

Drivable 3D Gaussian Avatars

–Supplemental Material–

Wojciech Zielonka^{1,2,3*}, Timur Bagautdinov³, Shunsuke Saito³, Michael Zollhöfer³,
Justus Thies^{1,2}, Javier Romero³

¹Max Planck Institute for Intelligent Systems, Tübingen, Germany

²Technical University of Darmstadt ³Codec Avatars Lab, Meta

<https://zielon.github.io/d3ga/>



Figure 1. D3GA enables motion transfer showing good generalizability while preserving each avatar’s high-quality details.

1. Appendix

This supplemental document presents additional results of our method in the context of garment decomposition and the effect of \mathcal{L}_{Neo} on geometry, qualitative evaluation of garment loss $\mathcal{L}_{Garment}$. Moreover, we show the effect of the corrective field Ψ applied to the input tetrahedrons presented in Figure 6. Finally, we present more information about the deformation gradient, color network ablation in the context of shadows and additional comparison to some NeRF-based models like NPC by Su et al. [10].

Compositionality: One of the important features of our architecture is its composition properties. We can arbitrarily decompose a given avatar to give segments of interest. Each of the given segments can undergo different specialized conditioning, for instance, expression codes or keypoints for face or motion vectors for face. Figure 8 shows

decomposed garments for five different avatars. Each garment part is independent and can be controlled separately.

Regularization Effects: We introduced \mathcal{L}_{Neo} to avoid geometry artifacts that could potentially displace the Gaussians. It prevents tetrahedra from losing too much volume, flipping, or diverging in size from the canonical shape. Optimization of layered garments will naturally struggle for regions that are either permanently or temporarily covered, resulting in geometric artifacts, which can be alleviated by \mathcal{L}_{Neo} regularization (See Supp. mat for more details).

In Figure 3, we show additional ablation of the regularization effect of cage usage. As can be seen, the avatar optimized using only LBS can exhibit artifacts due to incorrect 3D Gaussian orientation during the test time. Using tetrahedral cages and MLP-based correctives can improve their orientation significantly removing the artifacts.

Failure Cases: Human body avatar methods that rely solely on sparse signals, such as joint angle vectors, often struggle to accurately model more complex garment deformation.

*Work done while Wojciech Zielonka was an intern at the Codec Avatars Lab in Pittsburgh, PA, USA



Figure 2. Our color network replaces the Spherical Harmonics used in the 3DGS with a more compact view-dependent neural network. Here we present the effect of view and pose conditioning on the shadows modeling.



w/o correctives Ours

Figure 3. Effect of corrective networks. Here we disabled corrective networks and only deformed body triangular mesh with LBS.

tions that are independent of the body. Figure 4 illustrates the most common failure in modeling long garments. All MLP-based solutions use a coarse SMPL mesh to model the avatar. To achieve high-quality results, Animatable GL (AG) [5] requires a specialized template of the garment tracked per frame. In contrast, MLP-based solutions achieve more stable, albeit incorrect, results using only the SMPL mesh as guiding geometry.

Neo-Hookean Term: Figure 5 shows the regularization effect of the Neo-Hookean term [6] to prevent tetrahedrons from sheering or losing volume, especially in places where supervision is not available, e.g., under the garment.

In table 1 we additionally measured the relation between the number of Gaussians, quality, and speed. As can be seen, the best compromise is for 100k and 200k primitives as the tradeoff between speed and quality.

2. Cage Deformation Gradient

From Sumner et al. [11]: Our goal is to encode shape deformation through a differential specification, enabling us

Experiment	PSNR \uparrow	LPIPS \downarrow	SSIM \uparrow	FPS \uparrow	Δ FPS
25k Gaussians	29.938	0.058	0.960	28	107%
100k Gaussians	29.825	0.056	0.960	26	100%
200k Gaussians	29.864	0.056	0.960	23	88%
300k Gaussians	29.864	0.056	0.960	20	77%

Table 1. Average frame rate per second at 1024×667 resolution w.r.t to the amount of Gaussian measured on a Nvidia V100 GPU. 100k Gaussians provide the best rendering-time-to-quality ratio.

to create and use an algorithm that transfers differential changes. Continuum mechanics, which addresses the behavior of materials under external forces [3], offers established methods for representing large deformations of solids under load. The deformation gradient, a key concept in this field, provides the exact representation we require.

$$\tilde{\mathbf{p}} = \mathbf{U}(\mathbf{p}) = \begin{bmatrix} U_1(p_1, p_2, p_3) \\ U_2(p_1, p_2, p_3) \\ U_3(p_1, p_2, p_3) \end{bmatrix} \quad (1)$$

The deformation of an infinitesimal vector $d\mathbf{p}$ within the solid is determined by the deformation gradient $\frac{\partial \mathbf{U}}{\partial \mathbf{p}}$. Since \mathbf{U} maps from \mathbb{R}^3 to \mathbb{R}^3 and varies with position, its gradient is a second-order tensor field:

$$\frac{\partial \mathbf{U}}{\partial \mathbf{p}} = \begin{bmatrix} \frac{\partial U_1}{\partial p_1} & \frac{\partial U_1}{\partial p_2} & \frac{\partial U_1}{\partial p_3} \\ \frac{\partial U_2}{\partial p_1} & \frac{\partial U_2}{\partial p_2} & \frac{\partial U_2}{\partial p_3} \\ \frac{\partial U_3}{\partial p_1} & \frac{\partial U_3}{\partial p_2} & \frac{\partial U_3}{\partial p_3} \end{bmatrix} \quad (2)$$

However, in a more general case, we need to use an approximation of the Jacobian \mathbf{U} via discretization by triangulation or tetrahedralization for a given shape. In our method, when using cages, we have four vertices $\{i_1, i_2, i_3, i_4\}$ of a cage for which the deformation gradient can be defined as:

$$\begin{aligned} \mathbf{J}_j(\mathbf{v}_{i_2} - \mathbf{v}_{i_1}) &= \tilde{\mathbf{v}}_{i_2} - \tilde{\mathbf{v}}_{i_1} \\ \mathbf{J}_j(\mathbf{v}_{i_3} - \mathbf{v}_{i_1}) &= \tilde{\mathbf{v}}_{i_3} - \tilde{\mathbf{v}}_{i_1} \\ \mathbf{J}_j(\mathbf{v}_{i_4} - \mathbf{v}_{i_1}) &= \tilde{\mathbf{v}}_{i_4} - \tilde{\mathbf{v}}_{i_1} \end{aligned} \quad (3)$$

which in the matrix form equals:

$$\begin{aligned} \mathbf{J}_j \mathbf{V}_j &= \tilde{\mathbf{V}}_j \\ \mathbf{J}_j &= \mathbf{V}_j \tilde{\mathbf{V}}_j^{-1} \end{aligned} \quad (4)$$

where \mathbf{V}_j and $\tilde{\mathbf{V}}_j$ the 3×3 matrices and \mathbf{J}_j if the deformation gradient applied to the kernels of each Gaussian primitive j which encapsulates change between canonical and deformed tetrahedrons.

3. Broader Impact

Our project focuses on reconstructing a high-fidelity human body avatar from multiview videos, with the capability to extrapolate to poses not originally captured. While our technology is primarily intended for constructive purposes, such

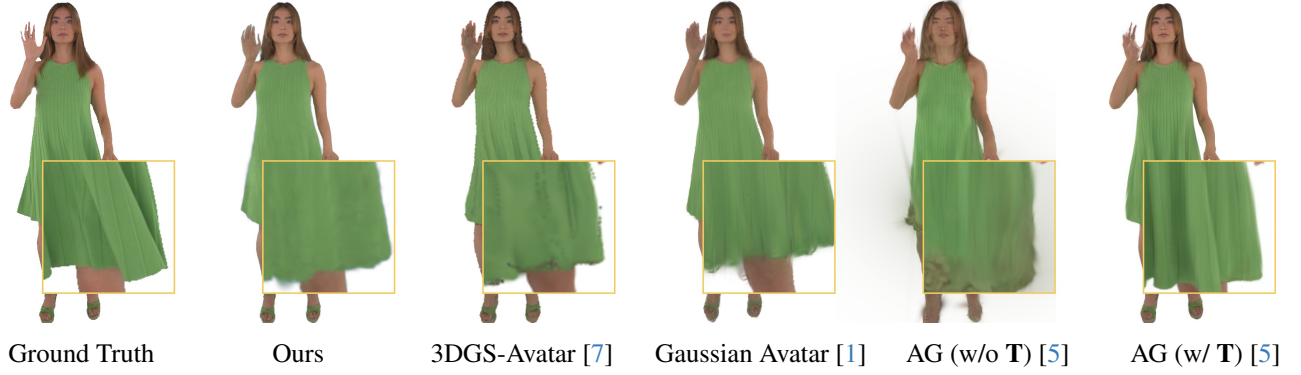


Figure 4. For pose-conditioned methods, common failure cases occur with subjects wearing long garments. Methods that do not use a specialized garment template (**T**), to which primitives are attached, often fail in these scenarios. Although Animatable Gaussians (**AG**) [5] achieves the best results when using such a template (/w **T**), it fails completely without one (w/o **T**). On the other hand, MLP-based methods

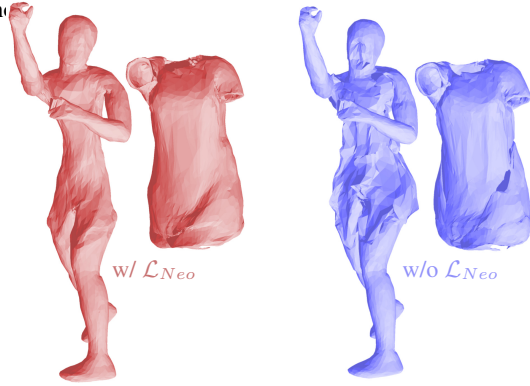


Figure 5. The effect of the tetrahedra regularization \mathcal{L}_{Neo} is mostly visible in the regions which lack supervision or undergo sliding, which covers them for most of the time.

as enhancing telepresence or mixed reality applications, we recognize the potential risks of its misuse. Hence, we advocate for advancements in digital media forensics [8, 9] to aid in detecting synthetic media. It is important to highlight that all individuals in our dataset have provided written consent for the use and release of their data.

As mentioned by Isik et al. [2] NeRF struggles with capturing long dynamic sequences due to its limited capacity. Evaluation of NPC [10] on ActorsHQ sequence shows significant artifacts depicted in Figure 9. Moreover, previous generation methods like TAVA [4] or ARAH [12] are prohibitively slow, especially for high-resolution images like in our case as they were designed to operate on the image with 256×256 size.

References

- [1] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In

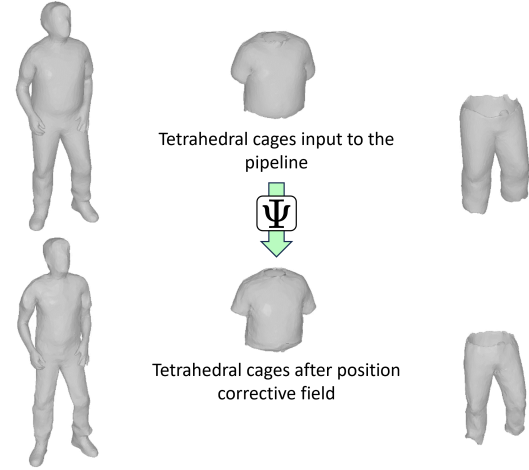


Figure 6. The effect of geometry corrective Ψ field shown on the input tetrahedral meshes before and after the pose corrective field is applied to the vertices.

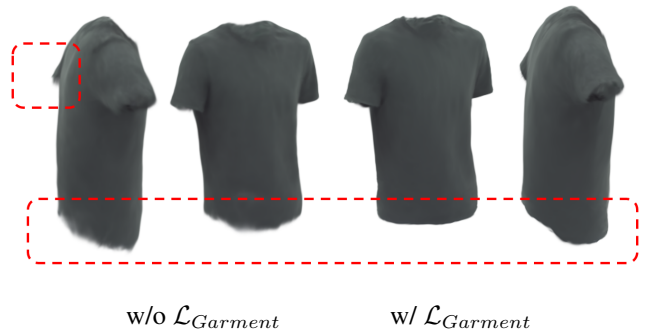


Figure 7. The additional supervision $\mathcal{L}_{Garment}$ improves the garment's shape by reducing semitransparent effects at the boundary.

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 3

- [2] Mustafa Isik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for



Figure 8. As each component of the avatar is modeled independently, it becomes straightforward to break down the avatar into individual layers. In this demonstration, we showcase the upper and lower segments of the garment. However, it’s important to note that we are not confined solely to the garment, enabling us to configure the initial layers in any desired arrangement.



Figure 9. Results of NPC [10] trained with 38 views on short and long dynamic sequences show how a NeRF representation struggles with capturing extended sequences.

Experiment	PSNR \uparrow	LPIPS \downarrow	SSIM \uparrow
Ours	29.825	0.056	0.960
w/o $\mathcal{L}_{\text{Garment}}$	30.140	0.057	0.961
Single layer	29.740	0.057	0.959

Table 2. Evaluation on our dataset; single-layer avatars incorrectly model sliding garments and garment loss improves the separation of the layers.

maps for high-fidelity human avatar modeling. *ArXiv*, abs/2311.16096, 2023. 2, 3

- [6] Miles Macklin and Matthias Müller. A constraint-based formulation of stable neo-hookean materials. *Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2021. 2
- [7] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024. 3
- [8] Andreas Rössl, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *ArXiv*, abs/1803.09179, 2018. 3
- [9] Andreas Rössl, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019. 3
- [10] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. *ArXiv*, abs/2304.02013, 2023. 1, 3, 4
- [11] Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM SIGGRAPH 2004 Papers*, 2004. 2
- [12] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdf. In *European Conference on Computer Vision*, 2022. 3

humans in motion. *ACM Trans. Graph.*, 42(4):160:1–160:12, 2023. 3

- [3] Wen C. Lai, David Rubin, and Erhard Kreml. *Introduction to Continuum Mechanics*. Pergamon Press, 3rd edition, 1993. 2
- [4] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *ArXiv*, abs/2206.08929, 2022. 3
- [5] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian