

SUPPLEMENTARY GASPACHO: GAUSSIAN SPLATTING FOR CONTROLLABLE HUMANS AND OBJECTS

Anonymous authors

Paper under double-blind review

1 OCCLUSION OF OBJECTS: WHY FEATURES?

The original 3DGS algorithm is designed to work with images of complete objects in static scenes. We observe that 3dGS doesn't converge when used to reconstruct objects under occlusion using object-only pixels. This motivates using features. They act as a regularizer and a smoothness prior for reconstruction under occlusion. This observation has been noted in prior work as well (Mihajlovic et al., 2024; Kocabas et al., 2023) which use features as a smoothness prior for 3DGS reconstruction. In our problem, even under dense camera setups, the object is occluded from a significant number of cameras. Fig.1 shows observed pixels of one object in some images from different cameras at a single timestep. In Fig. 3, we provide further illustration how features are used with 3D Gaussians object location for learning 3DGS. In Algorithm 1 we also provide pseudocode

2 COMPARISON WITH DYNAMICGAUSSIANS (LUITEN ET AL., 2024)

We want to highlight that unlike our method DynamicGaussians and follow ups are not controllable. We can control the human and object to generate renderings of new human-object interaction which is not possible with 4DGS methods which only focus on replay. To serve as a benchmark here we provide a qualitative comparison with DynamicGaussians in Fig. 5. We want to highlight that for replaying the original sequence our method produces sharper renderings than DynamicGaussians - but is much slower to train.

3 FURTHER OBJECTS

In Fig. 4, we provide further illustrations of the utility of Gaussian Maps for objects. Please zoom in. Without parametrizing Gaussians as 2D maps, for dynamic objects under occlusion sharp texture patterns are not learnt.

4 FURTHER COMPARISONS:

In Fig. 2, we provide further comparison with unmodified baselines as detailed in the paper on the BEHAVE dataset.

5 EVALUATION ON NEURAL DOME

We additionally evaluate on the Neural Dome two-sequence benchmark, which provides challenging human-object interaction scenarios. As shown in Table 1, our method consistently outperforms both AnimGaus and its modified object-aware variant. While AnimGaus+Obj (mod) improves over the vanilla baseline, it still struggles with accurate object reconstruction. In contrast, our approach achieves the best scores across all metrics. This trend follows the same pattern reported in the main paper, further validating the robustness of our method across diverse datasets.

Algorithm 1 Train 3D Gaussians from segmented multi-view images with features**Input:**

- Multi-view RGB images and corresponding object masks for several cameras.
- Initial 3D (x,y,z) locations for candidate object sites (learnable; optimized).
- A triplane feature grid with initial values (learnable).
- Small MLPs that converts a feature vector into 3D Gaussian parameters.
- Camera calibration for each view (needed for rendering).

Output: A feature grid, a set of optimized 3D Gaussian locations, and MLPs.

- Repeat for many training steps
- Pick one camera view and a small batch of 3D candidate sites
- For each site in the batch:
 - Use its current 3D position (x,y,z). Optionally apply tiny jitter for exploration.
 - Sample the feature grid at this 3D position to get a feature (tri-plane bilinear sampling).
 - Run the feature through the MLPs to get 3D Gaussian parameters: center (or small offset from the site), size, orientation, color, and opacity.
- Collect these Gaussians and render them from the chosen camera
- Produce a predicted RGB image and a predicted silhouette.
- Compare predictions to the real data inside the object mask only
 - Photometric term: average absolute RGB difference inside the mask.
 - Mask term: how well the predicted silhouette matches the mask.
- Combine these terms (with chosen weights) to get the total loss
- Backpropagate through the renderer and update
 - the feature grid values,
 - the **3D (x,y,z) site locations** (they are learnable and move during training),
 - and the MLP weights.
- Stop when validation stops improving or after a fixed number of steps
- Return the learned feature grid, the optimized 3D sites / Gaussian set, and the trained MLP



Figure 1: Occlusion Patterns due to presence of human

| Subject: Metric: | Human | | | Object | | | Full | | |
|--|--------------|---------------|--------------|--------------|---------------|--------------|--------------|---------------|--------------|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| AnimGaus (Li et al., 2024) | 26.85 | 0.8700 | 26.60 | 21.50 | 0.7581 | 34.40 | 24.99 | 0.8160 | 29.43 |
| AnimGaus + Obj (mod) (Li et al., 2024) | 27.65 | 0.8710 | 28.70 | 24.50 | 0.8431 | 33.40 | 26.11 | 0.8130 | 29.93 |
| Ours | 27.94 | 0.8741 | 25.80 | 26.39 | 0.8732 | 31.20 | 26.92 | 0.8724 | 28.10 |

Table 1: Quantitative results on the **Neural Dome (two sequences)** benchmark. Our method achieves the best performance across human, object, and full metrics. AnimGaus+Obj (mod) performs close overall but is weaker for object reconstruction, while AnimGaus is the weakest baseline.

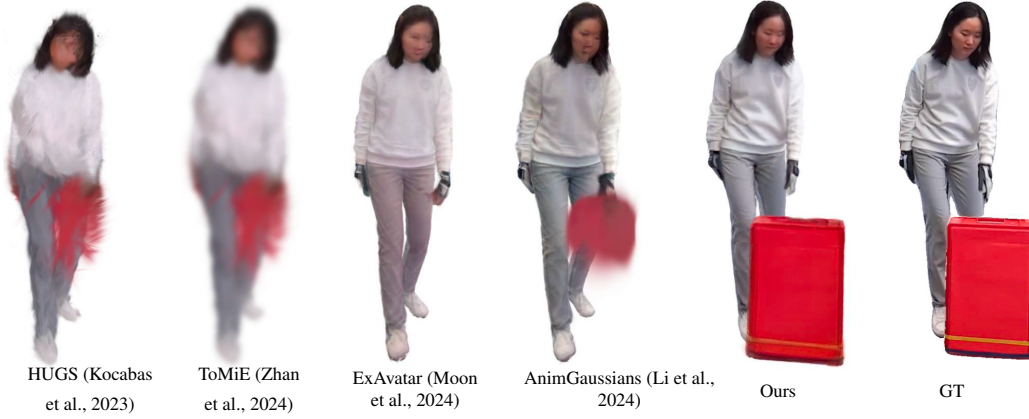


Figure 2: Quantitative comparison with unmodified existing methods for animatable human+object reconstruction on BEHAVE. Existing methods are unable to reconstruct animatable objects.

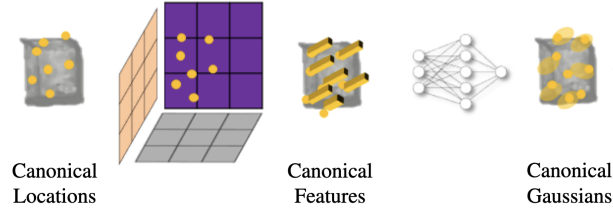


Figure 3: Learning Gaussians with features



Figure 4: Left: Learning Gaussian properties as 2D Gaussian Maps for Dynamic Objects under Occlusion; Right: Gaussian properties in 3D features space.



Figure 5: Left: Ours, Right: Dynamic Gaussians (segmented).

REFERENCES

- Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023.
- Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024.
- Marko Mihajlovic, Sergey Prokudin, Siyu Tang, Robert Maier, Federica Bogo, Tony Tung, and Edmond Boyer. SplatFields: Neural gaussian splats for sparse 3d and 4d reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.
- Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024.
- Yifan Zhan, Qingtian Zhu, Muyao Niu, Mingze Ma, Jiancheng Zhao, Zhihang Zhong, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Tomie: Towards modular growth in enhanced smpl skeleton for 3d human with animatable garments, 2024. URL <https://arxiv.org/abs/2410.08082>.