ProFuse: Efficient Open-Vocabulary 3D Gaussian Splatting with Early-Saturating Semantic Uplifting

Yen-Jen Chiou
National Yang Ming Chiao Tung University

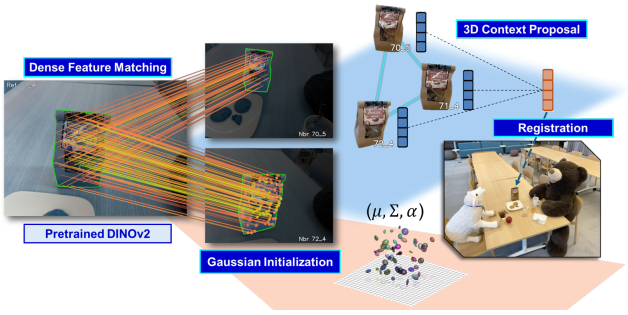

Figure 1. Overview of **ProFuse**. **Left:** A dense matcher supplies cross-view geometric and semantic correspondences. **Top:** Warped masks are grouped into 3D Context Proposals with a shared global feature. **Bottom:** Triangulated matches initialize a compact Gaussian scene, and proposal features are fused without render supervision for coherent open-vocabulary 3D semantics.

Abstract

We present ProFuse, a resource-efficient framework for open-vocabulary 3D understanding with 3D Gaussian Splatting. ProFuse uses dense multi-view correspondences to initialize a compact Gaussian scene without densification while simultaneously linking per-view masks into 3D Context Proposals. Each proposal aggregates a global feature from its member masks, and global features are attached to Gaussians through visibility-weighted accumulation along camera rays, producing coherent per-primitive semantics without render-supervised training or gradient-based optimization. This correspondence-guided design reduces scene preparation cost, yields a smaller Gaussian set for downstream querying, and reaches stable accuracy at shallow uplifting depth, which eliminates the need to sweep for an ideal depth across scenes during feature uplifting. Experiments on LERF and ScanNet show that ProFuse reduces offline deployment cost and accelerates online querying efficiency, while improving retrieval accuracy over prior 3D Gaussian baselines.

1. Introduction

Open-vocabulary 3D scene understanding seeks a representation that can localize and retrieve scene content from free-form language queries, enabling downstream uses such as robotics, navigation, and AR [2, 7, 20, 28, 29, 31]. Recent progress has made 3D Gaussian Splatting an attractive substrate for this problem because it provides an explicit scene representation together with efficient rendering [6, 8, 10, 11, 15–17, 22, 23, 25, 28, 30]. Recent approaches based on direct 3D semantic uplifting [9, 14, 15, 19, 27] further improve practicality by attaching language descriptors to scene primitives and answering queries in 3D space without a render-supervised semantic training loop [5, 8, 17, 18, 24, 30, 32]. In deployment, retrieval accuracy alone is insufficient. The cost of offline semantic scene preparation, the size of the resulting Gaussian scene, and the efficiency of online querying all matter when the system is used as a queryable 3D semantic memory.

Existing work such as Dr. Splat [9] offers an efficient alternative to render-supervised distillation through feature registration. Its semantic uplifting stage depends on the number of Gaussian contributors accumulated along each ray, and a larger accumulation depth increases the processing cost. Dense Gaussian scenes also increase the overhead of downstream retrieval and projection. In practice, stable performance may require a larger accumulation budget than is desirable, and deployment can involve repeated tuning of this depth across scenes. These limitations make efficiency, scalability, and predictable query cost central issues for open-vocabulary 3D systems.

We present ProFuse, a resource-efficient framework for open-vocabulary understanding with 3D Gaussian Splatting. ProFuse introduces a dense correspondence-guided pre-registration stage that serves two roles. Dense multi-view correspondences [1, 3, 4, 12, 13, 21, 26] initialize a compact Gaussian scene without iterative densification and associate per-view masks across images into object-level 3D Context Proposals. Each proposal aggregates a global feature from its member masks, and these features are attached to Gaussians through visibility-weighted accu-

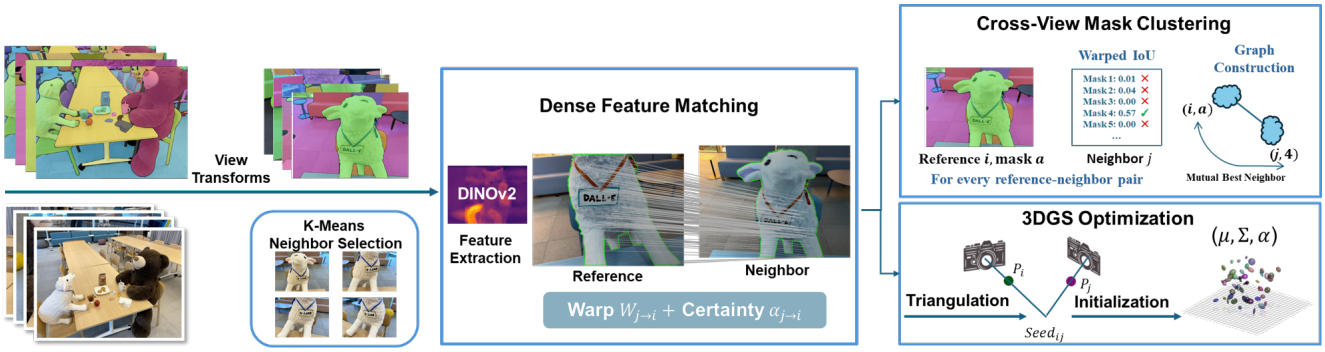

Figure 2. **Pre-registration.** For each reference view we select K neighbors via view clustering, dense matching then obtain per-pixel warps $W_{j\to i}$ and confidences $\alpha_{j\to i}$. **Bottom right:** Given the warps of a *pixel pair*, we triangulate a 3D seed point for Gaussian initialization. **Top right:** Warped IoU comparison on every reference–neighbor *mask pair*; masks that pass the selection form edges of a bipartite graph.

mulation along camera rays. The result is a coherent per-primitive semantic field obtained without render-supervised fine-tuning or gradient-based semantic optimization. This design strengthens cross-view semantic consistency while concentrating useful semantic mass on the leading contributors of each ray.

This operating regime improves efficiency at both stages of deployment. During offline semantic scene preparation, ProFuse reaches stable accuracy at shallow uplifting depth, which sharply reduces the need to sweep the Top-K accumulation depth across scenes. During online querying, the resulting scene remains substantially smaller, which lowers the cost of downstream retrieval and projection. Experiments on open-vocabulary 3D object selection and point cloud understanding show that ProFuse improves retrieval accuracy over prior 3D Gaussian baselines while reducing feature uplifting overhead, scene size, and query time. Our contributions are summarized as follows:

- We introduce a correspondence-guided pre-registration framework that jointly builds a compact Gaussian scene and cross-view 3D Context Proposals.
- We develop a proposal-level semantic uplifting strategy that yields coherent per-Gaussian semantics and reaches stable performance at shallow uplifting depth.
- We show that this design improves open-vocabulary retrieval accuracy together with scene compactness, offline deployment efficiency, and online query efficiency on standard 3D benchmarks.

2. Method

Given posed RGB views $\{I_i\}_{i=1}^N$ with known intrinsics and extrinsics, we construct a 3D Gaussian scene whose primitives store language-aligned descriptors and can be queried directly in 3D. The pipeline begins with *Dense Correspondence Pre-registration*, which initializes a Gaussian scene and forms *3D Context Proposals* that group masks across views. A *Feature Registration* stage then computes proposal-level global features and assigns a unit-normalized language descriptor to each Gaussian using visibility-based weights along camera rays, enabling direct 3D querying.

2.1. Dense Correspondence Pre-registration

We start from posed RGB images $\{I_i\}_{i=1}^N$ with known intrinsics and extrinsics. For each view I_i, SAM produces non-overlapping masks $\{M_i^k\}$, and CLIP encodes each mask region into a language-aligned feature $f_i^k \in \mathbb{R}^D$, forming $\mathcal{S}_i = \{(M_i^k, f_i^k)\}$.

Dense Feature Matching. Given two views (I_i, I_j), a dense matcher returns a warp and confidence map $C(I_i, I_j) \to (W_{j\to i}, \alpha_{j\to i})$. We retain correspondences with confidence above τ_α. The resulting correspondence fields are used to initialize the Gaussian seeds and to establish cross-view evidence for 3D Context Proposals.

Gaussian Initialization. For a confident match between (u_j, v_j) in view j and its mapped location (u_i, v_i) in view i, we back-project both rays and triangulate their intersection to seed a Gaussian center. Repeating this yields an initialized Gaussian set $\mathcal{G}_0 = \{g_n\}$, which is later refined through pruning without iterative densification.

3D Context Proposals. 3D Context Proposals are formed by grouping per-view masks that mutually support one another under dense correspondence into stable multi-view units. We build an undirected graph whose nodes are masks (i, k). For a candidate pair (M_i^a, M_j^b), we warp M_j^b into view i with $W_{j\to i}$ and evaluate agreement only on reliable correspondences using the confidence gate $\Gamma_{j\to i} = [\alpha_{j\to i} \geq \tau_\alpha] \wedge \texttt{vis\_mask}$. An edge is added only when the confidence-gated mask overlap and the bounding-box overlap both exceed their thresholds in *both* directions $(i \leftrightarrow j)$, which suppresses accidental matches and viewpoint-specific artifacts. Connected components define proposals $\mathcal{P} = \{P_m\}$, and small components are filtered by minimal size and minimal distinct-view support.

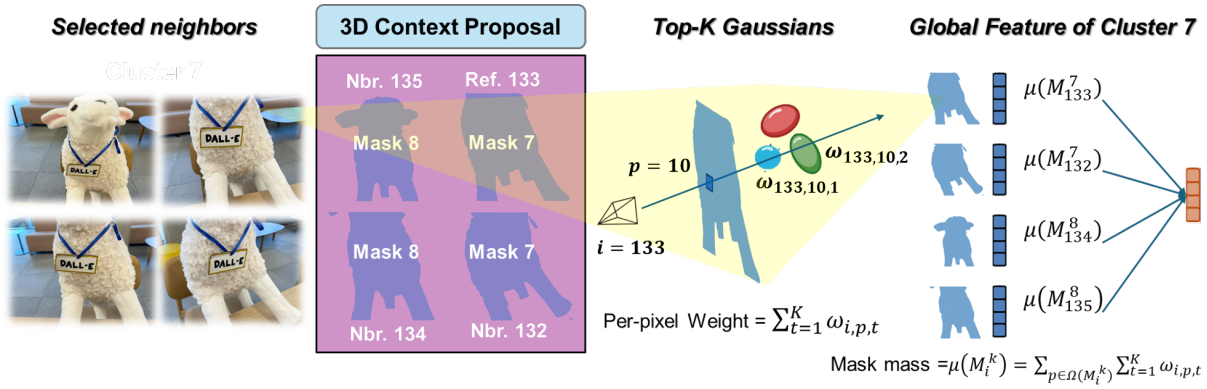

Figure 3. **From context proposal to global feature.** Left: masks of the same entity are grouped into a 3D Context Proposal. Center: for a pixel p, the renderer returns the top-K Gaussians with contributions $\{\omega_{i,p,t}\}_{t=1}^{K}$, from which the *mask mass* $\mu(M_i^k)$ is computed. Right: a mass-weighted pool of member mask embeddings forms the proposal feature, which is registered to Gaussians.

2.2. Feature Registration

The registration stage assigns a unit-normalized feature to every Gaussian. For view i and pixel p, the renderer returns the Top-K Gaussians along the ray $\{(g_{i,p,t}, \omega_{i,p,t})\}_{t=1}^{K}$ with blending contributions $\omega_{i,p,t} = T_{i,p,t}\alpha_{i,p,t}$ and $T_{i,p,t} = \prod_{s<t}(1-\alpha_{i,p,s})$.

Each proposal P_m contains masks from multiple views. We compute the *mask mass* $\mu(M_i^k) = \sum_{p\in\Omega(M_i^k)}\sum_{t=1}^{K}\omega_{i,p,t}$, then form a global language feature for each proposal by mass-weighted pooling

$$\bar{f}_m = \frac{\sum_{(i,k)\in P_m}\mu(M_i^k)\, f_i^k}{\left\|\sum_{(i,k)\in P_m}\mu(M_i^k)\, f_i^k\right\|_2}. \tag{1}$$

We construct per-view proposal label maps $L_i(p)$ that assign masked pixels to proposal IDs. For each pixel p with proposal $m = L_i(p)$ and each Top-K hit, we accumulate

$$\begin{aligned} A[g_{i,p,t}] &\leftarrow A[g_{i,p,t}] + \omega_{i,p,t}\,\bar{f}_m, \\ S[g_{i,p,t}] &\leftarrow S[g_{i,p,t}] + \omega_{i,p,t}. \end{aligned} \tag{2}$$

After all views, each Gaussian feature is $\hat{f}_g = \frac{A[g]}{\max(S[g],\varepsilon)}$ followed by ℓ_2 normalization.

2.3. Inference Procedure

A text query is encoded to $\hat{f}_q$ and each Gaussian is scored by cosine similarity $s_g = \hat{f}_q^{\top}\hat{f}_g$. Gaussians with $s_g \geq \tau_{\text{act}}$ are selected in 3D. For visualization in view i, we sum $\omega_{i,p,t}$ over active Top-K hits at each pixel and threshold to obtain the activation mask. To reduce memory overhead and keep query-time retrieval efficient on large Gaussian scenes, we store each per-Gaussian language descriptor using Product Quantization (PQ). PQ partitions the descriptor into low-dimensional subvectors and encodes each one with a codebook index. At query time, the stored codes are decoded into approximate descriptors for cosine-similarity retrieval.

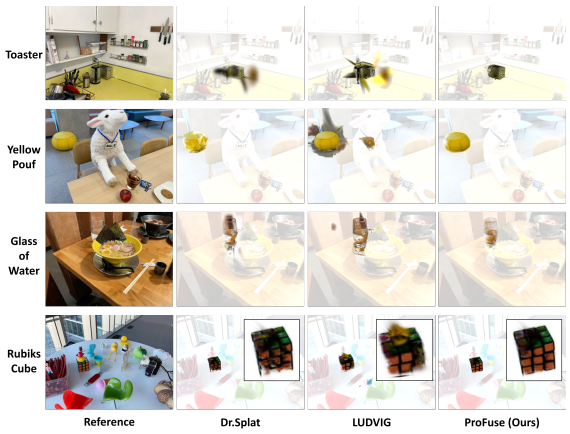

Figure 4. Qualitative comparison on 3D object selection.

Table 1. Evaluation of 3D object selection on LERF-OVS [11] dataset. Scores are averaged per scene and then across scenes. **mIoU** results; bold indicates the best performance.

Method	waldo kitchen	figurines	ramen	teatime	mean
LangSplat	9.18	10.16	7.92	11.38	9.66
LEGaussians	11.78	17.99	15.79	19.27	16.21
OpenGaussian	22.70	39.29	**31.01**	<u>60.44</u>	38.36
Dr. Splat	<u>39.07</u>	<u>53.36</u>	24.70	57.20	<u>43.58</u>
LUDVIG	30.40	37.80	27.80	38.20	33.60
ProFuse (Ours)	**41.93**	**57.74**	<u>29.16</u>	**65.86**	**48.67**

3. Experiments

3.1. Open-Vocabulary 3D Object Selection

We evaluate open-vocabulary 3D object selection on LERF-OVS [11] using official text queries and splits. Each method produces a per-frame activation mask for the queried concept, while our pipeline performs selection in 3D by retrieving Gaussians whose cosine similarity to the text embed-

Table 2. Wall-clock breakdown of deployment and query cost, with mean total Gaussians per scene.

Method	Initialization	Geometry	Semantics	Total Gaussians	Query
OpenGaussian	-	∼20 m	∼50 m	1.32 M	-
Dr. Splat	-	∼20 m	∼10 m	1.32 M	1.65 m
LUDVIG	-	∼20 m	∼10 m	0.60 M	-
ProFuse (Ours)	∼2.5 m	**∼15 m**	**∼0.5 m**	**0.54 M**	**1.00 m**

Table 3. Dataset-level Top-K ablation. For each variant, we report mean mIoU (↑) and the mean feature uplifting time (↓) under different Top-K settings.

Dataset/Method		LERF-OVS			ScanNet		
		K=10	K=20	K=40	K=10	K=20	K=40
Dr. Splat	mIoU↑	43.26	43.29	43.58	35.12	36.34	36.81
	time(s)↓	26	55	105	45	85	165
ProFuse	mIoU↑	48.67	48.67	48.67	39.74	39.74	39.74
	time(s)↓	28	28	28	25	25	25

ding exceeds a threshold. The selected Gaussian set is then projected back to each view through the renderer.

Table 1 reports per-scene and mean mIoU. ProFuse delivers consistently strong performance across all four scenes and achieves the highest mean score. Qualitative results are shown in Figure 4. Our method produces more accurate and cleaner object retrieval, showing sharper correspondence between the text query and the selected 3D content. These results establish that the efficiency gains reported in the following sections are not obtained by sacrificing accuracy.

3.2. Offline Deployment Efficiency

We report wall-clock measurements in Table 2. ProFuse reduces scene preparation overhead at multiple stages. The correspondence-guided initialization establishes both the initial Gaussian seeds and the 3D Context Proposals in about 2.5 minutes. This initialization removes the need for the densification process used in standard 3DGS training. During feature uplifting, proposal-level features are attached through visibility-weighted accumulation without render-supervised semantic optimization. As shown in Table 2, this correspondence-guided design substantially lowers the total deployment cost of semantic scene preparation.

Table 3 analyzes the Top-K accumulation depth used during feature uplifting. We vary the Top-K Gaussian contributors with $K \in \{10, 20, 40\}$ and report the mean mIoU together with the mean uplifting time. The mean accuracy of Dr. Splat improves as K increases at a sharply higher cost. ProFuse reaches its maximum accuracy at $K = 10$, and larger K does not provide additional gain while the uplifting time remains flat. This early saturation is consistent with proposal-level fusion, which concentrates useful semantic mass on the leading contributors of each ray and makes deeper Gaussians contribute negligible additional signal. This interpretation is also supported by our contribution analysis, where the per-view top-10 share ranges from 0.87 to 0.99, indicating that the most useful mass is already concentrated in a small subset of Gaussian contributors. This stable operating point at small K eliminates the need to sweep K across scenes during deployment.

3.3. Online Query Efficiency

After feature uplifting, each Gaussian stores a registered language descriptor with Product Quantization, and query-time retrieval operates directly on this representation. Query efficiency is therefore shaped by the size of the Gaussian scene and by the cost of projecting the selected 3D result back to image space.

Table 2 shows that ProFuse produces a substantially smaller final Gaussian scene than standard 3DGS-based lifting methods. Although correspondence-guided initialization is intentionally overcomplete, pruning removes redundant primitives and yields a lightweight representation for deployment. This compactness lowers the cost of downstream retrieval, indexing, and query-time operations that scale with the number of Gaussians. Under the same ScanNet evaluation pipeline, ProFuse completes full-scene retrieval faster than Dr. Splat while using a much smaller Gaussian set.

The same concentration effect also benefits query time projection. Since ProFuse concentrates the most useful semantic mass on the leading few contributors, the selected 3D result can be projected back to views without relying on a large contributor set. Combined with the more compact Gaussian scene, this yields a more efficient query-time operating point while preserving the best retrieval accuracy reported in Table 1.

4. Conclusion

We presented ProFuse, a resource-efficient framework for open-vocabulary understanding with 3D Gaussian Splatting. ProFuse combines correspondence-guided initialization with proposal-level semantic uplifting to build a compact and queryable 3D semantic scene without render-supervised fine-tuning or gradient-based semantic optimization. This design improves both offline deployment efficiency and online query efficiency: it shortens semantic scene preparation, reaches stable performance at shallow uplifting depth, and produces a substantially smaller Gaussian scene for downstream retrieval. Experiments demonstrate that ProFuse improves retrieval accuracy over prior 3D Gaussian baselines while providing a more efficient and scalable operating point for deployment.

References

[1] Naijian Cao, Renjie He, Yuchao Dai, and Mingyi He. Loflat: Local feature matching using focused linear attention transformer. *arXiv preprint arXiv:2410.22710*, 2024. 1

[2] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *CVPR*, 2025. 1

[3] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. *arXiv preprint arXiv:2202.00667*, 2022. 1

[4] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023. 1

[5] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024. 1

[6] Qingdong He, Jinlong Peng, Zhengkai Jiang, Kai Wu, Xiaozhong Ji, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Mingang Chen, and Yunsheng Wu. Unim-ov3d: Unimodality open-vocabulary 3d scene understanding with fine-grained feature representation. In *IJCAI*, 2024. 1

[7] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *ICRA*, London, UK, 2023. 1

[8] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023. 1

[9] Kim Jun-Seong, Kim GeonU, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *CVPR*, 2025. 1

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[11] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *ICCV*, 2023. 1, 3

[12] Dmytro Kotovenko, Olga Grebenkova, and Björn Ommer. Edgs: Eliminating densification for efficient convergence of 3dgs. *arXiv preprint arXiv:2504.13204*, 2025. 1

[13] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 1

[14] Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. Instancegaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In *CVPR*, 2025. 1

[15] Juliette Marrie, Romain Menegaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[16] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, 2024.

[17] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 1

[18] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, 2024. 1

[19] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. *arXiv preprint arXiv:2405.17596*, 2024. 1

[20] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *CoRL*, 2023. 1

[21] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1

[22] Hongyu Shen, Junfeng Ni, Yixin Chen, Weishuo Li, Mingtao Pei, and Siyuan Huang. Trace3d: Consistent segmentation lifting via gaussian instance tracing. In *ICCV*, 2025. 1

[23] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023. 1

[24] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. *arXiv preprint arXiv:2311.18482*, 2023. 1

[25] Wei Sun, Yanzhao Zhou, Jianbin Jiao, and Yuan Li. Cags: Open-vocabulary 3d scene understanding with context-aware gaussian splatting. *arXiv preprint arXiv:2504.11893*, 2025. 1

[26] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1

[27] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *NeurIPS*, 2024. 1

[28] 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. *arXiv preprint arXiv:2310.03923*, 2023. 1

[29] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *CVPR*, 2024. 1

[30] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *ECCV*, 2024. 1

[31] Hongjia Zhai, Xiyu Zhang, Boming Zhao, Hai Li, Yijia He, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Splatloc: 3d gaussian splatting-based visual localization for augmented reality. *arXiv preprint arXiv:2409.14067*, 2024. 1

[32] 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. 1