Metaverse 2025, 6(2), 3146.
https://doi.org/10.54517/m3146

Article

Garment-aware gaussian for clothed human modeling from monocular video

Zhihao Yang[1,2], Weilong Peng[1,2,*], Meie Fang[1,2]

[1] School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510000, China
[2] Metaverse Research Institute, Guangzhou University, Guangzhou 510000, China
*** Corresponding author:** Weilong Peng, wlpeng@gzhu.edu.cn

CITATION

Yang Z, Peng W, Fang M. Garment-aware gaussian for clothed human modeling from monocular video. Metaverse. 2025; 6(2): 3146. https://doi.org/10.54517/m3146

ARTICLE INFO

Received: 9 December 2024
Accepted: 14 March 2025
Available online: 27 March 2025

Abstract: Reconstructing the human body from monocular video input presents significant challenges, including a limited field of view and difficulty in capturing non-rigid deformations, such as those associated with clothing and pose variations. These challenges often compromise motion editability and rendering quality. To address these issues, we propose a cloth-aware 3D Gaussian splatting approach that leverages the strengths of 2D convolutional neural networks (CNNs) and 3D Gaussian splatting for high-quality human body reconstruction from monocular video. Our method parameterizes 3D Gaussians anchored to a human template to generate posed position maps that capture pose-dependent non-rigid deformations. Additionally, we introduce Learnable Cloth Features, which are pixel-aligned with the posed position maps to address cloth-related deformations. By jointly modeling cloth and pose-dependent deformations, along with compact, optimizable linear blend skinning (LBS) weights, our approach significantly enhances the quality of monocular 3D human reconstructions. We also incorporate carefully designed regularization techniques for the Gaussians, improving the generalization capability of our model. Experimental results demonstrate that our method outperforms state-of-the-art techniques for animatable avatar reconstruction from monocular inputs, delivering superior performance in both reconstruction fidelity and rendering quality.

Keywords: neural rendering; 3D reconstructing; 3D Gaussian splatting; clothing human modeling; animatable body

1. Introduction

Modeling animatable avatars is a highly challenging task due to the complex nature of human movement, clothing, and the wide range of non-rigid deformations that must be accurately captured. The difficulties arise from the need to represent dynamic, time-varying scenes while ensuring that the reconstructed avatars are visually realistic, temporally consistent, and animatable under various conditions. Previous methods based on image and video reconstruction for 3D human bodies [1–3] require stringent conditions, such as a large number of viewpoints and depth maps, making them difficult to deploy in real-world applications and personal use. Meanwhile, significant progress has been made in reconstructing digital humans from monocular videos [4,5]. Existing methods for human reconstruction based on neural radiance fields (NeRF) [6] can handle simple rigid transformations but struggle with severe non-rigid motion. If the video contains limited motion diversity, these methods will exhibit poor render quality for animated human movements and novel view synthesis. Besides, coordinate-based MLP methods for NeRF rely on regressing continuous fields but suffer from the low-frequency spectral bias of MLPs [7],

resulting in suboptimal outcomes.

Recently, 3D Gaussian splatting [8], an explicit and efficient point-based representation, has been proposed to achieve both high-fidelity rendering and real-time rendering speed. However, 3D Gaussian-based human reconstruction methods are constrained by the need for multi-view inputs [9–12]. Existing single-view 3D Gaussian splatting methods, such as GoMAvatar [13] and GauHuman [14], remain suboptimal under monocular video inputs.

To address the challenges faced by monocular video inputs, we propose a garment-aware 3D Gaussian splatting avatar. Our method builds on Animatable Gaussian [10], using orthogonal projection to parameterize 3D Gaussians on a canonical template and employing StyleUNet [15] to predict pose-dependent Gaussian maps, effectively reconstructing detailed human poses with 2D CNNs and 3D Gaussian splatting, thus avoiding the low-frequency spectral bias problem that plagues coordinate-based MLP methods in NeRF. Besides, additionally, existing approaches typically focus on modeling pose-dependent deformations but fail to effectively model clothing, resulting in overly smooth human surfaces. To address this, we introduce Learnable Cloth Features, which decouple clothing from non-rigid deformations and model it independently. This design enables the clothing on the human body to be modeled separately, leading to significantly improved rendering results.

Furthermore, we observed that rendering the synthesized avatar in the driving pose requires deforming the canonical 3D Gaussians into the posed space using Linear Blend Skinning (LBS). The accuracy of this transformation is crucial for achieving high-quality rendering results. However, since the initial LBS weights are based on a smooth body template, such as SMPL, applying these initial weights after the canonical 3D Gaussians have undergone non-rigid deformations related to pose and clothing results in significant inaccuracies. To address this issue, we introduce an Optimizable LBS Weights Module. Leveraging the Tri-planes representation, we efficiently predict adaptive LBS weights that account for these deformations. Specifically, we interpolate the Tri-planes at the center locations of the deformed 3D Gaussians to extract feature vectors, which are then processed by a weight decoder to predict LBS weights dynamically. This approach allows the LBS weights to adapt to non-rigid deformations, significantly improving the accuracy of the transformation and ultimately enhancing the rendering quality of the posed avatar.

Contributions:

(1) We propose a garment-aware 3D Gaussian splatting method that combines the strengths of 2D CNNs and 3D Gaussian splatting to achieve detailed human pose reconstruction from monocular video inputs.
(2) We introduce a decouple garment and pose- dependent non-rigid deformations while incorporating compact, optimizable LBS weights, significantly enhancing the reconstruction quality of monocular 3D human models.
(3) We demonstrate the effectiveness of our method through experiments on datasets with monocular video inputs, achieving superior rendering results and reconstruction accuracy compared to existing approaches.

2. Related work

Rendering and radiance fields: Early scene reconstruction from a collection of images, such as Structure-from-Motion (SfM) [16] and Multi-View Stereo (MVS) [17], often struggled to completely reconstruct scenes and dynamic scene modeling. Volumetric representations [18–20] for novel-view synthesis using volumetric ray marching have a significant cost due to the large number of samples required to query the volume. Neural Radiance Fields (NeRFs) [6] is one of the impressive works, representing scenes as implicit functions and synthesizing realistic images from arbitrary viewpoints through volumetric rendering. NeRF follow-up works have made progress in image synthesis quality and speed [21–23]. Some approaches focus on novel view synthesis from sparse input [24,25], while others extend to dynamic scene modeling [26,27].

Despite their advances, these approaches often struggle with real-time performance due to the heavy computational demands of volumetric rendering, leading to further exploration into more efficient representations like 3D Gaussians [8]. Due to its excellent performance in photo-realistic scene rendering and speed, 3DGS has been rapidly extended for digital human reconstruction [10,13,14,28] and dynamic scene modeling [29–31].

Human representation: The work of Alldieck et al. [32,33] utilizes the human parameter template SMPL [34] to provide human priors for human modeling. Zheng et al. [35] enhance geometric accuracy by extracting 3D features from the SMPL model and pixel-level features from images. Dong et al. [2] reconstruct from RGB-D image sequences, improving geometric precision with depth information and allowing for clothing reconstruction. Saito et al. [36,37] introduce a pixel-aligned implicit function for reconstructing 3D humans, capable of handling arbitrary human images and supporting single-view inputs. Jiang et al. [38] model human geometry and clothing separately, overcoming the limitations of parametric human models in representing clothing geometry. Xiu et al. [39] enhance local features based on the SMPL-X model [40] by incorporating normal information, achieving finer-grained 3D human models. These methods still struggle to animate dynamic human motions effectively.

Although NeRF has achieved excellent results in modeling static scenes, it still faces challenges in handling dynamic scenes, especially for high degrees of freedom and non-rigid deformations in humans. Peng et al. [4] synthesize novel view images of humans from sparse camera views and use the parametric human model SMPL for dynamic modeling. Weng et al. [41] propose the Vid2Actor, which connects deformation space with canonical space based on human pose movements, but the resulting rendering quality is poor. Weng et al. [5] focus on the free-viewpoint application, by introducing a pose correction module and a non-rigid deformation module to model moving people from monocular video. Weng et al. [42] reconstruct an editable human model by training on a collection of images with different views and textures, but it cannot generate continuous human motion images. Some works [43,44] aggregate features from input and introduce an efficient human representation to achieve higher dynamic human modeling quality. Recently, 3DGS human representation enabled high-quality rendering and speed. Specifically, Animatable

Gaussians leverage powerful 2D CNNs and 3DGS to create high-fidelity avatars. D3GA [9] utilizes cage-based deformation to model the motion of 3D Gaussians. Other approaches like GART [45] and GauHuman [14], employ linear blend skinning (LBS) to model 3DGS-based animatable avatars from monocular videos but still show results of limited visual quality.

3. Methods

Figure 1 given a monocular video, we first obtain the corresponding pose using an off-the-shelf SMPL parameter estimator and initialize canonical Gaussians based on the SMPL model. We then parameterize posed but unclothed Gaussians, anchoring them to a canonical template, and project them onto the front and back views using orthogonal projection. These projections are combined with optimizable cloth features, which are pixel-aligned with the former. Next, we employ 2D CNNs to generate a deformed Gaussian map. Subsequently, we apply deformation to the canonical Gaussians to obtain the deformed Gaussians, which are then animated using optimizable LBS weights and finally rendered into realistic images.

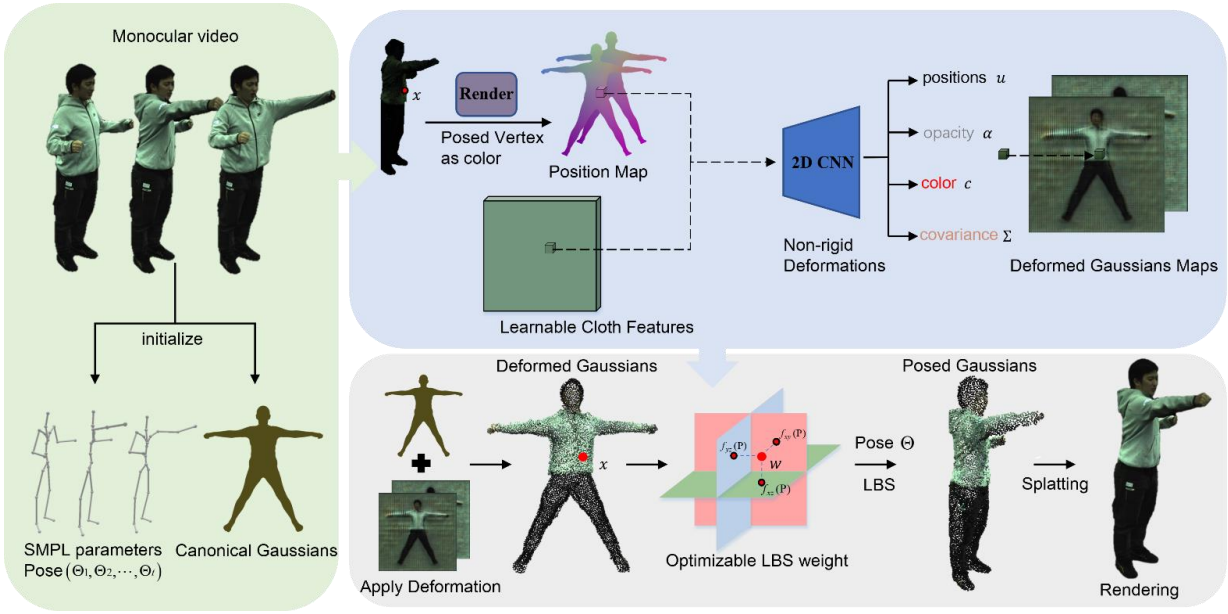

Figure 1. An overview of our methods.

In this section, given a monocular video of a human where the pose changes with each frame, we represent the canonical human body as a set of 3D Gaussians [8]. These Gaussians are defined by a full 3D covariance matrix Σ in world space and are centered at points (means) P.

$$G(x) = e^{-\frac{1}{2}(x-P)^T \Sigma^{-1}(x-P)} \tag{1}$$

To render the synthesized avatar in the driving pose, we deform the canonical 3D Gaussians into the posed space via LBS. Specifically, given a canonical 3D Gaussian, we transform its position P_c and covariance Σ^c with rotation matrix R and translation vector t.

$$P_o = R(p)P_c + t(p)\Sigma_o = R(p)\Sigma_c R^T(p) \tag{2}$$

Finally, we rendered the image using the splatting-based rasterization. The expected color *C* is computed by blending *N* ordered 2D Gaussians:

$$C = \sum_{i=1}^{N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{3}$$

where c_i is the color of each Gaussian and α_i is given by evaluating a 2D Gaussian with covariance Σ multiplied with the learned opacity.

3.1. Pose-dependent non-rigid deformation

To handle complex human movements and deformations, we decompose the motion field into two parts: the SMPL template with pose-dependent blend shapes, and the clothing. This is formalized as:

$$G_c(x, p) = G_{smpl}(x) + G_d(x, p) \tag{4}$$

where $G_{smpl}(x)$ represents the canonical human, and $G_d(x, p)$ starts from the human pose p to produce clothing and pose-dependent Gaussian deformation. However, MLPs are known to have a low-frequency bias, limiting their ability to capture high-frequency human dynamics. We follow Animatable Gaussians [10], parameterizing 3D Gaussians anchored to a canonical template onto front and back views via orthogonal projection and obtaining posed position maps $\left(P_f(p), P_b(p)\right)$. Then we use 2D CNNs to predict pose-dependent Gaussian maps based on pose conditions.

$$G_f(p), G_b(p) = F_s\left(P_f(p), P_b(p)\right) \tag{5}$$

where the posed position maps $\left(P_f(p), P_b(p)\right)$ are derived from posed but unclothed Gaussians, which contain pose-dependent information. For the $CNN(g)$, we use StyleUNet as our backbone. This technique allows us to not only reconstruct clothing details from the smooth parametric template but also model Gaussian deformations induced by different poses. For example, similar to SMPL pose blend shapes, this approach can handle expansions of the hips during specific movements [34].

3.2. Cloth-dependent non-rigid deformation

However, posed position maps only contain information related to human body poses and do not model the garment geometry. To address this, we introduce Learnable Cloth Features F_c, which are pixel-aligned with the posed position maps. This design enables us to model the clothing on the human body independently, achieving improved rendering results. Accordingly, the above equation can be reformulated as follows:

$$G_f(p), G_b(p) = F_s\left(P_f(p), P_b(p), F_c\right) \tag{6}$$

This design enables a more nuanced and independent representation of clothing, allowing for finer control and improved rendering quality. Unlike previous methods that often entangle garment details with body poses, this approach explicitly separates

the two, reducing artifacts and enhancing the realism of clothing depiction. Additionally, the pixel alignment ensures spatial consistency, which is crucial for high-quality results, and the independent modeling of garments allows for better generalization.

3.3. Optimizable LBS weights

We start with the LBS weights w^{smpl} from the SMPL model, which are initially based on a smooth body template. However, when offsets are applied to the canonical 3D Gaussians to account for deformations, using the original LBS weights to transform from the canonical space to the posed space leads to inaccuracies. To address this, we introduce an Optimizable LBS Weights Module. We represent the LBS weights as Tri-planes to reduce computational cost. Based on the center locations P of the 3D Gaussians, we interpolate the tri-planes to obtain feature vectors $\left(f_{xy}(P), f_{yz}(P), f_{xz}(P)\right)$. Specifically, we project the coordinates (*x*, *y*, *z*) onto the *xy*-planes, *yz*-planes and *xz*-planes, which are represented by feature planes, obtaining corresponding feature vectors for each projection. These feature vectors are then passed through a weight decoder, denoted as denoted as MLP_{pose}, to predict the corresponding LBS weight.

$$\widehat{w} = MLP_{pose}\left(f_{xy}(P), f_{yz}(P), f_{xz}(P)\right) \tag{7}$$

where $f_{xy}(P) \in R_x \times R_y \times D, f_{yz}(P) \in R_y \times R_z \times D,\ f_{xz}(P) \in R_x \times R_z \times D$, R_x,$R_y$and R_z are the resolutions of x, y and z axes, respectively, and *D* is the feature dimension. In our experiment, the resolution of each axis is set to 128, and the feature dimension is 32. This allows us to handle the deviations caused by these offsets and ensures more accurate deformation modeling. In practical applications, to ensure that the LBS weights approximate effective values, we predict offset weights relative to the SMPL model weights instead of learning global weights directly.

$$w = softmax\left(log(w^{smpl}) + \widehat{w}\right) \tag{8}$$

3.4. Optimization

Pose correction: Human poses are typically estimated from images, making them prone to inaccuracies. Therefore, we follow the approach of HumanNeRF [5] and introduce a pose refinement module MLP_{pose} that learns to correct the estimated poses.

$$p = p^{smpl} \otimes MLP_{pose}\left(p^{smpl}\right) \tag{9}$$

where $MLP_{pose}\left(p^{smpl}\right)$ produces a correction rotation to updated p^{smpl}.

Regularization: Under monocular input settings, the lack of constraints on non-rigid deformations leads to uneven distributions of 3D Gaussians, causing defects in the human geometry and adversely affecting the final rendering results. Inspired by Qian et al. [46], we incorporate an as-isometric-as-possible constraint to regulate the distances between points:

$$L_{aiap} = \sum_{i=1}^{N} \sum_{j \in B(j)} \left| d\left(P_c^i, P_c^j\right) - d\left(P_o^i, P_o^j\right) \right| \tag{10}$$

Considering that the density of 3D Gaussians varies across different body regions, for instance, the human face typically contains a higher density of points, we employ a ball query sampling method to ensure that the 3D Gaussians within a given area remain as isometric as possible. Here, $B(j)$ denotes the spherical neighborhood of a point *P*, and $d\left(P^i, P^j\right)$ represents the L_2 norm between points.

In addition, we apply a geometry norm regularization loss L_{geo} constrains the predicted Gaussian offsets, including both position and covariance, from becoming excessively large:

$$L_{geo} = \|\Delta(G)\|_2^2 \tag{11}$$

Thus, our overall regularization loss can be expressed as:

$$L_{reg} = L_{aiap} + L_{geo} \tag{12}$$

This formulation ensures a more uniform and stable distribution of 3D Gaussians, improving the accuracy of geometry modeling and rendering results.

Loss function: The total loss function includes the following components:

$$L = L_{color} + \lambda_1 L_{mask} + \lambda_2 L_{reg} \tag{13}$$

The color loss involves an mse loss and a perceptual loss between the rendered image C and the ground-truth image C_{GT}.

$$L_{color} = L_{mse}(C, C_{GT}) + L_{lpips}(C, C_{GT}) \tag{14}$$

L_{mask} are the L_1 loss between the accumulated volume density *M* and ground-truth subject masksM_{GT}.

$$L_{mask} = L_1(M, M_{GT}) \tag{15}$$

where λ are loss weights. Empirically, we set $\lambda_1 = 0.5$ and$\lambda_2 = 0.2$.

4. Experiment results and discussion

4.1. Evaluation dataset and metrics

We validate our method on two datasets: the ZJU-MoCap dataset [4]. For novel views synthesis evaluation, we select one camera as input, while the remaining cameras for evaluation. For novel pose synthesis. We provide qualitative results for animation on out-of-distribution poses. For quantitative experiments, we use Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) as evaluation metrics.

4.2. Baselines

We compare our method with state-of-the-art methods for human body reconstruction using monocular video, including NeuralBody [4], HumanNeRF [5], GART [45], Gauhuman [14], GoMAvatar [13]. NeuralBody and HumanNeRF are

based on NeRF (Neural Radiance Fields) reconstruction methods, while the remaining methods utilize 3D Gaussian Splatting (3DGS). We conduct experiments using the monocular setup for all methods and evaluate the results in comparison with our approach.

4.3. Comparison

Table 1 presents the quantitative comparison results of novel view synthesis between our method and the other approaches. Our method shows improvements across different metrics compared to the other methods.

Table 1. Quantitative results on the ZJU-MoCap dataset.

Method	SSIM↑	PSNR↑	LPIPS↓
NeuralBody [4]	0.9518	28.56	0.052
HumanNeRF [5]	0.9606	29.16	0.039
GoMAvatar [13]	0.9604	29.42	0.040
Gauhuman [14]	0.9600	29.25	0.045
GART [45]	0.9609	28.85	0.041
Ours	0.9610	29.52	0.035

In terms of qualitative results, as shown in **Figure 2**, our approach produces better texture details. Other methods, such as those shown in the first and third rows of the figure, exhibit blurring artifacts. Especially in terms of clothing textures, our method achieves better results compared to others.

Figure 2. Qualitative comparison for novel view synthesis to state-of-the-art.

Figure 3 compares the qualitative results of synthesizing novel poses between our method and the other approaches. Our method performs more realistically in terms

of both pose and texture.

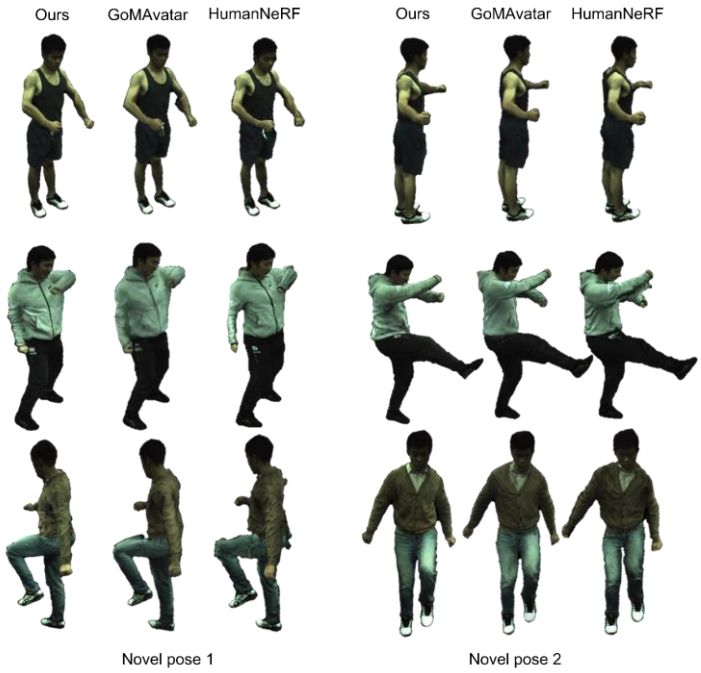

Figure 3. Novel pose qualitative results.

We also report a comparison of inference speed with NeRF-based methods and Gaussian-based methods, as shown in **Table 2**. Overall, Gaussian-based methods are faster than NeRF-based methods, such as NeuralBody and HumanNeRF. While Gaussian-based methods like Gauhuman offer fast speed, they compromise on texture quality and struggle with large deformations or fine details. In contrast, our method achieves a better balance between speed and detail with the potential for real-time performance with appropriate optimizations. The FPS values are computed on a single RTX 3090 GPU, rendering images at a resolution of 512 × 512.

Table 2. Comparison of inference speed.

NeRF-based	**FPS↑**	**3DGS-base**	**FPS↑**
NeuralBody	1.5	Ours	8
HumanNeRF	0.2	GoMAvatar	20
ARAH	0.07	Gauhuman	120

4.4. Ablation study

Through ablation studies, we validate the effectiveness of our contributions. **Table 3** presents the quantitative comparison results of novel view synthesis, highlighting the improvements between different modules and validating the effectiveness of our method.

Table 3. Ablation studies for novel view synthesis.

Method	SSIM↑	PSNR↑	LPIPS↓
Full model	0.9610	29.52	0.035
w/o Learnable cloth features	0.9578	29.06	0.037
w/o Optimizable LBS weights	0.9591	29.22	0.037
w/o Pose correction	0.9597	29.02	0.040
w/o L_{aiap}	0.9601	29.44	0.038

Figure 4 demonstrates the importance of Learnable Cloth Features. By modeling the cloth separately, our approach is better able to capture the fine texture details of human clothing.

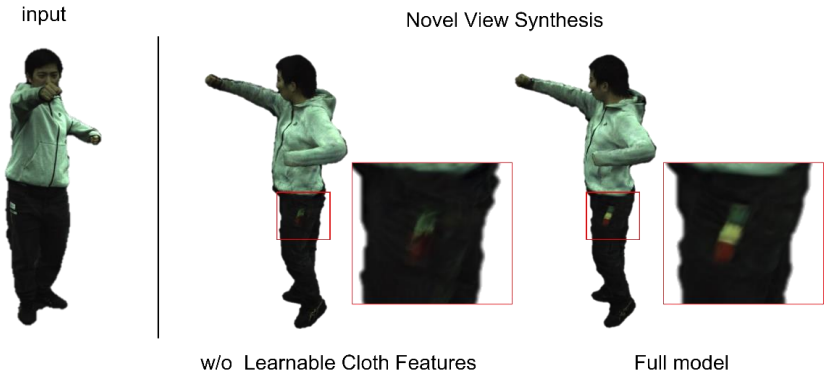

Figure 4. Optimizable cloth features improve novel view synthesis.

Due to the initial LBS weights being based on the smooth SMPL model, they cannot effectively handle the deformations of human clothing after non-rigid transformations, resulting in ghosting effects on the clothing. By using optimizable LBS weights, we can simultaneously update the LBS weights, improving the rendering results, as shown in **Figure 5**.

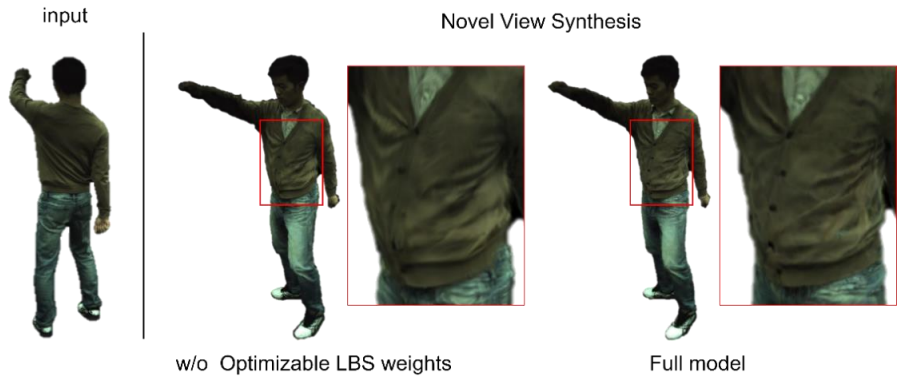

Figure 5. Optimizable LBS weights, by optimizing the Gaussian LBS weights, handle clothing deformations more effectively. For instance, the button on the clothing in the figure is more clearly visualized.

Figure 6 demonstrates how L_{aiap}, considering non-rigid deformations, it constrains the human geometry, resulting in more detailed human body features.

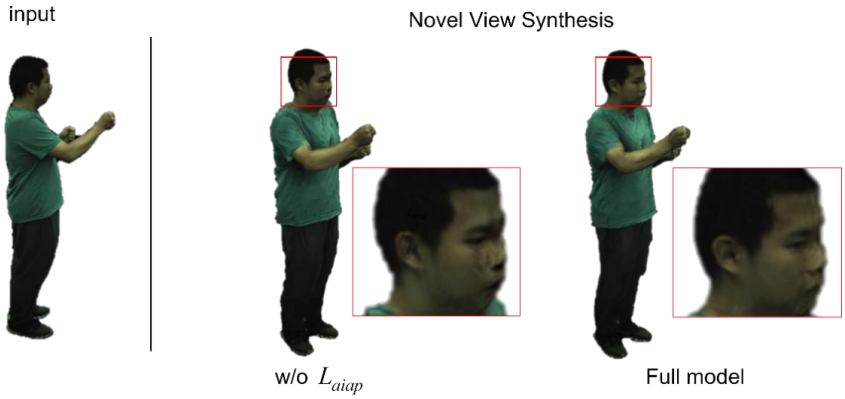

Figure 6. Without L_{aiap}, noticeable artifacts appear on the face. By maintaining the relative positions of the facial features, we can constrain the facial distortion caused by non-rigid deformations.

5. Conclusion

We present a garment-aware 3D Gaussian splatting method that effectively integrates the advantages of 2D CNNs and 3D Gaussian splatting, enabling precise clothed human pose reconstruction from monocular video inputs. By introducing a novel approach to decouple garment and pose-dependent non-rigid deformations, and incorporating compact, optimizable LBS weights, our method significantly improves the quality of monocular 3D human model reconstructions. Experimental results on datasets with monocular video inputs confirm that our approach achieves enhanced rendering and reconstruction accuracy under monocular video.

Author contributions: Conceptualization, WP and MF; methodology, ZY; software, ZY; validation, ZY, WP and MF; formal analysis, WP and MF; investigation, ZY; resources, ZY; data curation, ZY; writing—original draft preparation, ZY; writing—review and editing, WP and MF; visualization, ZY; supervision, WP and MF; project administration, YZ; funding acquisition, WP and MF. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62072126, the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012064, the Science and Technology. Program of Guangzhou under Grant SL2022A04J01112, the Fundamental Research Projects Jointly Funded by Guangzhou Council and Municipal Universities under Grant SL2023A03J00639, and Key Laboratory of Philosophy and Social Sciences in Guangdong Province of Maritime Silk Road of Guangzhou University (GD22TWCXGC15).

Conflict of interest: The authors declare no conflict of interest.

References

1. Casas D, Volino M, Collomosse J, et al. 4D video textures for interactive character appearance. Computer Graphics Forum. 2014; 33(2): 371-380. doi: 10.1111/cgf.12296
2. Dong Z, Guo C, Song J, et al. PINA: Learning a Personalized Implicit Neural Avatar from a Single RGB-D Video Sequence. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. pp. 20438-20448.
3. Guo K, Lincoln P, Davidson P, et al. The relightables. ACM Transactions on Graphics. 2019; 38(6): 1-19. doi: 10.1145/3355089.3356571
4. Peng S, Zhang Y, Xu Y, et al. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. pp. 9050-9059.
5. Weng CY, Curless B, Srinivasan PP, et al. HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. pp. 16189-16199.
6. Mildenhall B, Srinivasan PP, Tancik M, et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: Proceedings of the European Conference on Computer Vision; 2020. pp. 405–421.
7. Tancik M, Srinivasan P, Mildenhall B, et al. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020. pp. 7537–7547.
8. Kerbl B, Kopanas G, Leimkuehler T, et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics. 2023; 42(4): 1-14. doi: 10.1145/3592433
9. Zielonka W, Bagautdinov T, Saito S, et al. Drivable 3D Gaussian Avatars. arXiv. 2023; arXiv:2311.08581.
10. Li Z, Zheng Z, Wang L, et al. Animatable Gaussians: Learning Pose-Dependent Gaussian Maps for High-Fidelity Human Avatar Modeling. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 19711-19722.
11. Xu Y, Chen B, Li Z, et al. Gaussian Head Avatar: Ultra High-Fidelity Head Avatar via Dynamic Gaussians. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 1931-1941.
12. Jiang Y, Shen Z, Wang P, et al. HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19734-19745. doi: 10.1109/cvpr52733.2024.01866
13. Wen J, Zhao X, Ren Z, et al. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 2059-2069.
14. Hu S, Hu T, Liu Z. GauHuman: Articulated Gaussian Splatting from Monocular Human Videos. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 20418-20431.
15. Wang L, Zhao X, Sun J, et al. StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings; 2023. pp. 1-10.
16. Snavely N, Seitz SM, Szeliski R. Photo tourism. ACM Transactions on Graphics. 2006; 25(3): 835-846. doi: 10.1145/1141911.1141964
17. Schönberger JL, Zheng E, Frahm JM, et al. Pixelwise View Selection for Unstructured Multi-View Stereo. In: Proceedings of the European Conference on Computer Vision; 2016. pp. 501–518.
18. Li R, Tanke J, Vo M, et al. TAVA: Template-free Animatable Volumetric ActorsS. In: Proceedings of the European Conference on Computer Vision; 2022. pp. 419–436.
19. Niemeyer M, Mescheder L, Oechsle M, et al. Differentiable Volumetric Rendering: Learning Implicit 3D Representations Without 3D Supervision. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. pp. 3501-3512.
20. Varol G, Ceylan D, Russell B, et al. BodyNet: Volumetric Inference of 3D Human Body Shapes. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. pp. 20–36.

21. Garbin SJ, Kowalski M, Johnson M, et al. FastNeRF: High-Fidelity Neural Rendering at 200FPS. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021; 14326-14335.
22. Chen A, Xu Z, Geiger A, et al. TensoRF: Tensorial Radiance Fields. In: Proceedings of the European Conference on Computer Vision; 2022. pp. 333–350.
23. Fridovich-Keil S, Yu A, Tancik M, et al. Plenoxels: Radiance Fields without Neural Networks. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. pp. 5491-5500.
24. Chen A, Xu Z, Zhao F, et al. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. pp. 14104-14113.
25. Deng K, Liu A, Zhu JY, et al. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr52688.2022.01254
26. Park K, Sinha U, Hedman P, et al. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics. 2021; 40(6): 1–12.
27. Park K, Sinha U, Barron JT, et al. Nerfies: Deformable Neural Radiance Fields. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. pp. 5845-5854.
28. Hu L, Zhang H, Zhang Y, et al. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 634-644.
29. Yang Z, Gao X, Zhou W, et al. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); pp. 20331-20341.
30. Sun J, Jiao H, Li G, et al. 3DGStream: On-the-Fly Training of 3D Gaussians for Efficient Streaming of Photo-Realistic Free-Viewpoint Videos. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 20675-20685.
31. Wu G, Yi T, Fang J, et al. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 20310-20320.
32. Alldieck T, Pons-Moll G, Theobalt C, et al. Tex2Shape: Detailed Full Human Body Geometry from a Single Image. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019. pp. 2293-2303.
33. Alldieck T, Magnor M, Bhatnagar BL, et al. Learning to Reconstruct People in Clothing from a Single RGB Camera. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. pp. 1175-1186. doi: 10.1109/cvpr.2019.00127
34. Loper M, Mahmood N, Romero J, et al. SMPL: a skinned multi-person linear model. ACM Transactions on Graphics. 2015; 34(6): 1–16.
35. Zheng Z, Yu T, Liu Y, et al. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022; 44(6): 3170-3184. doi: 10.1109/tpami.2021.3050505
36. Saito S, Huang Z, Natsume R, et al. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019. pp. 2304-2314.
37. Saito S, Simon T, Saragih J, et al. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. pp. 81-90.
38. Jiang B, Zhang J, Hong Y, et al. BCNet: Learning Body and Cloth Shape from a Single Image. In: Proceedings of the European Conference on Computer Vision; 2020. pp. 18–35.
39. Xiu Y, Yang J, Tzionas D, et al. ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. pp. 13286-13296.
40. Pavlakos G, Choutas V, Ghorbani N, et al. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. pp. 10967-10977.
41. Weng C-Y, Curless B, Kemelmacher-Shlizerman I. Vid2Actor: Free-viewpoint Animatable Person Synthesis from Video in the Wild. arXiv. 2020; arXiv:2012.12884.
42. Weng CY, Srinivasan PP, Curless B, et al. PersonNeRF: Personalized Reconstruction from Photo Collections. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. pp. 524-533.

43. Noguchi A, Sun X, Lin S, et al. Neural Articulated Radiance Field. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. pp. 5742-5752.
44. Li Z, Zheng Z, Liu Y, et al. PoseVocab: Learning Joint-structured Pose Embeddings for Human Avatar Modeling. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings; 2023. pp. 1-11.
45. Lei J, Wang Y, Pavlakos G, et al. GART: Gaussian Articulated Template Models. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 19876-19887.
46. Qian Z, Wang S, Mihajlovic M, et al. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. pp. 5020-5030.