{M}_\alpha(x)$—are visually illustrated in Fig. 6, showing how each component contributes to the Gaussian-based rendering process.

Implementation Details This is the architecture of our encoder and decoder:

As depicted in Fig. 6, the three parallel encoders share an identical structure: an initial 3×3 convolutional layer with 32 channels, followed by six residual units with progressively increasing

channel dimensions. To enhance modality-specific encoding and channel selectivity, each residual unit includes a Squeeze-and-Excitation (SE) block. These SE modules adaptively recalibrate channel-wise responses, mitigating modality noise and reinforcing semantically important features across diverse inputs. The choice of a 3×3 kernel size and progressive channel widths (32, 64, 96, 128) balances efficient local feature extraction with scalable semantic representation. Skip connections are formed by concatenating intermediate encoder features at corresponding layer indices and forwarding them to the decoder. This skip alignment enriches the decoder with multi-level semantic information throughout the reconstruction process.

The decoder mirrors the encoder structure to complete the U-Net architecture, progressively upsampling and refining the feature maps. At the final decoder stage, three output heads are attached: a scale-rotation-opacity (SR-Opacity) branch for predicting Gaussian parameters, a depth branch to enforce depth consistency regularization, and an uncertainty branch for predicting per-pixel confidence maps. The depth branch outputs a normalized depth map via a Conv3x3(1) layer followed by a Tanh activation. The SR-Opacity branch applies a Conv3x3(8) layer, splitting outputs into 3 channels for scale (Softplus activation), 3 channels for rotation (normalization), and 2 channels for opacity (Sigmoid activation). The uncertainty branch outputs 2 channels via a Conv3x3(2) layer followed by Sigmoid activations, where Channel 1 corresponds to depth confidence and Channel 2 to SR-Opacity confidence.

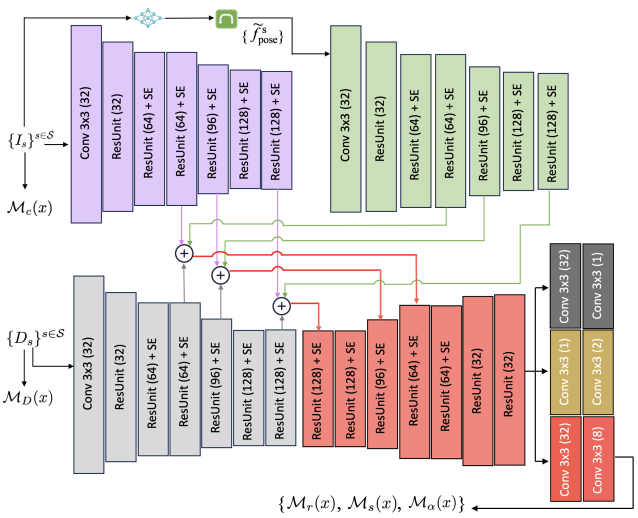

Figure 6: Architecture of the pose encoder-decoder network.

Detailed Configuration Variants and Their Impact

- **Config 1: Simplified Architecture**
  This variant uses only a single 5×5 Conv2D layer with 32 channels followed by one Residual Block. Due to its shallow depth and limited capacity, this configuration struggles to capture complex features, resulting in lower perceptual quality (LPIPS: 0.13) and reduced efficiency.
- **Config 2: Deeper Residual Learning with Skip Connection**
  This setup enhances the residual block structure by stacking two Conv2D layers with a skip connection between them. This deeper residual learning significantly improves image reconstruction, achieving the highest SSIM (0.92), lowest LPIPS (0.09), and PSNR of 26.0. These metrics indicate better preservation of structural and perceptual details compared to Config 1.
- **Config 3: Adding Batch Normalization and ReLU Activation**
  Building on Config 2, this variant introduces Batch Normalization layers and ReLU activations within the residual blocks. These additions improve feature representation and training stability, further enhancing the network's ability to generalize.
- **Config 4: No Downsampling Layers**
  In this variant, downsampling operations (e.g., strided convolutions or pooling) are removed to maintain spatial resolution throughout the encoder. While this results in a good SSIM score (0.91) and moderate PSNR (25.1), the LPIPS metric slightly degrades (0.11), reflecting a small perceptual quality drop compared to Config 2 and 3.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction clearly summarize the key contributions of the paper, including the proposed method, its performance improvements, and the scope of evaluation. The claims are consistent with the theoretical analysis and experimental results presented. Limitations and assumptions are also acknowledged, ensuring accurate representation of the work's impact and generalizability.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated section discussing limitations, clearly acknowledging challenges such as artifact persistence under extreme scenarios like fast motion and heavy occlusion. It also reflects on the scope and assumptions of the approach, providing transparency about current boundaries and future directions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper does not include any theoretical results, theorems, or formal proofs, so this question is not applicable.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the model architecture, training procedure, dataset preprocessing, and evaluation metrics necessary to reproduce the main experimental results. While the code and data may not be directly provided, the instructions and information included are sufficient for a knowledgeable researcher to replicate the key claims and conclusions.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The paper includes links to anonymized code and data in the supplementary material, along with detailed instructions to reproduce the main experimental results. We also plan to host the code and dataset on a public website for broader access after the review process.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper provides all essential experimental details, including data splits, model architectures, training procedures, hyperparameters, optimizer choices, and selection criteria. These are presented in the main text and further elaborated in the appendix to ensure clarity and reproducibility.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The paper reports error bars for the main experimental results, clearly specifying that they represent standard deviation across multiple random seeds. The method used to compute them (i.e., repeated runs with different initializations) is described in the experimental setup section. Additionally, relevant figures and tables include these error bars, and the paper explicitly states the assumptions behind the variability sources

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: The paper includes details about the computational resources used for each experiment in the Appendix. It specifies the type of compute used (e.g., NVIDIA A100 GPUs), memory configuration, and approximate run times. The total compute cost is estimated based on multiple runs, and the appendix also mentions that additional preliminary experiments were conducted but are not

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research complies with its principles. Our study does not involve human subjects, sensitive data, or deployment in high-risk settings. All methods and experiments were conducted responsibly, and we have taken appropriate steps to ensure transparency, reproducibility, and fairness.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: We include a dedicated section titled "Broader Impacts and Limitations" in our paper. This section discusses both the potential positive societal impacts of our work (e.g., advancing research in [insert brief example, such as accessible AI tools, healthcare, etc.]) as well as possible negative consequences (e.g., misuse risks, fairness concerns, etc.). We also outline potential mitigation strategies where relevant.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: Our paper does not involve the release of any models or datasets that pose a high risk for misuse. The models and data presented are task-specific, do not involve large-scale pretraining or scraping from the internet, and are unlikely to be repurposed for malicious applications.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All external assets used in the paper, including datasets and code libraries, are properly cited with corresponding references. We explicitly mention the licenses and terms of use for each asset where applicable. For example, we use [Dataset Name] under the CC-BY 4.0 license and [Library Name] under the MIT license, and include links to the official sources in the Appendix and supplementary materials.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We release new assets as part of this work, including [briefly name assets, e.g., a new dataset/model/codebase], and provide thorough documentation alongside them. The documentation includes details about data collection or model training, licensing terms, known limitations, and usage guidelines.

    Guidelines:

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

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

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

Answer: [Yes]

Justification: This research involves human subject testing for motion capture. Participants were informed according to institutional guidelines.

Guidelines:

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

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

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

Answer: [Yes]

Justification: Data collection involving human subjects was conducted with prior approval and oversight from the relevant university committee. The participants were informed about the nature of the study and any potential minimal risks. The IRB approval process is currently ongoing to formalize this approval in writing and will be completed before publication.

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

16. **Declaration of LLM usage**

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

Answer: [No]

Justification: The core methods and contributions of this research do not involve any use of large language models (LLMs). Any use of LLMs, if applicable, was limited to writing assistance or formatting and did not affect the research methodology or results.

Guidelines:

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