European Journal of AI, Computing & Informatics

Vol. 1 No.1 2025

Article **Open Access**

Dynamic Human-Scene Cooperative Novel View Synthesis Method Based on 3D Gaussian Splatting

Jinghan Wang [1,*]

Received: 09 March 2025
Revised: 15 March 2025
Accepted: 15 April 2025
Published: 18 April 2025

[1] North China Electric Power University, Beijing, 102206, China
* Correspondence: Jinghan Wang, North China Electric Power University, Beijing, 102206, China

Abstract: Dynamic human-scene cooperative novel view synthesis holds significant application value in fields such as Virtual Reality (VR), Augmented Reality (AR), film production, and digital humans. In Chapter 4, we implemented high-fidelity novel view synthesis of real human body surface details based on Neural Radiance Fields (NeRF). Although the synthesis of dynamic human surface details achieved promising results, the slow inference speed of NeRF and its implicit modeling of continuous space — lacking explicit geometric structures — make it difficult to decouple the human body from the scene. Consequently, NeRF fails to meet the requirements for dynamic human-scene cooperative novel view synthesis. Moreover, the absence of accurate semantic segmentation of humans and scenes in three-dimensional space poses a critical challenge in accurately decomposing dynamic human Gaussians and static scene Gaussians. To address these issues, this chapter proposes an efficient dynamic human-scene cooperative novel view synthesis framework based on the 3D Gaussian Splatting (3DGS) method. The framework standardizes the spatial coordinate systems of the human body and the scene to ensure geometric consistency and employs a triplane representation to reconstruct human Gaussians. Finally, a joint training strategy is adopted to simultaneously optimize the human and scene models. Comparative experiments on publicly available datasets demonstrate that the proposed method effectively corrects Gaussian misalignment caused by geometric coupling between the human body and the scene. This results in more accurate decoupling of the human body and the scene, enabling flexible recombination of human and scene elements without additional training, thereby achieving high-quality dynamic human-scene cooperative novel view synthesis.

Keywords: 3D reconstruction; natural scene; parametric model; 3D gaussian splatting; scene decoupling

1. Introduction

In recent years, computer vision and graphics have developed rapidly, and the concept of the metaverse has gained widespread attention, emerging as an influential topic in the development of the digital economy. As a crucial component of the metaverse, virtual digital humans are undoubtedly the foundation and core of this emerging field. Recent research has explored using Neural Radiance Fields (NeRF) to model 3D human avatars, typically relying on parametric body models as the structural basis for deformation during modeling [1-6].

However, in applications such as augmented reality (AR) and virtual scene live streaming, it is essential to efficiently decouple and model dynamic human bodies and

static scenes so that they can be flexibly recombined. This would enable new poses to drive human motion within different scenes. However, NeRF suffers from slow inference speeds and relies solely on implicit modeling of continuous spaces, which lacks explicit geometric structure. This makes it difficult to decouple the relationship between human bodies and scenes. Furthermore, the absence of accurate semantic segmentation between human bodies and scenes in three-dimensional space causes geometric coupling interference, making it challenging to separate dynamic human Gaussians from static scene Gaussians.

To address this challenge, this paper proposes a dynamic human–scene collaborative novel view synthesis method based on 3D Gaussian Splatting (3DGS). By standardizing the spatial coordinate systems of human bodies and scenes, the method ensures geometric consistency. A tri-plane representation is then used to reconstruct human Gaussians, and a joint training strategy is applied to train the human and scene models simultaneously. This approach enables high-quality dynamic human–scene collaborative novel view synthesis and allows for flexible human–scene recombination without additional training.

In summary, our main contributions are:

1) We propose a novel human Gaussian representation method based on tri-plane feature encoding. A multi-layer perceptron (MLP) predicts both static human Gaussian parameters and pose-dependent dynamic human Gaussian parameters from the tri-plane features. This ensures that the model can capture both local and global deformations induced by pose changes and their impact on color.
2) We propose a joint optimization strategy to separately represent and jointly optimize human and scene models.This allows the dynamic variations of human Gaussians to provide additional geometric constraints for scene Gaussians, preventing scene information loss due to occlusion. A depth loss is also introduced to constrain the joint optimization, ensuring correct occlusion relationships between the human body and the background, thereby improving the final novel view synthesis quality.
3) This method enables the fast creation and rendering of animatable human avatars and scenes from a small set (50–100 frames) of monocular videos captured in the wild. It achieves flexible human–scene recombination without requiring additional training and supports real-time rendering at 60 fps.

2. Related Work

Neural Radiance Fields (NeRF) introduced a joint representation of geometry and appearance for view synthesis using multi-view images, eliminating the need for complex capture setups [7]. Although NeRF was originally designed for static objects, recent works have extended NeRF to capture dynamic humans [1-4].

Xu et al. proposed H-NeRF, a dynamic 3D reconstruction method based on a structured implicit human model [8]. H-NeRF represents geometry using a signed distance field (SDF) and combines sparse multi-view synchronized videos with a parametric human model, significantly improving rendering sharpness and geometric integrity under complex human poses. HumanNeRF targets monocular video input, introducing a framework for joint optimization of a skeleton-driven motion field and a non-rigid motion field [3]. It first coarsely adjusts the deformation field through skeletal pose parameters, then refines local dynamics using a generic non-rigid field, enabling neural radiance field reconstruction without multi-view data. Similarly, NeuMan also targets monocular video input, using an SMPL parametric model to establish a mapping between canonical and observation spaces [2]. To enhance accuracy, it introduces an end-to-end SMPL optimization and correction network, allowing the model to learn more precise geometric information by training on geometric error estimation. Unlike the above methods that require subject-specific training, MPS-NeRF addresses the challenging task of novel view and

pose synthesis across different subjects using sparse multi-view static images [9]. It proposes a generalized NeRF framework, overcoming the reliance on single-subject training and multi-view video input.

Due to NeRF's slow training and rendering speed, recent methods have adopted 3D Gaussian splatting to represent scenes using a set of 3D Gaussians [10]. This approach significantly improves training and rendering efficiency by splatting and rasterizing Gaussians. Some methods have extended this approach to dynamic humans. Moreau et al. integrated the SMPL parametric human model with Gaussian representation, using linear blend skinning (LBS) to drive deformation of canonical Gaussian primitives initialized from SMPL [11]. Li et al. proposed a Gaussian modeling framework that combines explicit point-based representation with 2D CNNs, achieving high-fidelity digital human reconstruction through UV-space canonical Gaussian mapping [12].

Monocular video-based dynamic Gaussian human avatar reconstruction has also become a key branch of 3DGS digital human research. ParDy-Human further expands human Gaussian representation by introducing parent patch indexing and surface normal features [13]. It employs per-vertex deformation (PVD) to drive canonical Gaussians toward target poses and uses a deformation residual correction module to enhance the realism of non-rigid motion and clothing dynamics. Human101 by Li et al. focuses on balancing fast reconstruction and real-time rendering [14]. It initializes canonical Gaussian primitives from multi-view keyframe point clouds and binds Gaussians to neighboring SMPL triangles using a triangle-face rotation association mechanism, directly driving rotation and spherical harmonic coefficient updates.

While these methods have achieved remarkable progress in dynamic human avatar reconstruction, they overlook the need for simultaneously recovering both dynamic humans and static scenes from monocular video. This limits their ability to efficiently decouple human and scene representations and enable flexible recombination of dynamic human bodies and static backgrounds.

Our method builds upon the 3D Gaussian splatting framework [10]. By standardizing the spatial coordinate systems of human bodies and scenes, our approach ensures geometric consistency. A tri-plane representation is used to reconstruct human Gaussians, and a joint training strategy is employed to simultaneously optimize human and scene models. This enables high-quality dynamic human–scene collaborative novel view synthesis and allows for flexible recombination of human bodies and scenes without additional training.

3. Method

Given a monocular video containing camera motion, moving humans, and a static scene, our approach first standardizes the spatial coordinate systems of the human body and the scene to ensure geometric consistency between them. We then use a tri-plane representation to reconstruct human Gaussians and employ a joint training strategy to optimize both the human and scene models simultaneously.

3.1. Coordinate Alignment

Accurate alignment between the human body and the scene is a critical step in dynamic human–scene collaborative novel view synthesis based on 3D Gaussian Splatting (3DGS). Since human pose estimation typically operates within the camera coordinate system under an approximate orthographic projection model, coordinate alignment is necessary to ensure geometric consistency between the human body and the scene.

PnP (Perspective-n-Point) Solution: The SMPL model's 3D joint positions are obtained in the local camera coordinate system, while scene Gaussians are reconstructed in the global coordinate system using COLMAP. PnP is used to solve for the rotation and translation between the SMPL model and the COLMAP scene coordinate system.

Translation Optimization: Due to potential noise or mismatched feature points, the initial translation vector obtained from PnP may contain errors. A projection error loss function is minimized to refine the translation vector.

Scale Correction: PnP can solve for rotation and translation but cannot determine the absolute scale due to the nature of perspective projection. To resolve this, a ground plane equation is fitted using RANSAC from the scene point cloud. The scale factor is then estimated based on the intersection of the ray from the camera center and the ground plane.

By completing these steps, a final transformation matrix is obtained, enabling precise alignment between the SMPL coordinate system and the COLMAP scene coordinate system, ensuring accurate geometric consistency between the human body and the scene.

3.2. Human Gaussian Based on Tri-Plane Feature Representation

To enable animatable human Gaussians, we define a canonical space based on the SMPL human model's "Da-pose" and initialize human Gaussians using the SMPL mesh vertices. Since SMPL does not model hair and clothing, the density of human Gaussians is adaptively adjusted during training to capture these additional details. During rendering, linear blend skinning (LBS) weights are used to transform the human Gaussians from the canonical space to target poses, enabling novel view and pose synthesis.

To effectively capture complex human surface geometry and texture, we use a tri-plane feature representation. This approach organizes explicit 3D features into three orthogonal axis-aligned planes (XY, XZ, and YZ). Each plane's resolution R determines the spatial scale, while the channel count C stores feature information. The process is illustrated in Figure 1. For a human Gaussian, we project its position in the canonical space onto the three planes and use bilinear interpolation to obtain the corresponding feature vectors at those positions. These three feature vectors are then aggregated through summation to derive the final tri-plane feature vector. This vector is subsequently fed into Multi-Layer Perceptrons (MLPs) to learn the corresponding parameters of that human Gaussian.

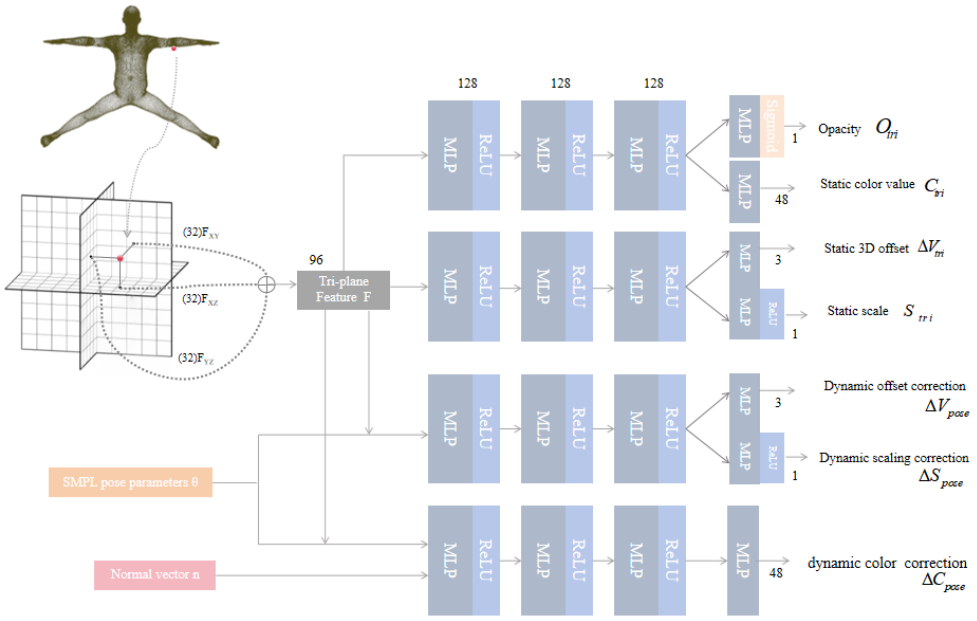


Figure 1. Tri-plane Feature Representation Structure Diagram.

Since the representation of the human body includes static (identity and environment) features and dynamic (pose-related) features, and different features have varying levels of complexity and learning difficulty, we employ four independent Multi-Layer Percep-

trons (MLPs) to separately regress the static and dynamic Gaussian parameters. This approach allows for more effective separation and modeling of identity features and pose features. To better generalize the human body to novel viewpoints, we constrain all human Gaussian primitives to be isotropic by limiting the degrees of freedom in scaling to 1 and setting the rotation to [1, 0, 0, 0].

Among these, the static color and opacity parameters are modeled by the first MLP:

$$C_{\text{tri}}, O_{\text{tri}} = \text{MLP}_1(F_i)$$

Where $C_{\text{tri}} \in \mathbb{R}^{48}$ represents the appearance color features of the human surface, primarily dependent on static environmental factors such as texture and lighting. And $O_{\text{tri}} \in \mathbb{R}^1$ uses the Sigmoid activation function to constrain the output within the physical range of [0, 1], indicating the degree from fully transparent to completely opaque. The static 3D offset and static scale are modeled by the second MLP:

$$\Delta V_{\text{tri}}, S_{\text{tri}} = \text{MLP}_2(F_i)$$

The static 3D offset $\Delta V_{\text{tri}} \in \mathbb{R}^3$ represents the center position of the human Gaussian in 3D space, capturing identity-related positional characteristics of the human body. Meanwhile, the static scale$S_{\text{tri}} \in \mathbb{R}^1$ is constrained to positive values using the ReLU activation function, ensuring that the shape of the human Gaussian remains physically plausible in 3D space.

We use the third MLP to model the dynamic offset correction and dynamic scaling correction, with inputs being the tri-plane features and the pose parameters excluding the root node:

$$\Delta V_{\text{pose}}, \Delta S_{\text{pose}} = \text{MLP}_3(F_i, \theta)$$

The dynamic offset correction $\Delta V_{\text{pose}} \in \mathbb{R}^3$is used to dynamically adjust the static position, modeling local deformations and positional changes caused by poses. The dynamic scaling correction $\Delta S_{\text{pose}} \in \mathbb{R}^1$is applied to modify the static scaling, capturing volume variations induced by different poses.

The final position and scaling are synthesized from both the static and dynamic components:

$$V_i = \mu_i + \Delta V_{\text{tri}} + \Delta V_{\text{pose}}$$
$$S_i = S_{\text{tri}} + \Delta S_{\text{pose}}$$

The dynamic color correction is modeled by the fourth MLP, with inputs including the tri-plane features, pose parameters, and normal vector information:

$$\Delta C_{\text{pose}} = \text{MLP}_4(F_i, \theta, n_i)$$

The final color is synthesized from both the static and dynamic components as follows:

$$C_i = C_{\text{tri}} + \Delta C_{\text{pose}}$$

When transforming Gaussian primitives from the canonical space to the observation space, we use k-nearest neighbor interpolation to retrieve the nearest 6 vertices from the SMPL model and compute the Linear Blend Skinning (LBS) weights for each Gaussian primitive through distance-based weighted averaging:

$$W_i = \sum_{j \in N_i} \frac{\omega_{j \to i}}{\omega_i} W_j$$

Where $\mathcal{N}_i$ represents the k nearest vertices in the SMPL mesh to the Gaussian center position, W_j denotes the LBS weight of the j vertex in the SMPL model, and $\omega_{j \to i}$ is the distance weighting term, indicating the influence of the SMPL mesh vertex on the Gaussian center. The distance weighting term is calculated using the following formula:

$$\omega_{j \to i} = exp\left(-\frac{\|p_i - v_j\|}{2\sigma^2}\right), \omega_i = \sum_{j \in \mathcal{N}(i)} \omega_{j \to i}$$

Where $\|p_i - v_j\|$denotes the Euclidean distance between the Gaussian center and the SMPL vertex.

3.3. Joint Optimization

This paper adopts a strategy of separate human–scene representation and joint optimization, simultaneously optimizing both scene Gaussians and human Gaussians (including tri-plane features and multi-layer perceptrons).

Joint loss: We first introduce a joint loss as the foundation of the joint optimization framework. This loss allows dynamic changes in the human body to provide additional geometric constraints for scene modeling, thereby improving the completeness and consistency of the synthesis results.

We combine human Gaussians with scene Gaussians and splat them onto the image plane to obtain a joint rendering result. The joint loss is computed using L1 loss, structural similarity (SSIM) loss, and perceptual (VGG) loss:

$$L_{Joint} = \lambda_1^{Joint} L_{L1}^{Joint} + \lambda_2^{Joint} L_{\text{ssim}}^{Joint} + \lambda_3^{Joint} L_{\text{vgg}}^{Joint}$$

Human body loss:Dynamic occlusion from the human body primarily affects the modeling of scene Gaussians, but it does not directly interfere with the modeling of human Gaussians. Therefore, more precise supervision can be applied separately using real images of the human region.

We introduce a human loss by comparing the rendered image of only human Gaussians against the real image containing only the human body on a plain background. The human loss is computed using L1 loss, structural similarity (SSIM) loss, and perceptual loss:

$$L_{\text{human}} = \lambda_1^{\text{human}} L_1^{\text{human}} + \lambda_2^{\text{human}} L_{\text{ssim}}^{\text{human}} + \lambda_3^{\text{human}} L_{\text{vgg}}^{\text{human}}$$

Depth Loss: In conventional Gaussian modeling, depth is generated through a weighted blending of multiple Gaussians along the ray direction. Due to the presence of multiple overlapping Gaussians, some Gaussians may appear in front of the human body or the scene because of depth errors, resulting in unrealistic occlusion relationships. Therefore, we introduce a depth loss to constrain the center positions of Gaussians, ensuring that their depth matches the depth map.

To generate the depth of the Gaussian closest to the ray direction, we apply a large opacity value to all Gaussians, thereby retaining only the most contributive Gaussian along the ray direction:

$$D(x_p) = \sum_{i \in N_r} (1-\tau)^{i-1} G_{\text{proj},i}(x_p) \|\mu_i - o\|_2^2$$

where x_p is the pixel position in the image, o is the camera center position, $N_r$represents the set of nearest Gaussians along the ray direction, $G_{\text{proj},i}$is the projection of the ith Gaussian on the image plane, and $\|\mu_i - o\|_2^2$ is the Euclidean distance from the camera to the center of the Gaussian. This amplifies the contribution of the nearest Gaussian during depth blending, reinforcing the constraints on occlusion relationships and geometric structure.

Based on this, we define the depth loss as:

$$L_{\text{depth}} = \|D(p) - \widehat{D}(p)\|_1$$

In summary, our total loss is formulated as:

$$L = \lambda_{Joint} L_{Joint} + \lambda_{\text{human}} L_{\text{human}} + \lambda_{\text{depth}} L_{\text{depth}}$$

4. Experiments

We conducted experiments using the Citron and Lab subsets from the Neuman dataset. The Citron subset is based on outdoor human motion videos, while the Lab subset is based on indoor human motion videos. Both subsets were divided into training, validation, and test sets in an 8:1:1 ratio, with specific details provided in the Table 1.

Table 1. Dataset Distribution.

Video Sequence	Total Frames / Frames	Training Set Frames / Frames	Validation Set Frames / Frames	Test Set Frames / Frames
Citron	102	30	4	3
Lab	103	82	11	10

4.1. Experimental Details

The experiments were conducted on a computer equipped with an AMD Ryzen 7 5800H processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 graphics card, running on the Ubuntu 20.04 operating system. The parameters λ_{Joint}, λ_{human}, and λ_{depth} were set to 0.6, 0.4, and 1, respectively, while λ_1^{Joint}, λ_2^{Joint}, and λ_3^{Joint} were configured as 0.3, 0.7, and 1, with the tuples(λ_1^{human}, λ_2^{human}, λ_3^{human}) and (λ_1^{Joint}, λ_2^{Joint}, λ_3^{Joint}) being identical. The ADAM optimizer was employed for network optimization, with an initial learning rate of 10^{-3} and a cosine learning rate decay strategy applied to dynamically adjust the learning rate during training. This setup ensured efficient and stable model training, leveraging the hardware capabilities and optimization techniques to achieve robust performance.

4.2. Qualitative Results

To validate the advantages of the proposed method in the task of dynamic human-scene collaborative novel view synthesis, experiments were conducted to compare our approach with the HUGS method and the Deformable-3DGS method. Both HUGS and Deformable-3DGS are capable of decoupling dynamic humans and static scenes from monocular human motion videos and performing collaborative rendering. The comparative analysis aims to demonstrate the superiority of our method in terms of rendering quality, scene consistency, and computational efficiency.

Figure 2 presents the results of different methods on our self-constructed dataset. The HUGS method proposes a scene-embedded dynamic human Gaussian representation to separate static scene Gaussians from dynamic human Gaussians. However, it only employs the SMPL model and SFM point clouds to coarsely decompose humans and scenes, failing to fully account for the mutual influence between human Gaussians and scene Gaussians. This leads to misalignment of Gaussians during collaborative rendering, such as scene Gaussians that should be behind the human appearing in front of the human Gaussians, as shown in the blue box in Figure 2(c). The Deformable-3DGS method learns a deformation field to forward-map 3D Gaussians from canonical space to observation space, achieving separation of static scenes and dynamic objects. However, due to the lack of a specialized human model (e.g., the SMPL model) as a prior, the deformation field struggles to capture high-frequency motion features of humans, resulting in poor reconstruction of dynamic humans, as illustrated in the orange box in Figure 2(d). In contrast to these methods, our approach aligns human and scene coordinates and decouples humans and scenes using tri-plane implicit representations for human Gaussians and explicit representations for static scene Gaussians. By adopting a joint optimization strategy and leveraging depth information, our method fully learns the interaction between humans and scenes, ensuring Gaussians appear in correct positions and avoiding erroneous occlusions. This ultimately achieves high-quality human-scene collaborative novel view synthesis, as demonstrated in the results shown in Figure 2 (b).

a) GT **b) Ours** **c) HUGS** **d) Deformable-3DGS.**

Figure 2. Qualitative Results.

4.3. Quantitative Results

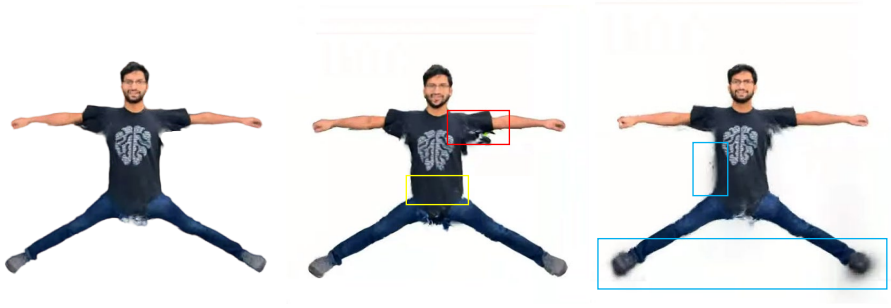
We employed three metrics — SSIM, PSNR, and LPIPS — to evaluate the results of human-scene collaborative novel view synthesis. The quantitative results on the Neuman dataset are presented in Table 2. As shown in the table, our method effectively alleviates the issues of insufficient decoupling and weak collaboration between humans and scenes. It achieves better decomposition of dynamic human Gaussians and static scene Gaussians while correcting the effects of Gaussian misalignment. Consequently, the rendered images are closer to the ground truth, demonstrating the superiority of our approach in terms of rendering quality and fidelity.

Table 2. Quantitative Results.

	Citron			**Lab**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours	25.92	0.866	0.085	26.37	0.921	0.062
HUGS	25.54	0.859	0.151	26.00	0.920	0.090
Deformable-3DGS	17.81	0.759	0.249	21.15	0.897	0.165

4.4. Ablation Experiments

In Figure 3, we present the impact of ablation studies on our method. First, we demonstrate the effect of tri-plane feature representation by directly optimizing the 3D Gaussian parameters of the human body instead of using feature tri-planes and MLPs to learn them. To deform individual Gaussians, we retrieve the nearest six vertices from the SMPL model using k-nearest neighbor interpolation and generate interpolated LBS weights through distance-based weighted averaging. We render the results of the human body in the canonical pose to showcase the ablation outcomes. The experimental results reveal that, as each Gaussian is optimized independently, the color and transparency of individual Gaussians tend to overfit the training frames, leading to color artifacts, as shown in the red box in Figure 3(b). Additionally, unnatural shrinkage occurs at the waist-leg junction of the human body, as illustrated in the yellow box in Figure 3(b).



a) Ours w/o **b) Tri-Plane Feature** **c) w/o Joint Optimization.**

Figure 3. Results of Human Canonical Pose in Ablation Experiments.

Furthermore, to validate the impact of the joint optimization strategy, we segment the human and scene using human masks and train human Gaussians with ground truth images containing only the human body (with the background set to random solid colors). Simultaneously, we train background Gaussians with ground truth images containing only the background (with the human region set to random solid colors). We also render the human body in the canonical pose to demonstrate the ablation results of this strategy. It can be observed that, due to the loss of boundary constraints from the scene on the human body, unnatural blurring appears around the human, as shown in the blue box in Figure 3(c).

Additionally, to verify the impact of this strategy on scene Gaussians, we render the human-scene collaborative synthesis results for ablation experiments. It is evident that, due to dynamic occlusion by the human body, certain scene regions cannot be fully observed from some viewpoints, leading to missing geometric and texture information in those areas. This results in floating or drifting artifacts, causing abnormal occlusion of the human body, as illustrated in the black box in Figure 4.

a) Ours **b) w/o Joint Optimization.**

Figure 4. Results of Human-Scene Collaborative Rendering in Ablation Experiments.

These ablation studies highlight the importance of tri-plane feature representation and the joint optimization strategy in achieving high-quality human-scene collaborative synthesis, ensuring both the fidelity of the human body and the consistency of the scene.

5. Conclusion

This paper addresses the challenges of geometric coupling and semantic interference between dynamic humans and static scenes by proposing a novel method for dynamic human-scene collaborative novel view synthesis based on 3D Gaussian Splatting. By standardizing the alignment between the human parametric model and the scene coordinate system and introducing a tri-plane feature representation for human Gaussian modeling, our approach enhances the disentanglement of human and scene elements while improving the robustness of human representation. Additionally, a joint human-scene optimization strategy is employed to alleviate the effects of dynamic occlusion on the scene,

thereby improving the accuracy of the synthesized views. This method effectively resolves issues such as Gaussian misalignment and incomplete scene reconstruction, achieving high-quality results in human-scene collaborative novel view synthesis.

Despite these advantages, the current method offers limited flexibility for further editing of the synthesized content, which poses a challenge for broader application scenarios. In practical settings such as virtual reality (VR) and augmented reality (AR), users may wish to modify human poses, adjust background elements, or reconstruct parts of the scene within dynamic human synthesis results. Existing methods, however, lack the adaptability to support such interactive operations. Future work could explore the integration of image- or video-based editing networks, combined with human pose parametric modeling and scene rendering controls, to enhance the interactivity and editability of synthesized content. These advancements would contribute to more versatile and user-friendly solutions, improving the method's adaptability and scalability in real-world applications.

References

1. C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges, "Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12858–12868, doi: 10.1109/CVPR52729.2023.01236.
2. W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan, "Neuman: Neural human radiance field from a single video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2022, pp. 402–418, doi: 10.1007/978-3-031-19824-3_24.
3. C. Weng, B. Curless, P. P. Srinivasan, J. T. Barron and I. Kemelmacher-Shlizerman, "HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video," in *2022 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 16189-16199, doi: 10.1109/CVPR52688.2022.01573.
4. S. Peng *et al.*, "Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 9050–9059, doi: 10.1109/CVPR46437.2021.00894.
5. A. W. Bergman, P. Kellnhofer, W. Yifan, E. R. Chan, D. B. Lindell, and G. Wetzstein, "Generative neural articulated radiance fields," *arXiv preprint arXiv:2206.14314*, 2022, doi: 10.48550/arXiv.2206.14314.
6. Z. Dong, X. Chen, J. Yang, M. J. Black, O. Hilliges, and A. Geiger, "AG3D: Learning to generate 3D avatars from 2D image collections," *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 14870–14881, doi: 10.1109/ICCV51070.2023.01370.
7. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, Eds., vol. 12346, Lecture Notes in Computer Science, Cham, Switzerland: Springer, 2020, pp. 565–580. ISBN: 9783030584511.
8. H. Xu, T. Alldieck, and C. Sminchisescu, "H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion," *arXiv e-prints*, arXiv:2110, doi: 10.48550/arXiv.2110.13746.
9. X. Gao, J. Yang, J. Kim, S. Peng, Z. Liu, and X. Tong, "MPS-NeRF: Generalizable 3D human rendering from multiview images," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2022.3205910.
10. B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph. (TOG)*, vol. 42, no. 4, pp. 1–14, 2023, doi: 10.1145/3592433.
11. A. Moreau, J. Song, H. Dhamo, R. Shaw, Y. Zhou, and E. Pérez-Pellitero, "Human Gaussian splatting: Real-time rendering of animatable avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2024, pp. 788–798, doi: 10.1109/CVPR52733.2024.00081.
12. Z. Li, Z. Zheng, L. Wang, and Y. Liu, "Animatable Gaussians: Learning pose-dependent Gaussian maps for high-fidelity human avatar modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2024, pp. 19711–19722, doi: 10.1109/CVPR52733.2024.01864.
13. H. Jung, N. Brasch, J. Song, E. Pérez-Pellitero, Y. Zhou, Z. Li, N. Navab, and B. Busam, "Deformable 3D Gaussian splatting for animatable human avatars," *arXiv preprint arXiv:2312.15059*, 2023, doi: 10.48550/arXiv.2312.15059.
14. M. Li, J. Tao, Z. Yang, and Y. Yang, "Human101: Training 100+ fps human Gaussians in 100s from 1 view," *arXiv preprint arXiv:2312.15258*, 2023, doi: 10.48550/arXiv.2312.15258.