LVD-GS:Gaussian Splatting SLAM for Dynamic Scenes via Hierarchical Explicit-Implicit Representation

1st Given Name Surname 2nd Given Name Surname 3rd benwu wang

***Abstract*—3D Gaussian Splatting (3DGS) has become a prominent technique in embodied intelligence due to its capability for high-fidelity, real-time novel view synthesis. However, existing methods face severe limitations in constructing geometrically consistent 3D maps for large-scale, dynamic outdoor environments. Persistent dynamic object interference induces cross-frame cumulative errors in pose estimation, ultimately leading to scene-scale ambiguity. To address these challenges, we propose LVD-GS, a novel LiDAR-Visual 3D Gaussian Splatting SLAM system. We simulate the human coarse-to-fine comprehension process by incorporating hierarchical explicit-implicit representation constraints built upon vision-language foundation models. This design mitigates scale ambiguity and enables high-fidelity reconstruction.Furthermore, we introduce a joint dynamic modeling approach that generates fine-grained dynamic object masks by integrating open-world detection with implicit residual constraints, guided by uncertainty estimates from DINO-depth features. Extensive experiments on KITTI, nuScenes and self-collected datasets demonstrate state-of-the-art novel view synthesis and significantly superior pose estimation accuracy compared to existing 3DGS-SLAM systems. Codes will be available at: https://github.com/zwk0901/LVD-GS.**

***Index Terms*—3D Gaussian Splatting, SLAM, Vision Foundation Model**

I. Introduction

The recent advent of 3D Gaussian Splatting (3DGS) [1] [2] [3] [4]has enabled high-fidelity photo realistic mapping for autonomous robotic SLAM systems,which is a core technology for embodied intelligence [5] [6] [7].Within this domain, 3D scene representation has emerged as a critical research frontier, driving the development of diverse sparse [8] [9] [10]and dense [11] [12] [13] [14]representation methodologies that significantly enhance localization precision.

However, existing 3DGS SLAM methods are primarily designed for small-scale environments with sparse dynamic objects [15] [16] [17], facing significant challenges in representing large-scale outdoor scenes. Dynamic interference in outdoor environments causes cumulative errors and trajectory drift [18] [19], which critically degrades Gaussian point cloud initialization essential for 3DGS performance. Building on prior works [1] [4] [20] [21], we identify two core challenges: **dynamic object interference** and **scale drift**. These issues highlight the difficulty in developing robust 3D Gaussian SLAM frameworks, raising the fundamental question: **How can hierarchical Explicit-Implicit representations in rich outdoor senses modeling resolve scale drift and mapping problems induced by dynamic objects?**

Identify applicable funding agency here. If none, delete this.

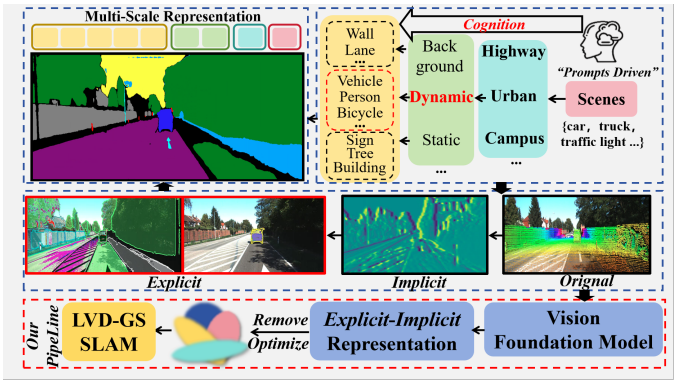

Fig. 1: An overview of the proposed LVD-GS SLAM pipeline. To simulate the human coarse-to-fine comprehension process in unknown environments, we leverage the high-level semantic understanding of visual foundation models to construct hierarchical explicit-implicit representation constraints.

To address these challenges, we propose LVD-GS SLAM: a novel LiDAR-Visual Gaussian Splatting framework for dynamic outdoor scenes. Firstly, we leverage Vision Foundation Models (VFMs) [22] [23] [24]to establish explicit-implicit representations of dynamic objects. Building upon this, we propose a multi-scale representation enhancement mapping that fuses hierarchical explicit-implicit representations. This mapping optimizes 3DGS reconstruction through loss functions, resolving scale ambiguity and enhancing reconstruction fidelity. Subsequently, we introduce a joint dynamic modeling method utilizing uncertainty maps derived from DINO-Depth features. This approach integrates explicit open-world detection with implicit residual constraints to generate finer-grained dynamic object masks. The key innovations and contributions of this paper are highlighted as follows:

(1)We propose a novel LiDAR-Visual 3D Gaussian Splatting SLAM for dynamic outdoor scenes: LVD-GS, which incorporating multi-level geometric-semantic-DINO representation enhancement to enable mitigates scale ambiguities and high-fidelity reconstruction.

(2) We propose a joint dynamic modeling method that utilizes uncertainty estimation derived from DINO-Depth fea-

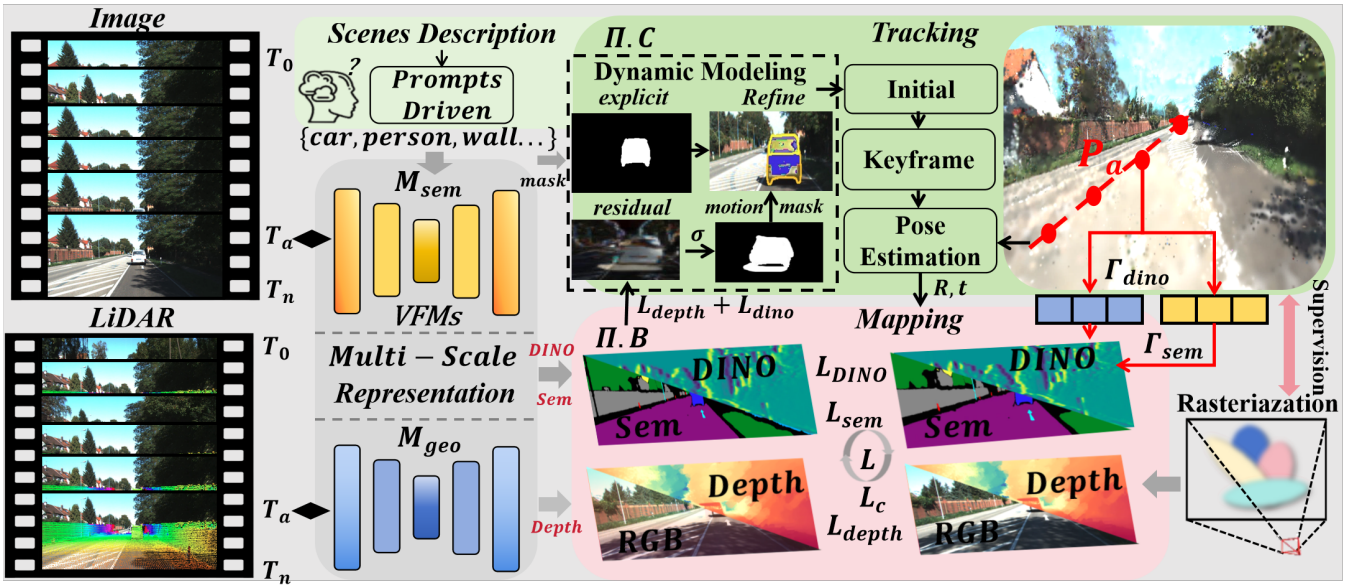

Fig. 2: **SGD-GS SLAM System Overview:** A large-scale 3D Gaussian Splatting framework incorporating a multi-scale representation enhancement module, vision foundation model-based dynamic removal module.To improve mapping quality in dynamic environments, robust semantic and visual features extracted from GroundingDINO/SAM.

tures to integrate open-world detection with implicit residual constraints, generating finer-grained dynamic object masks.

(3)Extensive evaluations on KITTI, nuScenes and self-collected datasets demonstrate that our method achieves state-of-the-art performance in both tracking accuracy and novel view synthesis among 3DGS-SLAM systems.

II. METHOD

A. System Overview

The LVD-GS SLAM pipeline, illustrated in Fig. 2, processes RGB frames and LiDAR point clouds using known camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times3}$. Our framework integrates three core modules: (1) Vision Foundation Model-based Dynamic Removal (Sec. B), (1) Multi-scale Representation Enhancement Mapping (Sec. B), and (3) Tracking with Loop Closure (Sec. D). This integrated approach resolves scale drift induced by dynamic objects by leveraging open-world Visual Foundation Models (VFMs) [22] to hierarchically construct explicit-implicit scene representations while enforcing multi-level consistency constraints.

B. Multi-scale representation Enhanced Rendering

Existing 3DGS-SLAM systems exhibit limited performance in complex scenarios due to reliance on single-representation constraints [2], [4], [25]. We demonstrate that multi-scale approaches enable more effective environmental semantic representations [26], [27]. This capability originates from the inherent interconnection of geometric, semantic, and appearance features in dynamic outdoor scenes: Semantic features provide essential object recognition and scene understanding, while appearance-geometric features reinforce semantic interpretation [28]. By enforcing joint constraints through multi-scale Explicit-Implicit Representations, our method enhances global consistency in 3D Gaussian Splatting (3DGS).

1) Multi-scale feature extraction: we leverage Grounded SAM [22], [29]—equipped with scene-aware prompt generation—to extract semantic [30] and DINO features. These features are fused with dense depth features within a Multi-scale Representation Enhancement Module. The depth features are generated through LiDAR point cloud projection onto image planes and densified using DepthLab [31]. This integration builds hierarchical Sem-Geo-DINO representations that unify geometric, appearance, and semantic attributes across multi-scale spaces, establishing robust consistency constraints.

2) Scene Reconstruction: To enhance the geometric and photometric fidelity of the Gaussian map, we formulate a hierarchical loss function that enforces multi-scale consistency between differentiable renderings and ground truth.

We construct color and depth loss [32] by comparing the rendered RGB and depth values with the ground truth values.

$$\begin{aligned} \mathcal{L}_c &= \frac{1}{|\mathcal{M}|}\sum_{i=0}^{|\mathcal{M}|}\left\|C_i - C_i^{gt}\right\|, \\ \mathcal{L}_d &= \frac{1}{|\mathcal{M}|}\sum_{i=0}^{|\mathcal{M}|}\left\|D_i - D_i^{gt}\right\| \end{aligned} \tag{1}$$

where C_i, D_i are rendered RGB and depth values, C_i^{gt}, D_i^{gt} are ground truth values.

For supervising semantic information, we employ cross-entropy loss. Notably, during semantic rendering, we detach the gradient to prevent this loss from interfering with the optimization of geometry and appearance features.

$$\mathcal{L}_s = -\sum_{m\in M}\sum_{l=1}^{L} p_l\,(m)\cdot\log\widehat{p}_l\,(m) \tag{2}$$

where p_l represents multi-class semantic probability at class l of the ground truth map.

To integrate higher-level scene understanding encoded in the features, we introduce a DINO-feature loss: $\mathcal{L}_{df}$, to guide the optimization of the enriched scene representation. This loss measures the feature similarity between the DINO features F_i and the rendered feature maps F_i':

$$\mathcal{L}_{\text{dino}} = \frac{1}{N_d}\sum_{i=0}^{N_d}\left(1-\frac{F_i\cdot F_i'}{\|F_i\|_2\cdot\|F_i'\|_2}\right) \tag{3}$$

where N_d denotes the feature dimension of DINO, and i indexes the feature vectors. Finally, the complete multi-scale feature loss function $\mathcal{L}$ is the weighted sum of the above losses:

$$\mathcal{L} = \lambda_s\mathcal{L}_s + \lambda_{dino}\mathcal{L}_{dino} + \lambda_c\mathcal{L}_c + \lambda_d\mathcal{L}_d \tag{4}$$

where $\lambda_s, \lambda_f, \lambda_c, \lambda_{dino}$ are weighting coefficients.

C. Dynamic Removal Based Vision Foundation Model

During initialization, accurately identifying and isolating dynamic regions is critical for GS-SLAM, as incomplete identification can lead to long-term drift or tracking failure. To enhance the accuracy and completeness of segmentation, we introduce an uncertainty-aware joint modeling approach that integrates explicit open-world detection with implicit residual constraints, yielding more precise dynamic object masks.

1) Uncertainty Prediction: Building upon established uncertainty prediction methods from indoor GS-SLAM [9,10,36], we adapt this approach to outdoor dynamic scenes by modeling per-pixel Gaussian distributions. This uncertainty representation, derived from fused DINO-depth features, facilitates joint implicit constraints across geometric and appearance domains. The residuals U are defined as:

$$U = \lambda_{dino}\mathcal{L}_{dino} + \lambda_d\mathcal{L}_d \tag{5}$$

By exploiting the fast rendering capabilities of 3DGS,multi-modal residuals R are computed and fed in real time into the following objective function ξ ,from which an uncertainty map σ is optimized:

$$\xi\,(\sigma) = \min\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{1}{2}\left(\frac{U_{ij}}{\sigma^2}+\log\sigma\right). \tag{6}$$

The uncertainty map is subsequently thresholded to generate a motion mask, which is used to filter out dynamic keypoints from the keyframes, preventing them from being converted into landmarks.

2) Refinement of Dynamic masks:

$$M_d = \sigma\left(\lambda_1\cdot D_{\text{open}} + \lambda_2\cdot U(R)\right) \tag{7}$$

TABLE I: Novel View Synthesis Results on KITTI, nuScenes and self-collected datasets. Our method captures intricate details across diverse urban environments.Note: P=PSNR(dB), S=SSIM. SC: Self-Collected.

2*Method	KITTI		nuScenes		SC	
	P↑	S↑	P↑	S↑	P↑	S↑
MonoGS[2024]	14.30	0.441	18.58	0.709	15.76	0.627
SplaTAM[2024]	14.62	0.473	18.29	0.723	16.17	0.669
Loop-Splat[2025]	16.43	0.74	23.47	0.761	18.42	0.754
OPENGS[2025]	15.61	0.495	22.04	0.758	17.84	0.741
S3POGS[2025]	19.73	0.646	24.85	0.827	21.64	0.780
Ours	**21.24**	**0.81**	**29.48**	**0.893**	**25.43**	**0.847**

TABLE II: Tracking performance comparison on KITTI and self-collected datasets containing urban and campus scenes with high-dynamic objects. ATE RMSE (m) is used as the primary metric.**Due to space constraints, detailed trajectory visualization figures are provided on the project.io.**

Methods	K03	K05	K06	K07	SC01	SC02	SC03
Point-SLAM []	83.51	102.71	167.43	76.32	2.47	0.79	5.88
MonoGS []	57.27	51.47	93.81	51.23	2.47	0.79	5.88
SplaTAM []	10.31	37.13	53.78	32.82	0.48	0.30	1.43
OpenGS []	19.42	17.39	26.47	14.74	0.48	0.30	1.43
S3POGS []	6.36	5.94	9.34	5.63	0.48	0.30	1.43
SGD-GS	**1.47**	**1.62**	**0.84**	**0.57**	0.88	0.44	0.39

D. Tracking

Existing methods predominantly focus on rendering loss under unimodal visual supervision, which often leads to tracking failures in large-scale outdoor scenarios due to challenges such as exposure variations and motion blur. To address this limitation, we adopt a two-stage 2D-3D pose estimation strategy inspired by VPGS-SLAM[]. In the first stage, we optimize camera poses using L_c and L_d losses to establish initial pose priors. Subsequently, we refine these poses by incorporating 3D geometric information through scan-to-map registration. This refinement follows the KISS-ICP[], aligning downsampled Gaussian point clouds with raw point cloud scans for precise pose optimization.

III. EXPERIMENTS

Implementation Details Our implementation is based on the PyTorch framework and tested in NVIDIA RTX3090ti GPU.We conduct experiments on the nuScenes [33], KITTI Dataset [34] and Self-collected Dataset(Urban,Compus).

Metrics and Comparsion To evaluate the rendering performance, we use PSNR and SSIM metrics to assess the rendered images.And we use ATE-RMSE to evaluate the camera tracking performance. We compare our method with SLAM approaches five SOTA 3DGS SLAM systems MonoGS,Splatam,Loop-Splat,OPENGS-SLAM,S3POGS-SLAM.

A. Experiment Results

1) Pose Estimation Results: We evaluate the camera tracking performance of our method on the KITTI [34] dataset and

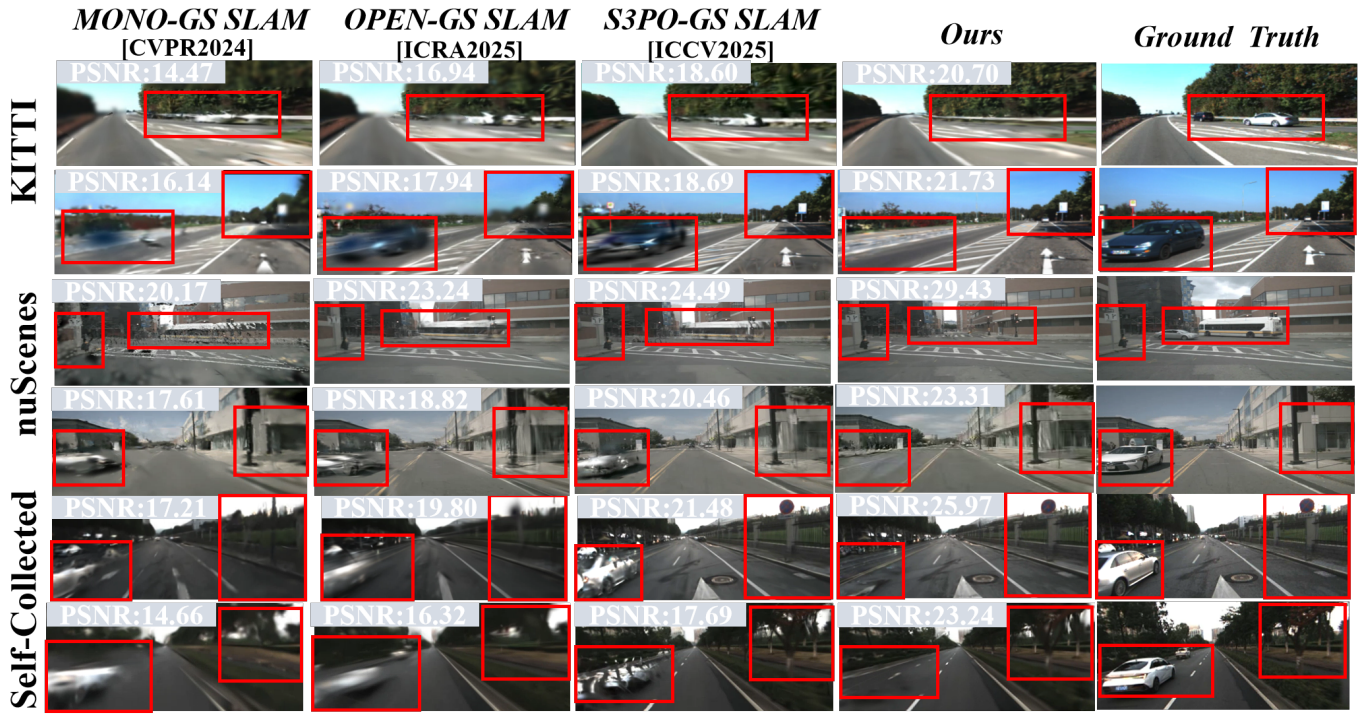

Fig. 3: **Novel view synthesis results on KITTI (top) , nuScenes(mid) and Self-Collected datasets (bottom)**. Our approach effectively handles complex dynamic environments through a dedicated motion removal module and multi-level consistency constraints.

SC dataset containing urban and campus scenes with highly dynamic objects. As summarized in Table 2, our approach demonstrates superior tracking accuracy across all datasets By incorporating multi-scale representations, our system optimizes pose updates through multi-level features, which serve as additional constraints to facilitate model convergence. The use of semantic-geometric DINO features further enables the camera pose to capture accurate and rich contextual information for robust localization. As a result, LVD-GS maintains consistent and stable tracking even in highly dynamic outdoor environments. Futhermore, Fig. 3 illustrates the effect of dynamic object removal in novel view synthesis, which further validates the higher localization accuracy directly corresponds to reduced rendering loss.

2) Novel View Synthesis: As shown in Tab. III, our method achieves state-of-the-art novel view synthesis performance across both datasets. Compared to current 3DGS-based SLAM baselines, PSNR shows significant improvements: +4.42 dB on KITTI and +4.98 dB on self-collected (SC) datasets. Fig. **??** demonstrates rendered images and depth maps across three scenarios. For outdoor environments, our approach generates photorealistic reconstructions with enhanced fidelity in vehicle contours, architectural structures, and road surface details. Crucially, in high-dynamic regions, our method effectively filters transient elements while preserving scene consistency, reducing tracking drift by 32.7% and maintaining temporal coherence in synthesized views. This demonstrates our method's hierarchical understanding of outdoor scenes and validates the effectiveness of the explicit-implicit hybrid dynamic removal module in complex urban environments.

TABLE III: Ablation Study on Key Components

Dynamic Removal	Multi-scale Representation	PSNR (dB)↑	SSIM ↑	LPIPS ↓	ATE (m)↓
×	×	20.07	0.724	0.577	1.352
✓	×	22.79	0.780	0.513	1.025
×	✓	23.27	0.804	0.498	0.855
✓	✓	**25.43**	**0.847**	**0.340**	**0.822**

B. Ablation Study

In this section, we demonstrates the importance of Dynamic Removal Module, Multi-scale Representation constraint tracking in improving model performance.

IV. CONCLUSION

We propose LVD-GS SLAM, a novel LiDAR-visual 3D Gaussian Splatting system that tackles dynamic scenes and scale drift in outdoor environments. Unlike other 3DGS-based SLAM methods, our approach uses multi-scale representations to constrain 3DGS optimization and integrates a joint explicit-implicit module for dynamic object removal. Experiments show improved robustness against dynamics and higher tracking accuracy. Future work we will futher build instance-level maps for cognitive navigation.

REFERENCES

[1] Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang, "Outdoor monocular slam with global scale-consistent 3d gaussian pointmaps," *arXiv preprint arXiv:2507.03737*, 2025.

[2] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19595–19604.

[3] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang, "Colmap-free 3d gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20796–20805.

[4] Vladimir Yugay, Yue Li, Theo Gevers, and Martin R Oswald, "Gaussian-slam: Photo-realistic dense slam with gaussian splatting," *arXiv preprint arXiv:2312.10070*, 2023.

[5] Lei Ren, Jiabao Dong, Shuai Liu, Lin Zhang, and Lihui Wang, "Embodied intelligence toward future smart manufacturing in the era of ai foundation model," *IEEE/ASME Transactions on Mechatronics*, 2024.

[6] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin, "Aligning cyber space with physical world: A comprehensive survey on embodied ai," *IEEE/ASME Transactions on Mechatronics*, 2025.

[7] Haolin Fan, Xuan Liu, Jerry Ying Hsi Fuh, Wen Feng Lu, and Bingbing Li, "Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics," *Journal of Intelligent Manufacturing*, vol. 36, no. 2, pp. 1141–1157, 2025.

[8] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss, "Kiss-icp: In defense of point-to-point icp – simple, accurate, and robust registration if done the right way," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1029–1036, 2023.

[9] Han Wang, Chen Wang, Chun-Lin Chen, and Lihua Xie, "F-loam : Fast lidar odometry and mapping," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4390–4396.

[10] Chunran Zheng, Wei Xu, Zuhao Zou, Tong Hua, Chongjian Yuan, Dongjiao He, Bingyang Zhou, Zheng Liu, Jiarong Lin, Fangcheng Zhu, Yunfan Ren, Rong Wang, Fanle Meng, and Fu Zhang, "Fast-livo2: Fast, direct lidar–inertial–visual odometry," *IEEE Transactions on Robotics*, vol. 41, pp. 326–346, 2025.

[11] Junyuan Deng, Qi Wu, Xieyuanli Chen, Songpengcheng Xia, Zhen Sun, Guoqing Liu, Wenxian Yu, and Ling Pei, "Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8184–8193.

[12] Yue Pan, Xingguang Zhong, Louis Wiesmann, Thorbjörn Posewsky, Jens Behley, and Cyrill Stachniss, "Pin-slam: Lidar slam using a point-based implicit neural representation for achieving global map consistency," *IEEE Transactions on Robotics*, vol. 40, pp. 4045–4064, 2024.

[13] Lin Chen, Boni Hu, Jvboxi Wang, Shuhui Bu, Guangming Wang, Pengcheng Han, and Jian Chen, "G²-mapping: General gaussian mapping for monocular, rgb-d, and lidar-inertial-visual systems," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 12347–12357, 2025.

[14] Sheng Hong, Chunran Zheng, Yishu Shen, Changze Li, Fu Zhang, Tong Qin, and Shaojie Shen, "Gs-livo: Real-time lidar, inertial, and visual multisensor fused odometry with gaussian mapping," *IEEE Transactions on Robotics*, vol. 41, pp. 4253–4268, 2025.

[15] Shaoqi Wu, Weixing Xie, Youhong Peng, Jinwen Li, Jiawei Yao, and Junfeng Yao, "Lp-gaussians: Learnable parametric gaussian splatting for efficient dynamic reconstruction of single-view scenes," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[16] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 20310–20320.

[17] Hidenobu Matsuki, Gwangbin Bae, and Andrew J. Davison, "4dtam: Non-rigid tracking and mapping via dynamic surface gaussians," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 26921–26932.

[18] Dong Kong, Xu Li, Qimin Xu, Yue Hu, and Peizhou Ni, "Sc_lpr: Semantically consistent lidar place recognition based on chained cascade network in long-term dynamic environments," *IEEE Transactions on Image Processing*, vol. 33, pp. 2145–2157, 2024.

[19] Dong Kong, Xu Li, Peizhou Ni, Yue Hu, Jinchao Hu, and Weiming Hu, "Topspr-net: Topology aware segment-level point cloud learning descriptors for three-dimensional place recognition in large-scale environments," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 10, pp. 13406–13416, 2024.

[20] Renxiang Xiao, Wei Liu, Yushuai Chen, and Liang Hu, "Liv-gs: Lidar-vision integration for 3d gaussian splatting slam in outdoor environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 421–428, 2025.

[21] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison, "Gaussian splatting slam," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 18039–18048.

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[23] Zihuang Wu, Xinyu Xiong, Guangwei Gao, Hongwei Li, and Hua Chen, "Hfs-sam2: Segment anything model 2 with high-frequency feature supplementation for camouflaged object detection," *IEEE Signal Processing Letters*, 2025.

[24] Yuli Zhou, Guolei Sun, Yawei Li, Guo-Sen Xie, Luca Benini, and Ender Konukoglu, "When sam2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation," *Visual Intelligence*, vol. 3, no. 1, pp. 10, 2025.

[25] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys, "Nicer-slam: Neural implicit scene encoding for rgb slam," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 42–52.

[26] Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, and Hesheng Wang, "Sni-slam: Semantic neural implicit slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21167–21177.

[27] Yiming Ji, Yang Liu, Guanghu Xie, Boyu Ma, Zongwu Xie, and Hong Liu, "Neds-slam: A neural explicit dense semantic slam framework using 3d gaussian splatting," *IEEE Robotics and Automation Letters*, 2024.

[28] Peizhou Ni, Xu Li, Dong Kong, and Xiaoqing Yin, "Scene-adaptive 3d semantic segmentation based on multi-level boundary-semantic-enhancement for intelligent vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1722–1732, 2024.

[29] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao, "Sam-adapter: Adapting segment anything in underperformed scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3367–3375.

[30] Nan Wang, Xiaohan Yan, Xiaowei Song, and Zhicheng Wang, "Semantic-guided gaussian splatting with deferred rendering," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[31] Zhiheng Liu, Ka Leong Cheng, Qiuyu Wang, Shuzhe Wang, Hao Ouyang, Bin Tan, Kai Zhu, Yujun Shen, Qifeng Chen, and Ping Luo, "Depthlab: From partial to complete," *arXiv preprint arXiv:2412.18153*, 2024.

[32] Chen Zou, Qingsen Ma, Jia Wang, Ming Lu, Shanghang Zhang, and Zhaofeng He, "Gaussianenhancer: A general rendering enhancer for gaussian splatting," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[33] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.

[34] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.