# HYBRID GAUSSIANS FOR ROBUST OPEN-VOCABULARY 3D SEGMENTATION WITH MULTI-VIEW OBJECT ASSOCIATION AND BOUNDARY REFINEMENT

*Xueqi Qiu*[1]*, Yueming Sun*[1]*, Tianyu Zhang*[1]*, Yuxuan Xia*[2]*, Yang Long*[1]

[1]Department of Computer Science, Durham University, Durham, United Kingdom
[2]Shanghai Jiao Tong University, Shanghai, China



## ABSTRACT

Open-vocabulary 3D segmentation localizes objects from free-form text queries, but remains challenging in real image sequences: incomplete or noisy 2D supervision destabilizes multi-view identity assignment, while full-scene semantic learning weakens object-level discriminability. We introduce Hybrid Gaussians, a unified 3D representation jointly modeling object association and language-aligned semantics. Its Multi-View Object Association mechanism combines Observation Fusion and Semantic Contrastive Learning to improve identity consistency and semantic discrimination. Boundary Reconstruction Optimization further refines local boundary structure to improve contour quality. Experiments on LERF and 3D-OVS demonstrate strong quantitative and qualitative performance. Our method achieves 59.1% mIoU on LERF, yielding a 13.4% relative gain over the baseline. Project page: `https://nora202.github.io/hybridgaussians`.

***Index Terms*—** Open-Vocabulary Segmentation, 3D Gaussian Splatting, Object-Aware Supervision

## 1. INTRODUCTION

Open-vocabulary 3D segmentation localizes and segments objects in 3D scenes using free-form text queries rather than predefined labels. It supports embodied perception, robotic interaction, immersive media, and language-guided scene understanding [1, 2]. Existing methods lift supervision from 2D foundation models into point-based 3D representations [3, 4, 5], while NeRF [6] and 3D Gaussian Splatting (3DGS) [7] provide rendering-friendly alternatives. LERF, 3D-OVS, Feature-3DGS, LangSplat, and LEGaussians incorporate language-aligned features for text-driven querying and segmentation [8, 9, 10, 11, 12]. Object-level modeling has also been explored through identity-aware grouping, object-specific Gaussian sets, and object-centric anchors [13, 14, 15]. More recent methods combine instance and semantic cues through instance-aware grouping, identity-guided feature learning, cross-modal semantic alignment, instance-to-language mapping, or multi-view semantic aggregation [16, 17, 18, 19, 20]. Nevertheless, incomplete, noisy, or intermittent 2D cues still hinder consistent segmentation across views.

We therefore introduce Hybrid Gaussians, a unified 3D representation that equips each Gaussian with an object identity and a learnable language-aligned semantic embedding. Unlike prior methods that primarily use identity or semantic features as auxiliary attributes, our representation jointly exploits them for object-conditioned rendering and supervision. Building on this representation, Multi-view Object Association comprises two complementary components. Observation Fusion aggregates object-level evidence across views to update Gaussian identities, while Semantic Contrastive Learning combines object-conditioned semantic rendering with cross-view contrastive alignment to improve semantic discrimination. We further introduce Boundary Reconstruction Optimization, which aligns rendered gradients with masked image observations near object contours to improve boundary quality. In summary, our contributions are:

- We introduce a hybrid Gaussian formulation that jointly maintains object identities and language-aligned semantic embeddings within the same primitives, enabling object-conditioned rendering and learning.
- We address inconsistent cross-view object association and weak object-level semantic discrimination through Multi-view Object Association.
- We improve contour quality using a lightweight Boundary Reconstruction Optimization objective.
- Experiments on LERF and 3D-OVS demonstrate strong qualitative and quantitative performance.

## 2. METHOD

### 2.1. Hybrid Gaussians

Given multi-view images $\mathcal{I} = \{I_u^{gt}\}_{u\in\mathcal{U}}$, we represent the scene using $N$ Gaussian primitives following 3DGS [7]:

$$\mathcal{G} = \{g_n = (l_n, \Sigma_n, \delta_n, c_n, z_n, f_n)\}_{n=1}^{N}, \tag{1}$$

where $l_n \in \mathbb{R}^3$, $\Sigma_n \in \mathbb{R}^{3\times3}$, $\delta_n$, and $c_n \in \mathbb{R}^3$ denote the center, covariance, opacity, and color, respectively. We augment

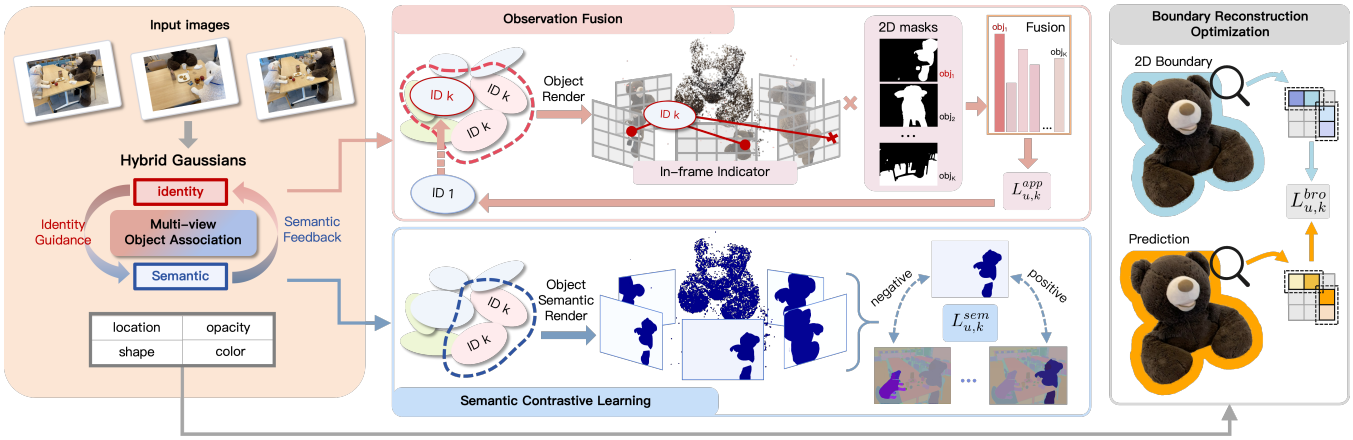


**Fig. 1**. **Training pipeline of our method.** We jointly optimize Hybrid Gaussians through Multi-view Observation Fusion for identity reassignment, Semantic Contrastive Learning for object-aware semantic discrimination, and Boundary Reconstruction Optimization for sharper object boundaries, enabling robust open-vocabulary 3D segmentation.

each Gaussian with an object assignment $z_n \in \{0, \ldots, K\}$ and a learnable language-aligned embedding $f_n \in \mathbb{R}^D$. These identity and semantic attributes form Hybrid Gaussians for object-conditioned rendering and learning.

### 2.2. Multi-view Object Association

**Observation Fusion (OF).** We consider $K$ foreground objects and a background label indexed by $k \in \{0, \ldots, K\}$. SAM [21] initializes the object masks, which are propagated across views using SAM2 [22]. After confidence filtering, the foreground mask of object $k$ in view $u$ is

$$M_{u,k} = \mathbb{I}\big[\zeta(I_u^{gt}, k) \geq \tau_1\big], \qquad k \in \{1, \ldots, K\}, \tag{2}$$

and the background is defined as $M_{u,0} = 1 - \bigvee_{k=1}^{K} M_{u,k}$.

For Gaussian $g_n$ with center $l_n$, its projected location in view $u$ is obtained as

$$\tilde{p}_{n,u} = K_u(R_u l_n + t_u), \qquad p_{n,u} = \left(\frac{\tilde{p}^x_{n,u}}{\tilde{p}^z_{n,u}}, \frac{\tilde{p}^y_{n,u}}{\tilde{p}^z_{n,u}}\right). \tag{3}$$

We define $V_{n,u} \in \{0, 1\}$ to indicate whether $g_n$ has positive depth and its projected center lies within the image domain. The multi-view support for assigning $g_n$ to object $k$ is then

$$s_{n,k} = \frac{\sum_{u \in \mathcal{U}} w_u V_{n,u} M_{u,k}(p_{n,u})}{\sum_{u \in \mathcal{U}} w_u V_{n,u} + \epsilon}, \tag{4}$$

where $w_u$ denotes the learnable normalized reliability of view $u$. For Gaussians with valid observations, the fused evidence is converted into a soft object association with temperature $T$:

$$q_{n,k} = \frac{\exp(s_{n,k}/T)}{\sum_{j=0}^{K} \exp(s_{n,j}/T)}. \tag{5}$$

The hard identity $z_n$ is assigned to the object with the highest association probability. For Gaussians without valid observations, both the soft association and hard identity are retained from the preceding iteration.

Unlike single-view hard assignment, the soft association aggregates consistent object evidence across views and reduces the influence of missing, occluded, or locally ambiguous masks. The association probability directly modulates the contribution of $g_n$ when rendering object $k$:

$$I^{pred}_{u,k}(p) = \sum_{n=1}^{N} \left\{ q_{n,k} c_n \alpha_{n,u}(p) \prod_{m<n} (1 - \alpha_{m,u}(p)) \right\}. \tag{6}$$

The object-conditioned rendering is supervised using

$$\mathcal{L}^{app}_{u,k} = (1 - \lambda)\|I^{pred}_{u,k} - I^{gt}_{u,k}\|_1 + \lambda[1 - \mathrm{SSIM}(I^{pred}_{u,k}, I^{gt}_{u,k})], \tag{7}$$

where $I^{gt}_{u,k} = I^{gt}_u \odot M_{u,k}$. This rendering objective couples multi-view object association with appearance reconstruction, allowing the soft associations to be progressively refined during optimization and the corresponding hard identities $z_n$ to be updated accordingly.

**Semantic Contrastive Learning (SCL).** Although Observation Fusion improves object identity consistency, directly lifting dense 2D semantic features into 3D may still introduce cross-object feature contamination [16]. We therefore use the fused object associations to perform object-conditioned semantic contrastive learning.

For view $u$, we extract a dense language-aligned reference map from the spatial features of a frozen CLIP image encoder $\Phi_i(\cdot)$ [23]:

$$S^{ref}_u = \mathrm{Up}\big(\Phi^{\mathrm{dense}}_i(I^{gt}_u)\big) \in \mathbb{R}^{D \times H \times W}, \tag{8}$$

$\mathrm{Up}(\cdot)$ upsamples the patch-level features to the image resolution. Using the soft object association $q_{n,k}$ obtained from Observation Fusion, we render an object-conditioned semantic map:

$$S^{pred}_{u,k}(p) = \sum_{n=1}^{N} \left\{ q_{n,k} f_n \alpha_{n,u}(p) \prod_{m<n} (1 - \alpha_{m,u}(p)) \right\}. \tag{9}$$

where the semantic embeddings are composited using the standard 3DGS alpha-blending weights.

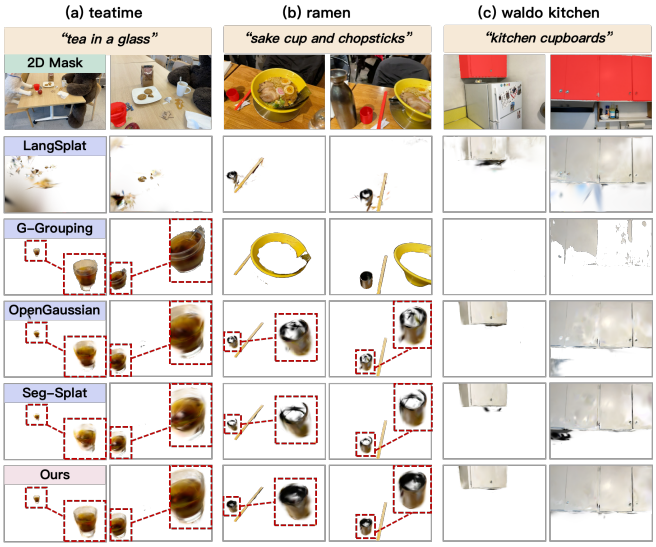


**Fig. 2. Qualitative comparison.** For each text query, we visualize the corresponding object segmentation results from different views. Gaussian Grouping [13] and Segment-then-Splat [14] are abbreviated as G-Grouping and Seg-Splat.

To remove view-dependent spatial layouts, we aggregate the semantic features within the propagated object mask:

$$z_{u,k}^{ref} = \frac{\sum_p M_{u,k}(p) S_u^{ref}(p)}{\sum_p M_{u,k}(p)+\epsilon},\ z_{u,k}^{pred} = \frac{\sum_p M_{u,k}(p) S_{u,k}^{pred}(p)}{\sum_p M_{u,k}(p)+\epsilon}. \tag{10}$$

We further apply $\ell_2$ normalization to obtain $\bar{z}_{u,k}^{ref}$ and $\bar{z}_{u,k}^{pred}$. For an anchor $\bar{z}_{u,k}^{pred}$, the reference embedding of the same object in another valid view $v \neq u$ forms a positive pair, whereas the embeddings of other foreground objects form negatives. The semantic contrastive objective is

$$\mathcal{L}_{u,k}^{sem} = -\log \frac{\exp\left(\mathrm{sim}(\bar{z}_{u,k}^{pred}, \bar{z}_{v,k}^{ref})/\tau_2\right)}{\sum_{k'=1}^{K} \exp\left(\mathrm{sim}(\bar{z}_{u,k}^{pred}, \bar{z}_{v,k'}^{ref})/\tau_2\right)}, \tag{11}$$

$\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and $\tau_2$ is the temperature.

Observation Fusion maintains consistent object identities for Gaussians across views, while Semantic Contrastive Learning improves intra-object consistency and inter-object separability. Together, they form Multi-view Object Association, enabling stable object identities and discriminative language-aligned representations.

### 2.3. Boundary Reconstruction Optimization (BRO)

Although $\mathcal{L}_{u,k}^{app}$ and $\mathcal{L}_{u,k}^{sem}$ provide region-level appearance and semantic supervision, they do not explicitly emphasize object contours, where 2D-to-3D lifting errors frequently cause cross-object semantic ambiguity and geometric misalignment [26, 27]. We therefore introduce BRO to selectively enhance reconstruction within object-boundary neighborhoods. We first render the soft object opacity map

$$A_{u,k}^{pred}(p) = \sum_{n=1}^{N} \left\{ q_{n,k}\alpha_{n,u}(p) \prod_{m<n} (1-\alpha_{m,u}(p)) \right\}. \tag{12}$$

Boundary bands are extracted from the predicted opacity and propagated object mask using the morphological gradient:

$$B_{u,k}^{pred} = \mathcal{D}(A_{u,k}^{pred}) - \mathcal{E}(A_{u,k}^{pred}),\ B_{u,k}^{gt} = \mathcal{D}(M_{u,k}) - \mathcal{E}(M_{u,k}), \tag{13}$$

where $\mathcal{D}$ and $\mathcal{E}$ denote dilation and erosion, respectively. We combine the two boundary bands as

$$B_{u,k} = \mathrm{clip}\left(B_{u,k}^{pred} + B_{u,k}^{gt}, 0, 1\right) \tag{14}$$

to cover both the target contour and potentially misaligned predicted boundaries.

Let $\Delta_x$ and $\Delta_y$ denote the horizontal and vertical forward finite-difference operators. The corresponding directional boundary weights are obtained by aligning $B_{u,k}$ with the spatial support of each finite difference and are denoted by $W_{u,k}^x$ and $W_{u,k}^y$. The BRO loss is defined as

$$\mathcal{L}_{u,k}^{bro} = \frac{\sum_{d\in\{x,y\}} \left\| W_{u,k}^d \odot \left(\Delta_d I_{u,k}^{pred} - \Delta_d I_{u,k}^{gt}\right)\right\|_1}{C \sum_{d\in\{x,y\}} \|W_{u,k}^d\|_1 + \epsilon}, \tag{15}$$

**Table 1. Quantitative comparison on LERF and 3D-OVS in terms of average mIoU (%).** The best and second-best results are highlighted in bold and underlined, respectively.

| Methods | LERF | 3D-OVS |
|---|---|---|
| LERF [8] | 37.4 | 54.8 |
| LEGaussians [12] | 46.6 | 53.5 |
| G-Grouping [13] | 29.6 | 89.0 |
| 3D-OVS [9] | 42.2 | 86.8 |
| OpenGaussian [16] | 38.4 | – |
| LangSplat [11] | 51.4 | 93.4 |
| Feature-3DGS [10] | – | 87.8 |
| Laser [24] | 53.0 | 89.3 |
| SAGA [25] | – | <u>96.0</u> |
| Seg-Splat [14] | <u>52.1</u> | 91.3 |
| COS3D [20] | 50.8 | – |
| Ours | **59.1** | **96.1** |

**Table 2. Efficiency comparison on LERF and 3D-OVS.** Results are reported as training time (ms/iter) / inference time (ms/iter) / peak memory (MB). Lower values are better.

| Methods | Train / Infer / Memory | |
|---|---|---|
| | LERF | 3D-OVS |
| G-Grouping [13] | 105.85 / 252.88 / 10037.63 | 60.18 / 173.94 / 3603.59 |
| OpenGaussian [16] | 84.77 / **125.91** / **2099.80** | 46.25 / <u>66.84</u> / **1400.16** |
| LangSplat [11] | 89.71 / <u>135.39</u> / <u>2546.68</u> | 51.60 / **66.42** / <u>1433.34</u> |
| Seg-Splat [14] | **65.48** / 135.46 / 2744.55 | **28.66** / 71.78 / 1582.40 |
| Ours | <u>75.99</u> / 278.56 / 3721.02 | <u>38.98</u> / 108.92 / 2903.34 |

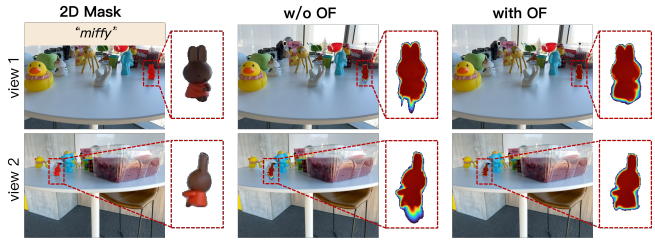


**Fig. 3**. **Ablation study of OF.** We visualize the depth maps of the queried objects across multiple views.

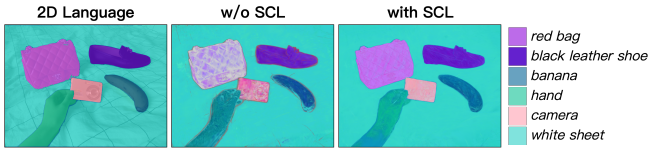


**Fig. 4**. **Ablation study of SCL.** We compare the rendered 2D language maps without and with SCL against the 2D language supervision.

where $C$ is the number of image channels and $\epsilon$ ensures numerical stability. By restricting gradient reconstruction to a narrow contour neighborhood, BRO encourages sharper and more accurately localized object transitions without redundantly constraining homogeneous interior regions. The overall training objective is

$$\mathcal{L} = \lambda_{\text{app}}\mathcal{L}_{\text{app}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{bro}}\mathcal{L}_{\text{bro}}, \qquad (16)$$

where each loss denotes the average of its corresponding per-view, per-object objective in Equations (7), (11) and (15).

## 3. EXPERIMENTS

### 3.1. Implementation Details

We evaluate our method on LERF [8] and 3D-OVS [9]. LERF contains complex indoor scenes and emphasizes fine-grained object localization, while 3D-OVS provides language-driven annotations for indoor scenes. 2D object masks are initialized using SAM [21] and propagated across views using SAM2 [22]. Image and text semantic features are extracted using OpenCLIP ViT-B/16 pretrained on LAION-2B [28]. All experiments are conducted on an NVIDIA RTX 4090 GPU.

### 3.2. Results and Analysis

As shown in Figure 2, our method produces more complete object regions and cleaner boundaries across different scenes and views. For the fine-grained queries "tea in a glass" and "sake cup and chopsticks", our method better preserves the queried objects while reducing interference from surrounding regions. In Table 1, our method achieves the best average mIoU on both benchmarks, reaching 59.1% on LERF and 96.1% on 3D-OVS. This exceeds the second-best results by 7.0 and 0.1 points, respectively.

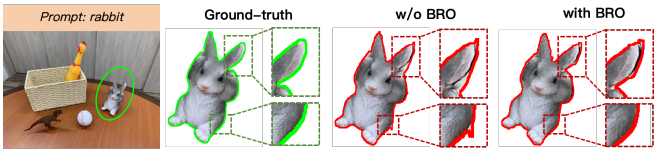


**Fig. 5**. **Ablation study of BRO**. We highlight the extracted boundary bands.

**Table 3**. **Ablation study of the proposed components on LERF and 3D-OVS (%).**

| OF | SCL | BRO | mAcc / mIoU / mBIoU LERF | mAcc / mIoU / mBIoU 3D-OVS |
|---|---|---|---|---|
| – | – | – | 78.01 / 52.15 / 56.52 | 97.45 / 91.30 / 80.18 |
| ✓ | – | – | 79.53 / 53.03 / 56.77 | 97.95 / 92.30 / 81.99 |
| – | ✓ | – | 79.59 / 53.74 / 55.45 | 97.87 / 90.41 / 80.09 |
| – | – | ✓ | 82.22 / 53.42 / 58.53 | 97.28 / 91.67 / 82.25 |
| ✓ | ✓ | – | 80.04 / 56.83 / 57.23 | 98.01 / 92.68 / 82.08 |
| ✓ | – | ✓ | 81.13 / 57.56 / 59.22 | 98.15 / 94.00 / 85.65 |
| – | ✓ | ✓ | 83.11 / 57.77 / 59.14 | 98.08 / 95.04 / 85.20 |
| ✓ | ✓ | ✓ | **84.36 / 59.13 / 60.54** | **98.59 / 96.12 / 88.27** |

Regarding efficiency, Table 2 shows that our method achieves the second-fastest training speed on both datasets. Compared with OpenGaussian, the training time is reduced by 10.4% and 15.7%, respectively. Although our method introduces additional inference and memory costs, it offers a practical trade-off between computational overhead and segmentation performance.

### 3.3. Ablation Studies

Table 3 evaluates the contributions of OF, SCL, and BRO on LERF and 3D-OVS. OF consistently improves performance across both benchmarks, while BRO notably enhances boundary quality. Although SCL alone yields mixed results, it provides clear complementary benefits when combined with the other components. Integrating all three modules achieves the best performance across all metrics, reaching 59.13% and 96.12% mIoU and 60.54% and 88.27% mBIoU on LERF and 3D-OVS, respectively. Qualitative ablations in Figures 3 to 5 further illustrate the effects of OF, SCL, and BRO on cross-view consistency, semantic discrimination, and boundary refinement, respectively.

## 4. CONCLUSION

In this paper, we presented a robust open-vocabulary 3D segmentation framework based on Hybrid Gaussians with Multi-View Object Association. Hybrid Gaussians provide a unified representation for object-conditioned rendering and learning, Observation Fusion and Semantic Contrastive Learning improve identity consistency and semantic discriminability, and Boundary Reconstruction Optimization further refines contour quality and local spatial precision.

## 5. REFERENCES


[1] Zhijie Yan, Shufei Li, Zuoxu Wang, Lixiu Wu, Han Wang, Jun Zhu, Lijiang Chen, and Jihong Liu, "Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation," *IEEE Robotics and Automation Letters*, 2025.

[2] Shuting He, Peilin Ji, Yitong Yang, Changshuo Wang, Jiayi Ji, Yinglin Wang, and Henghui Ding, "A survey on 3d gaussian splatting applications: Segmentation, editing, and generation," *PAMI*, 2026.

[3] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al., "Openscene: 3d scene understanding with open vocabularies," in *CVPR*, 2023, pp. 815–824.

[4] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann, "OpenMask3D: Open-Vocabulary 3D Instance Segmentation," in *NeurIPS*, 2023.

[5] Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le, "Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation," in *ICRA*. IEEE, 2024, pp. 9411–9417.

[6] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[8] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik, "Lerf: Language embedded radiance fields," in *CVPR*, 2023, pp. 19729–19739.

[9] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu, "Weakly supervised 3d open-vocabulary segmentation," *NeurIPS*, vol. 36, pp. 53433–53456, 2023.

[10] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in *CVPR*, 2024, pp. 21676–21685.

[11] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister, "Langsplat: 3d language gaussian splatting," in *CVPR*, 2024, pp. 20051–20060.

[12] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan, "Language embedded 3d gaussians for open-vocabulary scene understanding," in *CVPR*, 2024, pp. 5333–5343.

[13] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *ECCV*. Springer, 2024, pp. 162–179.

[14] Yiren Lu, Yunlai Zhou, Yiran Qiao, Chaoda Song, Tuo Liang, Jing Ma, Huan Wang, and Yu Yin, "Segment then splat: Unified 3d open-vocabulary segmentation via gaussian splatting," *NeurIPS*, 2025.

[15] Ruijie Zhu, Mulin Yu, Linning Xu, Lihan Jiang, Yixuan Li, Tianzhu Zhang, Jiangmiao Pang, and Bo Dai, "Objectgs: Object-aware scene reconstruction and scene understanding via gaussian splatting," in *ICCV*. IEEE, 2025, pp. 8350–8360.

[16] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al., "Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding," *NeurIPS*, vol. 37, pp. 19114–19138, 2024.

[17] SungMin Jang and Wonjun Kim, "Identity-aware language gaussian splatting for open-vocabulary 3d semantic segmentation," in *ICCV*, 2025, pp. 20467–20476.

[18] Changzhou Li, Xinyu Yang, Weiguo Yang, and Xinyi Li, "Vaflangsplat: Voxel-aware fusion language gaussian splatting," in *Proceedings of the 33rd ACMMM*, 2025, pp. 4952–4961.

[19] Jiazhong Cen, Xudong Zhou, Jiemin Fang, Changsong Wen, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian, "Tackling view-dependent semantics in 3d language gaussian splatting," in *ICML*, 2025.

[20] Runsong Zhu, Ka-Hei Hui, Zhengzhe Liu, Qianyi Wu, Weiliang Tang, Shi Qiu, Pheng-Ann Heng, and Chi-Wing Fu, "Cos3d: Collaborative open-vocabulary 3d segmentation," *NeurIPS*, 2025.

[21] 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 *ICCV*, 2023, pp. 4015–4026.

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al., "Sam 2: Segment anything in images and videos," in *ICLR*, 2025, vol. 2025, pp. 28085–28128.

[23] 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 *ICML*. PmLR, 2021, pp. 8748–8763.

[24] Xingyu Miao, Haoran Duan, Yang Bai, Tejal Shah, Jun Song, Yang Long, Rajiv Ranjan, and Ling Shao, "Laser: Efficient language-guided segmentation in neural radiance fields," *PAMI*, 2025.

[25] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian, "Segment any 3d gaussians," in *AAAI*, 2025, vol. 39, pp. 1971–1979.

[26] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian, "Multi-class part parsing with joint boundary-semantic awareness," in *ICCV*, 2019, pp. 9177–9186.

[27] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov, "Boundary iou: Improving object-centric image segmentation evaluation," in *CVPR*, 2021, pp. 15334–15342.

[28] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023, pp. 2818–2829.