

iSegMan: Interactive Segment-and-Manipulate 3D Gaussians

Appendix

1. Details of Experimental Settings

1.1. Dataset Description

The datasets used in the experiments are described below:

- **Mip-NeRF 360** [1]. This dataset contains 9 scenes, 5 outdoors and 4 indoors, each of which contains a central object or area with a detailed background.
- **Instruct-N2N** [7]. This dataset consists of 6 scenes, each of which provides manually captured multi-view natural images, camera poses, and camera paths.
- **LERF** [9]. This dataset consists of 9 scenes, each of which provides multi-view images, camera poses, and camera paths.
- **LLFF** [11]. This dataset consists of both renderings and real images of natural scenes. The real images are 24 scenes captured by a handheld cellphone.
- **NVOS** [15]. The source data used in this dataset comes from the LLFF dataset, which contains 7 scenes and annotated segmentation masks with 8 instances (two instances are annotated in the “horn” scene). This dataset provides a 2D mask ground-truth of the target viewpoints.
- **SPIn-NeRF** [12]. This dataset contains segmentation annotations of 10 scenes, each of which provides 100 multi-view images and corresponding camera poses. For each scene, the first 40 images are the ground-truth captures without the unwanted object, and the rest of the images are the training views with the object present.

1.2. Implementation Details

All of the original 3D Gaussians in our experiments are trained utilizing the method presented in [8], with raw data from publicly available datasets, and rendered during training using the highly optimized renderer proposed in [8]. For the epipolar-guided interaction propagation, the default feature extractor for interaction matching employs DINO-small [3] with a patch size of 16. To improve the efficiency, we perform a $2\times$ downsampling operation on the input image of the feature extractor. For the visibility-based Gaussian voting, we utilize SAM [10] equipped with the ViT-Huge [5] as the interactive segmenter. The predetermined threshold of normalized votes is set to 0.8. For semantic editing, we employ Instruct-Pix2Pix [2] as the image editor and train each editing instruction for 1500-2000 steps. We prohibit Gaussian densification during the editing process. Note that ablation studies are performed on the SPIn-NeRF dataset by default. We use PyTorch for implementation and a single 32GB NVIDIA V100 GPU for all experiments.

2. Evaluation Details of Semantic Editing

User Study. The detailed evaluation criteria of user study are presented in Tab. A. We ask the participants to score from three dimensions: accuracy of instruction comprehension, rationality of editing results, and quality of editing results. The scoring criteria for each dimension are quantified on a scale of 1 to 5 inclusive, with no allowance for decimal increments. Finally, we take the average of the scores of three dimensions as the user study score and provide the 95% confidence interval. The user study results reported are the average scores of a total of 30 participants.

CLIP Directional Similarity. CLIP directional similarity [6] refers to the cosine similarity between the change of the images and captions in the CLIP [14] embedding space during the editing process. CLIP directional similarity measures the consistency of the change between images and captions. The higher the value, the more the edited image matches the editing instructions, and vice versa. The calculation is presented in Eq. (A).

$$\begin{aligned}\Delta I &= E_I(\mathcal{I}_v^e) - E_I(\mathcal{I}_v), \\ \Delta T &= E_T(t_e) - E_T(t_{ori}), \\ \text{CLIP}_{dir} &= \frac{\Delta I \cdot \Delta T}{|\Delta I| |\Delta T|},\end{aligned}\tag{A}$$

where E_I and E_T represent the image and text encoders of CLIP, respectively. \mathcal{I}_v^e represents the image rendered from the edited scene, \mathcal{I}_v represents the image rendered from the original scene. t_e represents the caption of the edited image, t_{ori} represents the original image caption. v represents the rendering viewpoint, and we compute the average metric over all viewpoints for each scene.

3. Results of Ablation Study

Epipolar Constraint. To verify the effectiveness of the epipolar constraint, we remove it and evaluate the accuracy and execution time of the region selection, cf. Tab. B. The results show that removing the epipolar constraint does produce incorrectly matched interactions due to the noise introduced by significantly increasing the search space, thus reducing accuracy.

Iterative Inspection Mechanism. To verify the effectiveness of the iterative inspection mechanism, we also remove it and evaluate the accuracy and execution time, cf. Tab. C. Since the iterative inspection mechanism only works when the target region is occluded or out of view, we select four scenes with such situations for evaluation, namely “bicycle” and “counter” from the Mip-NeRF 360 [1] dataset, and

Dimension	#Point	Description
Accuracy	1	Very poor, the system barely understands the instructions and does not match the user's intention at all.
	2	Rather poor, the understanding of the instructions is not very accurate, and there are irrelevant areas that are obviously changed.
	3	Acceptable, the understanding of the instructions is basically correct, and there are basically no irrelevant areas that are obviously changed.
	4	Fairly good, the understanding of the instructions is relatively accurate, and there are basically no irrelevant areas that have been changed, but there is still room for improvement.
	5	Very good, the system understands the instructions very accurately and there are no obvious shortcomings.
Rationality	1	Very poor, the result is very unreasonable, there is severe distortion or the original features are completely lost.
	2	Rather poor, the result is relatively unreasonable, the original features are rarely retained, and irrelevant areas are significantly distorted.
	3	Acceptable, the result is basically reasonable, the original features are basically identifiable, and the distortion in irrelevant areas is not obvious.
	4	Fairly good, the result is reasonable, the original features can be accurately identified, and there is a small amount of negligible distortion.
	5	Very good, the result is clearly reasonable, the original features are fully identifiable and there is no obvious distortion.
Quality	1	Very poor, texture detail is very blurred, color distribution anomalous.
	2	Rather poor, texture detail is blurred, color distribution is sometimes anomalous.
	3	Acceptable, texture detail is slightly blurred, color distribution is basically normal.
	4	Fairly good, texture detail is relatively clear, color distribution is normal.
	5	Very good, texture detail is very clear, color distribution is very reasonable.

Table A. The detailed evaluation criteria of the user study.

Epipolar Constraint	mIoU (%)	mAcc (%)	Execution Time	
			Feature	Segment
\times	88.7	98.5	52s	7s
\checkmark	92.4	99.1	52s	6s

Table B. Ablation on epipolar constraint.

IIM	mIoU (%)	mAcc (%)	Execution Time	
			Feature	Segment
\times	83.9	96.4	46s	5s
\checkmark	90.1	98.2	46s	6s

Table C. Ablation on iterative inspection mechanism.

Feature Extractor	mIoU(%)	mAcc(%)
DINO [3]	92.4	99.1
DINOv2 [13]	92.3	99.1
MoCov3 [4]	92.0	98.9

Table D. Ablation on feature extractor.

epipolar-guided interaction propagation, we employ different feature extractors for the ablation, *cf.* Tab. D. We employ DINO [3], DINOv2 [13] and MoCov3 [4] respectively for evaluation, and the results indicate that the proposed method is robust to the feature extractor.

“bouquet” and “figurines” from the LERF [9] dataset. We report the average results of four scenes and adopt a uniform sampling rate of 25% for each scene to maintain efficiency. The results indicate that removing the iterative inspection mechanism introduces noise matching interactions that cause incorrect 2D segmentations to participate in the voting, resulting in a decrease in accuracy.

Feature Extractor. To test the generalizability of the

4. Preliminary: 3D Gaussian Splatting

3DGS (Gaussian Splatting) [8] models a 3D scene as a set of 3D Gaussian primitives, which are initialized from the sparse point clouds obtained by Structure from Motion (SfM) [16]. Each Gaussian Θ_i is parameterized by a center point x and a covariance matrix Σ_i , which represents



Figure A. **Additional visualization results.** Orange arrows indicate interactive 3D segmentation, and blue arrows indicate semantic editing.

the distribution as:

$$\Theta_i(\mathbf{x}) = e^{-\frac{1}{2}\mathbf{x}^T \Sigma_i^{-1} \mathbf{x}}. \quad (\text{B})$$

To derive a physically meaningful covariance matrix that is necessarily positive semi-definite, the subsequent equivalent representation is employed:

$$\Sigma_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^T \mathbf{R}_i^T, \quad (\text{C})$$

where the covariance matrix Σ_i is decomposed into a scaling factor \mathbf{S}_i and a rotation quaternion \mathbf{R}_i . Moreover, an opacity σ_i is employed to control the influence of each Gaussian when blending across the scene, and a color \mathbf{c}_i is applied to represent its appearance.

To summarize, each 3D Gaussian is parameterized by a set of attributes: position $\mu_i \in \mathbb{R}^3$, scaling factor $\mathbf{S}_i \in \mathbb{R}^3$, rotation quaternion $\mathbf{R}_i \in \mathbb{R}^4$, opacity $\sigma_i \in \mathbb{R}$, and color $\mathbf{c}_i \in \mathbb{R}^k$ (where k indicates the degrees of freedom). Each 3D scene can be formally represented by a 3D Gaussian set: $\Theta = \{(\mu_i, \mathbf{S}_i, \mathbf{R}_i, \sigma_i, \mathbf{c}_i)\}_{i=1}^N$, where N indicates the number of 3D Gaussians. These 3D Gaussians can be effectively rendered to compute the color \mathbf{C} by blending N ordered Gaussians overlapping the pixel:

$$\mathbf{C} = \sum_{i \in N} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (\text{D})$$

where α_i is calculated by evaluating Θ_i with Eq. (B) multiplied by its opacity σ_i .

5. Additional Visualization Results

We present additional visualization results, cf. Fig. A. For semantic editing, we provide text editing instructions,

while for other manipulation requirements, we provide requirement descriptions and specify the tools to be invoked (marked in blue). The extensive and impressive visualization results demonstrate that our iSegMan provides precise region control and excellent manipulation performance, significantly enhancing the controllability, flexibility and practicality of existing 3D manipulation systems.

References

- [1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 1
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2
- [4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 2
- [5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1
- [6] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-

- guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. [1](#)
- [7] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. [1](#)
- [8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. [1](#), [2](#)
- [9] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. [1](#), [2](#)
- [10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. [1](#)
- [11] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. [1](#)
- [12] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinstein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. [1](#)
- [13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. [2](#)
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [1](#)
- [15] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G Schwing, and Oliver Wang. Neural volumetric object selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2022. [1](#)
- [16] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. [2](#)