2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MASTD3R-SLAM: Monocular Adaptive Semantic Tracking and Dynamic Reconstruction SLAM

Fengwei Yang[1*] Qingran Lin[2*] Chaolun Zhu[3*†]

***Abstract*— The challenge of dynamic scenes has long been one of the core issues in the application and generalization of SLAM systems. Traditional visual SLAM systems often rely on depth sensors and prior camera parameters, making it difficult to correct dynamic challenges from arbitrary input images while simultaneously constructing dense maps. Recently, view-oriented point cloud prediction foundation models have attracted significant attention. Their impressive capability of performing 3D reconstruction without requiring camera priors has led to the emergence of SLAM systems such as SLAM3R and MASt3R-SLAM. However, these systems face challenges when applied to dynamic scenes and cannot directly use traditional methods for correction, such as semantic masking or optical flow segmentation. To address this issue, we propose MASTD3R-SLAM, a SLAM method specifically designed for dynamic scenes that supports arbitrary video inputs. The method combines fused mask-based processing with coarse-to-fine pointmap alignment and optimization to achieve point cloud–to–pose re-mapping correction, and further performs Gaussian rendering to remove rendering artifacts and suppress dynamic mapping interference. Compared to the original baseline, our approach improves tracking ATE accuracy by more than 20% and successfully restores the correct 3D map.**

I. INTRODUCTION

SLAM (Simultaneous Localization and Mapping) is a comprehensive approach to addressing challenges in robotic perception. Among the various solutions, methods based on visual sensors are considered an important pathway due to their low cost, ease of acquisition, and high information density. However, in real-world applications, SLAM systems often face uncertainties and disturbances, such as dynamic objects and scene changes. To address these issues, traditional visual SLAM methods typically require precise camera parameters and additional sensor data (e.g., depth and LiDAR) to effectively constrain dynamic interference in images.

In recent years, neural network-based methods have attracted increasing attention. These approaches can directly predict point clouds from sparse images, or even a single view, enabling real-time dense SLAM without the need for prior camera intrinsic calibration. Representative works such as MASt3R-SLAM [1] and SLAM3R [2] have demonstrated the ability to reconstruct 3D scenes from arbitrary video inputs, even in extremely sparse settings (e.g., only a few dozen or fewer images). However, these methods still face significant challenges in dynamic scenes: under dynamic disturbances, they often suffer from severe accumulated errors or even tracking loss, while traditional dynamic SLAM solutions are difficult to directly apply. Traditional dynamic SLAM methods [3] typically achieve performance gains by combining semantic segmentation and optical flow masks with geometric tracking frameworks such as ORB features [4]. However, for systems based on MASt3R, such direct mask-based corrections are difficult to apply because the pose prediction process does not rely on geometric structures. In addition, conventional approaches often assume predefined priors, including known camera intrinsics and dynamic objects.

To address these issues, we propose **MASTD3R-SLAM**. First, our method employs a DINOV2-based [5] segmentation model to generate semantic masks, and further integrates depth anomaly detection to adaptively identify dynamic pixels, producing fused dynamic masks. We then introduce an uncertainty-aware coarse-to-fine pointmap correction, including point cloud alignment and global optimization. This process directly re-estimates camera poses from 3D point clouds and remaps corrected point clouds, which is highly consistent with the characteristics of two-view 3D reconstruction networks and fundamentally different from traditional 2D-3D approaches. Finally, to obtain high-fidelity models and reduce rendering artifacts, we incorporate 3D Gaussian reconstruction [6] and impose rendering constraints directly on the fused point clouds, without introducing any additional dynamic loss terms, thereby improving generalization. Experimental results show that our method can achieve stable tracking and 3D reconstruction from arbitrary dynamic video inputs without requiring any pre-calibrated camera parameters or pretrained priors.

The main contributions of this work are summarized as follows:

1) We propose MASTD3R-SLAM, a complete SLAM framework for arbitrary dynamic video inputs, which includes mask fusion, pointmap correction, tracking, and rendering optimization modules. By constructing uncertainty-aware corrections from fused masks, we generate corrected 3D point clouds and design mapping, direct correction, and pruning starting from 3D results. This enables pose correction and Gaussian initialization, ultimately achieving high-fidelity dynamic-free reconstruction.
2) We introduce an adaptive mask based on the fusion of depth and semantic cues, which can identify dynamic pixels through depth anomaly detection and provide uncertainty guidance for two-view networks, thereby separating dynamic point clouds.
3) We present a coarse-to-fine pointmap reconstruction

*:Equal contribution. †:Corresponding author.
[1]Duke University [2]Georgia Tech [3]Waseda University

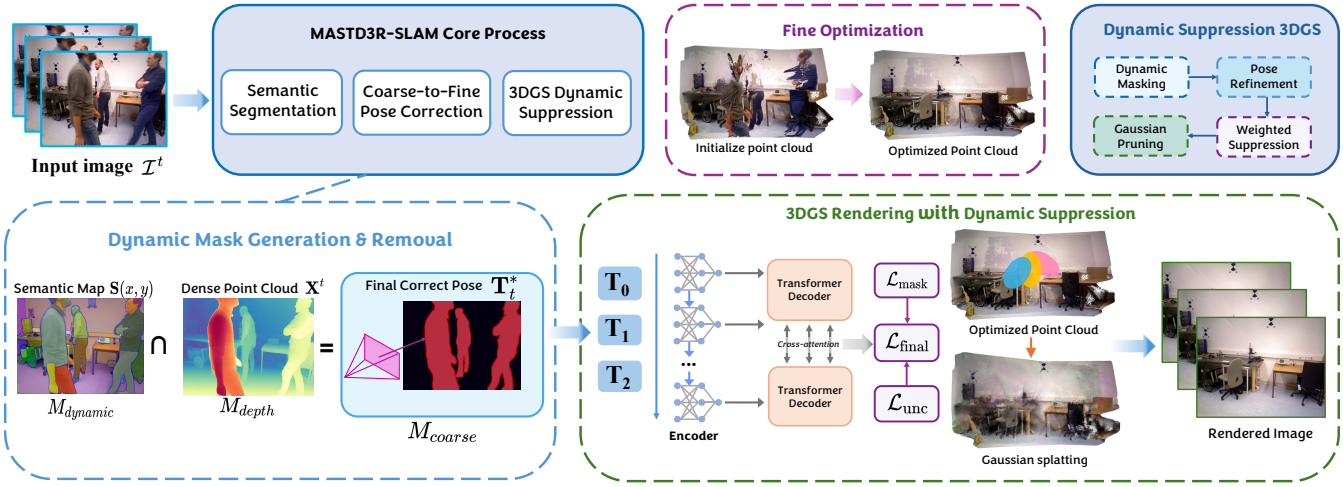

Fig. 1. The framework of our MASTD3R-SLAM is illustrated in the figure. Our pipeline consists of three main contributions: Semantic Segmentation, Coarse-to-Fine Pose Correction, and 3DGS Dynamic Suppression. We first obtain fused depth through a fusion module to assist in dynamic removal. Then, dynamic interference is addressed in a coarse-to-fine manner via a remapping process. Finally, the Gaussian rendering module corrects dynamic artifacts and refines rendering to achieve high-quality visualization.Our pipeline is end-to-end and maintains a running speed of 13 FPS.

method that employs a two-step optimization process to eliminate accumulated errors, including local-to-global point cloud alignment and pruning, as well as Gaussian rendering based on fused masks and point clouds. Our method does not rely on any additional rendering losses or specially designed matching strategies. We conduct experiments on three dynamic datasets, including highly challenging outdoor dynamic scenes. The results demonstrate that our method outperforms baseline approaches in PSNR and ATE accuracy.

II. RELATED WORK

Recent advances in neural 3D reconstruction have significantly improved visual perception systems for SLAM. Traditional multi-view reconstruction methods rely on geometric priors and triangulation across multiple views, which require accurate camera intrinsics and sufficient view overlap. However, in real-world scenarios, illumination changes, sparse textures, and unknown camera parameters often limit their applicability. To address these issues, DUSt3R [7] introduced a two-view reconstruction framework that directly predicts registered point clouds by jointly reasoning about correspondences, camera poses, and scene geometry. MASt3R [1] further improves feature matching and localization to enhance generalization.

Building upon these reconstruction priors, recent works have explored SLAM systems that leverage learned pointmap representations. MASt3R-SLAM [1] introduces incremental pointmap matching and global optimization for real-time tracking and mapping. Similarly, SLAM3R [2] employs improved I2P and L2W modules for stable local-to-global reconstruction, while Droid-Splat [8] integrates optical-flow-based tracking with Gaussian splatting. Despite their effectiveness, these approaches generally assume static scenes and remain vulnerable to drift and mapping errors under dynamic interference.

Traditional dynamic SLAM methods [9], [10] typically introduce geometric constraints with semantic segmentation during the tracking stage for correction. However, in deep reconstruction-based SLAM systems, tracking is embedded within network inference, making such direct feature removal or mask-based correction difficult to apply. To address this limitation, we propose a mapping-to-tracking correction scheme that enables uncertainty-aware point cloud pruning and fusion without additional post-processing.

Recent multimodal reasoning frameworks also explore enhanced visual understanding through agent-based reasoning, vision-language-action models, and tool-augmented video analysis [11]–[13].

III. METHOD

The method in this paper consists of three core components: semantic segmentation and dynamic mask removal, coarse-to-fine point cloud matching, and tracking rendering bundle adjustment. We fully leverage the two-view 3D reconstruction priors of MASt3R, while introducing new strategies to enhance robustness and accuracy. A framework diagram of the method is shown in Figure 1. In Sec III-A we introduce how to obtain more accurate and adaptive segmentation masks. In SecIII-B we present the coarse-to-fine mapping and pose correction process. In SecIII-C we describe the tightly coupled adaptive dynamic Gaussian rendering.

A. Semantic Segmentation and Dynamic Mask Removal

To effectively identify and remove interference regions in dynamic environments, we design a dynamic mask generation method that combines semantic segmentation with depth anomaly detection. In contrast to traditional approaches that rely on geometric consistency of predicted point clouds, our method leverages semantic priors and statistical modeling

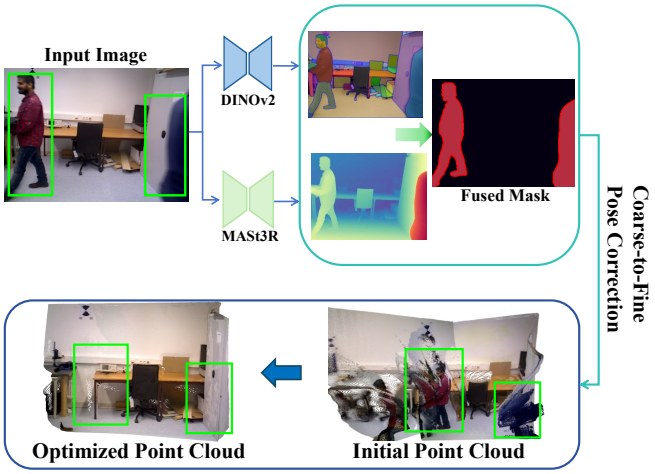

Fig. 2. Our fusion mask illustration demonstrates how the fusion process enables dynamic object recognition and segmentation, while simultaneously obtaining a more accurate base point cloud for pose recovery and Gaussian initialization.

to enhance the detection of dynamic regions. Moreover, we exploit the inherent depth estimation capability of the MASt3R model, avoiding the need for an additional depth estimation network.

First, we apply the generalizable DINOV2 model to the input image $\mathcal{I}^t$ to obtain a semantic response map $\mathbf{S}(x, y)$. Different from conventional methods that explicitly divide pixels into dynamic categories (e.g., pedestrians, vehicles) and static background, we do not predefine such boundaries. Instead, $\mathbf{S}(x, y)$ is interpreted as a semantic prior probability map that reflects the uncertainty of each pixel belonging to different categories.

Then, the MASt3R model is used to regress a dense point cloud $\mathbf{X}^t \in \mathbb{R}^{H\times W\times 3}$ and the corresponding depth map $\mathbf{Z}^t \in \mathbb{R}^{H\times W}$. Dynamic regions typically appear as anomalies inconsistent with their local neighborhood in depth. To capture this property, we compute the local depth variance $\sigma^2_{\text{local}}(x, y)$ and assume that background depths follow a global Gaussian distribution $N(\mu_{\text{depth}}, \sigma^2_{\text{depth}})$. The depth anomaly response is then formulated as:

$$\mathbf{M}_{\text{depth}}(x, y) = \begin{cases} 1 & \text{if } \sigma^2_{\text{local}}(x, y) > \tau_{\text{var}} \\ & \quad \text{or } |\mathbf{Z}^t(x, y) - \mu_{\text{depth}}| > k \cdot \sigma_{\text{depth}}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

To adaptively distinguish between dynamic and static pixels, we further employ a Bayesian framework that fuses semantic responses with depth anomalies. The posterior probability of pixel (x, y) belonging to a dynamic region D is defined as:

$$P(D \mid \mathbf{Z}, \mathbf{S}) = \frac{P(\mathbf{Z} \mid D)\, P(D \mid \mathbf{S})}{P(\mathbf{Z} \mid D)\, P(D \mid \mathbf{S}) + P(\mathbf{Z} \mid \bar{D})\, P(\bar{D} \mid \mathbf{S})}. \tag{2}$$

Here, $P(D \mid \mathbf{S})$ represents the semantic prior probability directly obtained from the normalized DINOV2 output rather than a predefined binary label. $P(\mathbf{Z} \mid D)$ and $P(\mathbf{Z} \mid \bar{D})$ denote the likelihood of observing the depth under dynamic and static hypotheses, respectively. This formulation avoids deterministic labeling and instead provides an uncertainty-based estimation of whether a pixel is dynamic or static.

Finally, by applying the threshold τ_{bayes}, we obtain the dynamic mask:

$$\mathbf{M}_{\text{dynamic}}(x, y) = \begin{cases} 1 & \text{if } P(D \mid \mathbf{Z}, \mathbf{S}) > \tau_{\text{bayes}}, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

which is then combined with the depth anomaly mask to generate the final coarse mask:

$$\mathbf{M}_{\text{coarse}} = \mathbf{M}_{\text{dynamic}} \cap \mathbf{M}_{\text{depth}}. \tag{4}$$

Unlike traditional binary mask removal strategies, $\mathbf{M}_{\text{coarse}}$ not only identifies potential dynamic regions but also serves as an *uncertainty-aware label* integrated into the network optimization process. Specifically, the mask provides confidence weights for dynamic/static regions in both point cloud updating and pose estimation, allowing the network to adaptively adjust the loss function during training and optimization.

B. Coarse-to-Fine Pose Correction

Based on the combined mask $\mathbf{M}_{\text{coarse}}$, we design a coarse-to-fine pose correction procedure. The core idea is to first obtain a robust initial alignment using static point clouds, then correct for monocular scale drift, and finally perform a unified optimization to refine the pose.

For two consecutive frames $(\mathcal{I}^{t-1}, \mathcal{I}^t)$, the MASt3R model regresses the corresponding point clouds $\mathbf{X}^{t-1}, \mathbf{X}^t$ and feature descriptors $\mathbf{D}^{t-1}, \mathbf{D}^t$. Dynamic regions are filtered using $\mathbf{M}^{t-1}_{\text{coarse}}$ and $\mathbf{M}^t_{\text{coarse}}$, resulting in static point clouds:

$$\widetilde{\mathbf{X}}^{t-1} = \{\mathbf{X}^{t-1}(p) \mid \mathbf{M}^{t-1}_{\text{coarse}}(p) = 0\} \tag{5}$$

$$\widetilde{\mathbf{X}}^{t} = \{\mathbf{X}^{t}(q) \mid \mathbf{M}^{t}_{\text{coarse}}(q) = 0\}. \tag{6}$$

In the coarse stage, initial matches are established using feature similarity:

$$s_{p,q} = \mathbf{D}^{t-1}(p)^\top \mathbf{D}^t(q), \tag{7}$$

and the initial matching set $\widetilde{\mathcal{M}}_{\text{coarse}}$ is obtained. We then apply the Iterative Closest Point (ICP) method in a point-to-point formulation to estimate the initial pose:

$$\mathbf{T}_{\text{coarse}} = \arg\min_{\mathbf{T}\in SE(3)} \sum_{(p,q)\in\widetilde{\mathcal{M}}_{\text{coarse}}} \left\| \widetilde{\mathbf{X}}^{t-1}(p) - \mathbf{T}\cdot\widetilde{\mathbf{X}}^{t}(q) \right\|^2. \tag{8}$$

The ICP algorithm iteratively alternates between nearest-neighbor search (to update correspondences) and least-squares optimization (to update $\mathbf{T}$) until convergence. Since dynamic points have already been removed, the ICP correspondences are more reliable, leading to a robust initialization $\mathbf{T}_{\text{coarse}}$.

To further correct scale drift, we estimate a patch-wise scale factor from the median depth:

$$s_k = \frac{\text{med}(z_k^t)}{\text{med}(z_k^{\text{pre}})}, \quad s_{\text{global}} = \text{median}\{s_k\}, \tag{9}$$

and align the point cloud as

$$\widetilde{\mathbf{X}}^t_{\text{aligned}} = s_{\text{global}} \cdot \widetilde{\mathbf{X}}^t. \tag{10}$$

In the fine stage, we take $\mathbf{T}$coarse as initialization and directly perform nonlinear optimization using the corrected point clouds. Different from methods that only rely on feature constraints, we explicitly correct the drifted poses through point cloud re-mapping. Specifically, the corrected point cloud $\widetilde{\mathbf{X}}^t$aligned is mapped back to the previous frame $\widetilde{\mathbf{X}}^{t-1}$ in the global coordinate system, and the geometric residuals between them are minimized:

$$\mathbf{T}t^* = \arg\min \mathbf{T} \in SE(3) \tag{11}$$

$$\sum_{(p,q)\in\widetilde{\mathcal{M}}\text{fine}} \left|\widetilde{\mathbf{X}}^{t-1}(p) - \mathbf{T}\cdot\widetilde{\mathbf{X}}^t\text{aligned}(q)\right|^2 \lambda_{\text{cham}} E_{\text{cham}}(\mathbf{T}). \tag{12}$$

Here, the first term is a point-to-point constraint based on point cloud re-mapping, which explicitly eliminates drift caused by monocular estimation; the second term E_{cham} is a bidirectional Chamfer distance that further enforces global geometric consistency. Through this point cloud–driven re-mapping correction, the pose drift can be gradually compensated during the iterative optimization process, yielding the final optimized pose $\mathbf{T}_t^*$.

C. 3DGS Rendering with Dynamic Suppression

To achieve high-fidelity scene reconstruction in dynamic environments, we introduce 3D Gaussian Splatting (3DGS) as a unified rendering and optimization backend. 3DGS models the scene using a set of anisotropic Gaussian primitives, each parameterized by position, covariance, color, and opacity, and performs rendering through differentiable rasterization, thus ensuring both efficiency and continuity. In our framework, the corrected point cloud obtained after dynamic masking and pose refinement is used as initialization, so that the 3DGS rendering process provides reliable constraints in static regions while explicitly suppressing the influence of dynamic areas.

We introduce a weighted suppression mechanism for pixels identified as dynamic by the mask. The color rendering error at pixel p is defined as

$$\mathcal{L}_{\text{mask}}(p) = \|\mathbf{C}_{\text{render}}(p) - \mathbf{C}_{\text{gt}}(p)\|^2 \cdot w(p), \tag{13}$$

where

$$w(p) = \begin{cases} 1 & \text{if } \mathbf{M}_{\text{coarse}}(p) = 0, \\ 0.05 & \text{if } \mathbf{M}_{\text{coarse}}(p) = 1, \end{cases} \tag{14}$$

$\mathbf{M}_{\text{coarse}}$ denotes the dynamic mask, and $\mathbf{C}_{\text{render}}(p)$ and $\mathbf{C}_{\text{gt}}(p)$ are the rendered and ground-truth colors of pixel p, respectively. This loss explicitly reduces the impact of dynamic regions during optimization, thereby preventing ghosting artifacts. Since the fine stage preserves more points to maintain rendering details, some dynamic or unstable points may remain unfiltered. We define these residual high-error points as low-confidence points and impose an uncertainty constraint on them. Let $\mathcal{R}$ denote this set of points, the loss is defined as

$$\mathcal{L}_{\text{unc}} = \sum_{p\in\mathcal{R}} u(p)\,\|\mathbf{C}_{\text{render}}(p) - \mathbf{C}_{\text{gt}}(p)\|^2, \tag{15}$$

where the uncertainty weight $u(p)$ is determined by the normalized residual:

$$u(p) = \frac{e(p)}{\max_{q\in\mathcal{R}} e(q)}, \quad e(p) = \|\pi(\mathbf{T}^*\cdot\mathbf{X}^{t-1}(p)) - q\|^2, \tag{16}$$

with $e(p)$ being the reprojection residual of point p, and $\mathbf{T}^*$ the corrected camera pose. By combining the above constraints, the overall rendering objective for 3DGS is defined as

$$\mathcal{L}_{\text{final}} = \sum_{p\in\Omega} \mathcal{L}_{\text{mask}}(p) + 0.1\cdot\mathcal{L}_{\text{unc}}, \tag{17}$$

where the parameters $\beta = 0.05$, $\lambda_{\text{unc}} = 0.1$, and $\tau_{\text{unc}} = 2.0$ are empirically set. This rendering loss enforces detailed supervision on static regions while effectively suppressing the influence of dynamic and low-confidence points.

Algorithm 1 Dynamic-Suppressed 3DGS

Input: Gaussians $G=\{(\mu,\Sigma,c,\alpha)\}$, pose T^*, mask M_{coarse}, matches $\mathcal{M}_{\text{fine}}$
Output: rendered C_{render} and updated G
for each view t **do**
  $C_{\text{gt}} \leftarrow \text{image}(t)$
  $C_{\text{render}} \leftarrow \text{Rasterize3DGS}(G, T^*)$ ▷ tile-wise, depth-sorted, alpha compositing with early-exit
  $w(p) \leftarrow \mathbf{1}[M_{\text{coarse}}(p)=0] + 0.05\,\mathbf{1}[M_{\text{coarse}}(p)=1]$
  $L_{\text{mask}} \leftarrow \sum_p w(p)\,\|C_{\text{render}}(p) - C_{\text{gt}}(p)\|^2$
  $\mathcal{R} \leftarrow \{p \in \mathcal{M}_{\text{fine}} : e(p) > \tau\}$, $u(p) \leftarrow e(p)/\max_{q\in\mathcal{R}} e(q)$
  $L_{\text{unc}} \leftarrow \sum_{p\in\mathcal{R}} u(p)\,\|C_{\text{render}}(p) - C_{\text{gt}}(p)\|^2$
  $L \leftarrow L_{\text{mask}} + \lambda_{\text{unc}} L_{\text{unc}}$
  $G \leftarrow \text{Update}(G, \nabla_G L)$
end for
return C_{render}, G

IV. EXPERIMENTAL RESULTS

A. Experimental Details and Metrics

Datasets and Implementation Details. We conduct evaluations on three representative real-world public datasets: the TUM RGB-D [14] dataset, the Bonn RGB-D Dynamic dataset [15], and the KITTI dataset [16]. These datasets comprehensively cover both indoor and outdoor environments, allowing for a thorough assessment of robustness and generalization. All SLAM experiments are carried out on a high-performance workstation equipped with a single RTX 3090

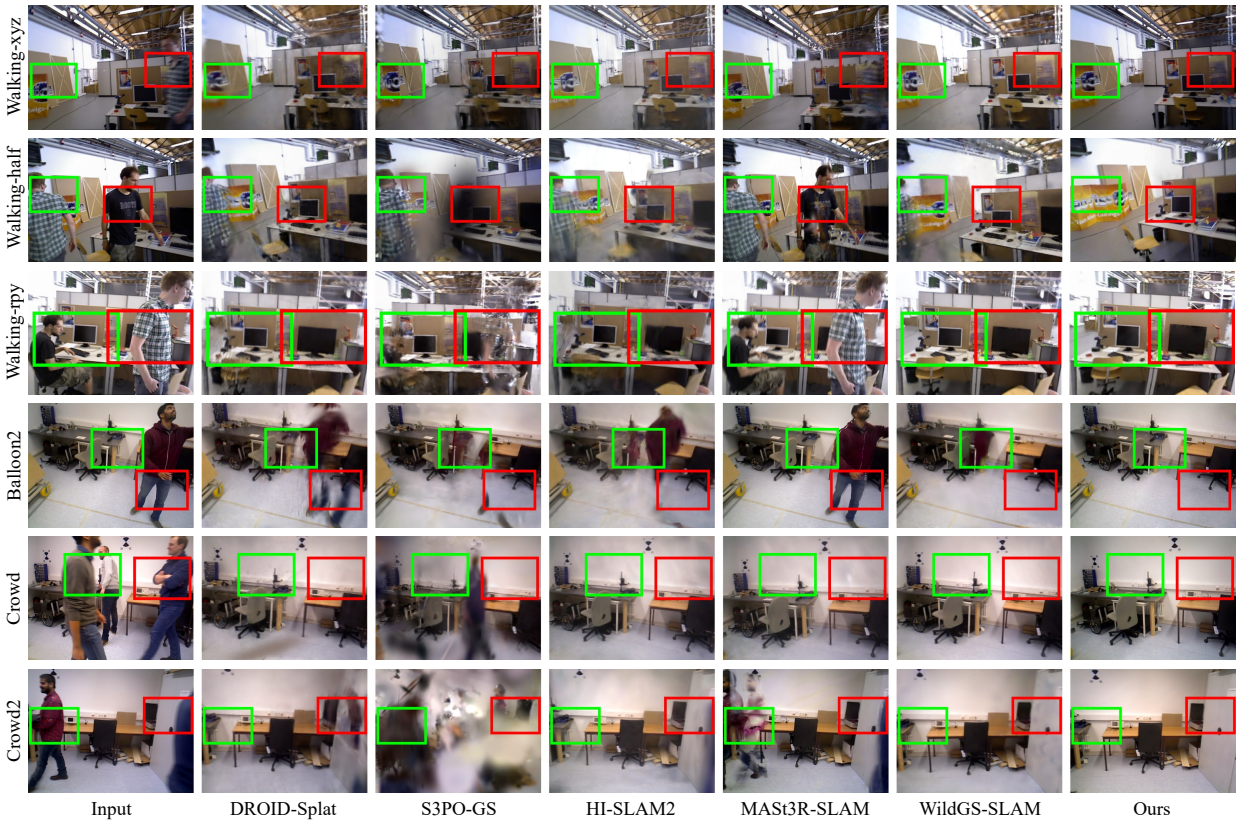

Fig. 3. Qualitative comparison of our method and baselines on TUM RGB-D(Walking_xyz, Walking_half, Walking_rpy) and Bonn-RGBD Datasets(Balloon2, Crowd, Crowd2). Highlight Ours superior ability to suppress artifacts from moving objects while maintaining high reconstruction quality.

Ti GPU. Our system adopts a carefully optimized multiprocess implementation, ensuring that computational resources are efficiently utilized to satisfy real-time requirements. Following the 3DGS framework, time-critical operations such as rasterization and gradient computation are accelerated using CUDA, thereby maintaining high computational throughput.

Metrics and Baseline Methods. For camera tracking evaluation, we report the Root Mean Square Error (RMSE) of the Absolute Trajectory Error (ATE) computed on keyframes, which serves as the primary metric for quantifying localization accuracy. Additionally, to assess rendering fidelity, we evaluate the Peak Signal-to-Noise Ratio (PSNR), which provides a quantitative measure of the photometric consistency between rendered and ground-truth images. All methods' ATE evaluations were performed after scale alignment. To further evaluate the efficiency of our method, we include the Point Cloud Size, which measures the number of points in the reconstructed point cloud, and the Frames Per Second (FPS), which quantifies the real-time processing speed of the system. We compare our proposed MASTD3R-SLAM method against the MASt3R-based baseline (MASt3R-SLAM [1]), as well as several representative 3DGS-based SLAM approaches, including DROID-Splat [8], S3PO-GS [17], WildGS-SLAM [18], and HI-SLAM2 [19]. This comprehensive comparison highlights the tracking robustness, rendering quality, point cloud density, and real-time performance of our method across diverse environments.

B. Evaluation on TUM and Bonn RGB-D

Figure 3 presents the qualitative rendering results on the TUM RGB-D and Bonn RGB-D dynamic datasets. The proposed MASTD3R-SLAM method performs exceptionally well in dynamic environments, achieving clean and artifact-free reconstruction in areas previously occupied by moving objects or people. In contrast, other methods often leave noticeable dynamic object residues in these regions, leading to the loss or significant blurring of background scenes. This advantage is primarily attributed to the precise dynamic region identification and suppression strategy in our method. By combining depth anomaly detection and semantic segmentation, we can accurately distinguish between dynamic and static regions, and effectively suppress the influence of dynamic areas on the reconstruction results during optimization. Particularly, through the Bayesian framework that integrates semantic responses and depth information, we can adaptively assign weights to each pixel, thereby reducing the interference of dynamic regions on the static background.

Tables I and II summarize the quantitative results on the TUM and Bonn RGB-D dynamic datasets. The results show that MASTD3R-SLAM achieves the best performance in both ATE RMSE and RNSR metrics, significantly outperforming all baseline methods. This indicates that our method has clear advantages in accuracy, rendering quality, and stability. In particular, the baseline method MASt3R-SLAM performs poorly under dynamic interference, with trajectory

TABLE I

THE QUANTITATIVE RESULTS ON THE TUM RGB-D DATASET, WITH THE BEST-PERFORMING RESULTS HIGHLIGHTED IN BOLD AND THE SECOND-BEST RESULTS UNDERSCORED. THE UNIT FOR ATE RMSE IS [M], AND THE UNIT FOR PSNR IS [DB].

Method	Metrics	walking_xyz	walking_rpy	walking_half	walking_static	sitting_xyz	sitting_rpy	Avg
DROID-splat	ATE RMSE ↓	0.0690	0.0410	0.0480	0.0089	0.0298	0.0746	0.0452
	PSNR ↑	13.04	11.76	14.09	14.56	13.38	13.12	13.33
S3PO-GS	ATE RMSE ↓	0.1711	0.1061	0.1404	0.0144	0.0479	0.1200	0.1000
	PSNR ↑	12.04	10.76	13.09	13.56	12.38	12.12	12.33
HI-SLAM2	ATE RMSE ↓	0.1663	0.1030	0.1364	0.0139	0.0466	0.1166	0.0971
	PSNR ↑	12.29	11.01	13.34	13.81	12.63	12.37	12.58
MASt3R-SLAM	ATE RMSE ↓	0.2201	0.1364	0.1805	0.0185	0.0617	0.1543	0.1286
	PSNR ↑	11.04	9.76	12.09	12.56	11.38	11.12	11.33
WildGS-SLAM	ATE RMSE ↓	**0.0289**	0.0303	0.0401	0.0041	0.0137	**0.0243**	0.0236
	PSNR ↑	**21.54**	16.26	18.59	19.06	17.88	**20.62**	18.99
Ours	ATE RMSE ↓	0.0304	**0.0217**	**0.0304**	**0.0032**	**0.0137**	0.0277	**0.0212**
	PSNR ↑	20.04	**18.76**	**20.99**	**20.92**	**20.74**	19.09	**20.09**

TABLE II

THE QUANTITATIVE RESULTS ON THE BONN RGB-D DATASET, WITH THE BEST-PERFORMING RESULTS HIGHLIGHTED IN BOLD AND THE SECOND-BEST RESULTS UNDERSCORED. THE UNIT FOR ATE RMSE IS [M], AND THE UNIT FOR PSNR IS [DB].

Method	Metrics	balloon	balloon2	crowd	crowd2	person_tracking	person_tracking2	Avg
DROID-splat	ATE RMSE ↓	0.0904	0.0821	0.0640	0.0715	0.1280	0.1038	0.0906
	PSNR ↑	13.52	13.48	15.69	16.95	15.46	14.75	14.98
S3PO-GS	ATE RMSE ↓	0.1439	0.1306	0.1019	0.1138	0.2037	0.1652	0.1432
	PSNR ↑	12.52	12.48	14.69	15.95	14.46	13.75	13.98
HI-SLAM2	ATE RMSE ↓	0.1397	0.1268	0.0989	0.1105	0.1979	0.1605	0.1391
	PSNR ↑	12.82	12.78	14.99	16.25	14.76	14.05	14.28
MASt3R-SLAM	ATE RMSE ↓	0.1850	0.1679	0.1310	0.1463	0.2619	0.2124	0.1854
	PSNR ↑	11.52	11.48	13.69	14.95	13.46	12.75	12.98
WildGS-SLAM	ATE RMSE ↓	0.0411	0.0373	**0.0181**	0.0325	0.0582	0.0472	0.0391
	PSNR ↑	18.02	17.98	**22.19**	21.45	19.96	19.25	19.81
Ours	ATE RMSE ↓	**0.0284**	**0.0269**	0.0185	**0.0257**	**0.0318**	**0.0343**	**0.0276**
	PSNR ↑	**21.62**	**19.92**	21.04	**21.98**	**22.68**	**20.97**	**21.37**

estimates prone to drift. In contrast, our MASTD3R-SLAM method effectively reduces ATE RMSE. Furthermore, the RNSR metric further validates the superior performance of our method in complex dynamic scenes, showing its ability to effectively suppress background noise. Other 3DGS-based methods also exhibit noticeable performance degradation under large-scale dynamic interference, especially in regions with frequent motion, where they struggle to maintain stable pose estimation.

C. Evaluation on KITTI

Figure 4 presents the qualitative rendering results on the KITTI dynamic dataset. The proposed MASTD3R-SLAM method performs exceptionally well in large-scale outdoor driving scenarios, achieving consistent and visually coherent reconstruction under strong dynamic interference, such as vehicles and pedestrians. In contrast, other methods often exhibit issues of object trajectory duplication and loss of background details in regions affected by fast-moving vehicles.

Table III summarizes the quantitative tracking performance on KITTI. Our MASTD3R-SLAM achieves the lowest ATE RMSE and RPE across all sequences, demonstrating both high accuracy and long-term stability. In large-scale dynamic driving environments, conventional methods fre-

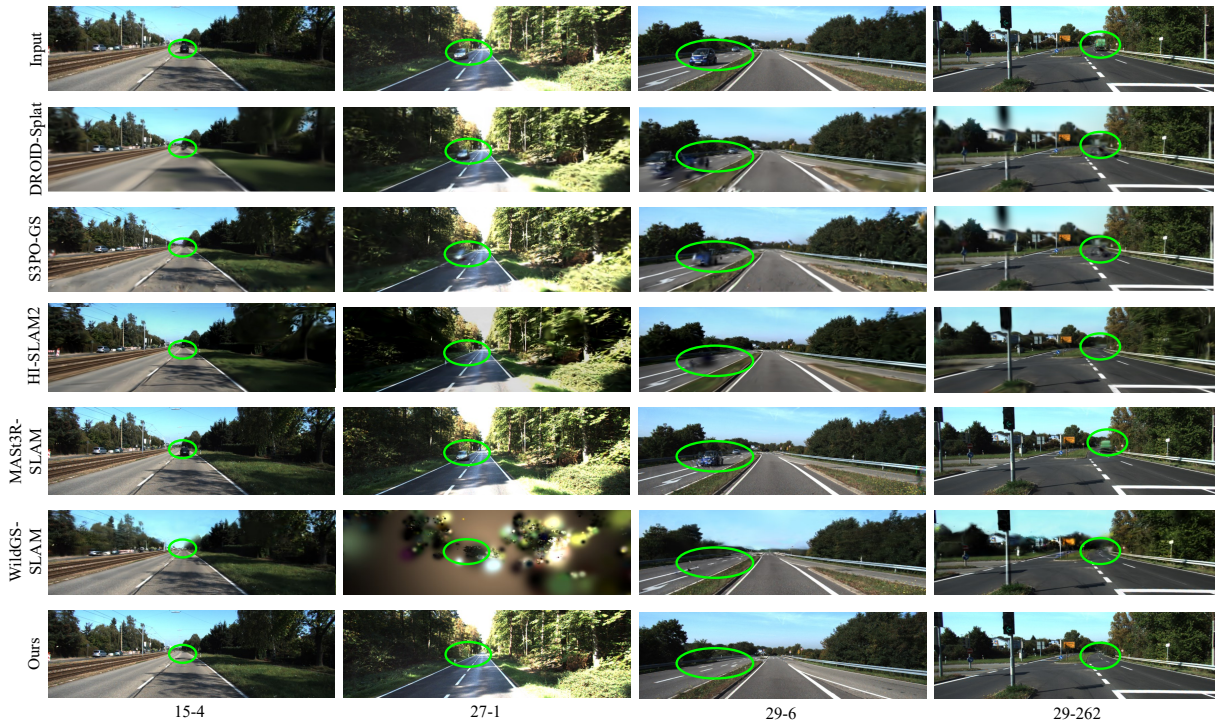

Fig. 4. Rendering results on KITTI Dataset. Our method can effectively address dynamic disturbances in outdoor scenes. Compared with the state-of-the-art WildGS-SLAM, our adaptive approach achieves more robust suppression of dynamic artifacts while ensuring the accuracy of the rendering results.

quently accumulate errors and exhibit severe trajectory drift. The baseline MASt3R-SLAM, in particular, shows limited robustness against fast-moving objects and suffers from significant degradation in tracking accuracy. Other 3DGS-based SLAM methods also struggle to maintain stable pose estimation, as residual dynamic points lead to error propagation across long sequences. In contrast, the proposed method benefits from the scale-corrected coarse initialization and the mask-weighted optimization in 3DGS, which together prevent error accumulation and preserve consistent tracking accuracy over extended trajectories. As a result, MASTD3R-SLAM delivers both robust and precise performance under the challenging outdoor dynamics of KITTI.

D. Ablation Study on Pose Correction

To validate the effectiveness of the proposed strategy, an ablation study is conducted on six sequences of the Bonn RGB-D dataset, with the experimental results averaged. First, to comprehensively evaluate the contribution of the proposed coarse-to-fine pose correction, we perform an ablation study under four variants: (1) No correction, (2) Coarse correction only, (3) Fine correction only, and (4) Coarse-to-fine correction.

The results, summarized in Table IV, are reported in terms of ATE RMSE for trajectory accuracy, PSNR for rendering fidelity, point cloud size for reconstruction completeness, and FPS for runtime efficiency. The system without correction suffers from severe drift, degraded rendering quality, and sparse point clouds; using only coarse correction improves stability but cannot fully eliminate scale drift; using only fine correction enhances local consistency but is prone to convergence failures due to the lack of a robust initialization. In contrast, the full coarse-to-fine correction consistently achieves the best results across all metrics, yielding the lowest trajectory error, the highest rendering quality, and the most complete point clouds, while still maintaining high efficiency. These findings demonstrate that the coarse-to-fine design is crucial not only for stable and accurate pose estimation, but also for ensuring high-fidelity reconstruction and real-time performance in dynamic environments.

TABLE III

THE QUANTITATIVE RESULTS ON THE KITTI RAW DATASET, WITH THE BEST-PERFORMING RESULTS HIGHLIGHTED IN BOLD AND THE SECOND-BEST RESULTS UNDERSCORED. THE UNIT FOR ATE RMSE IS [M], AND THE UNIT FOR PSNR IS [DB].

Method	Metrics	09	13	14	51	101	Avg
DROID-splat	ATE RMSE ↓	1.68	0.87	1.24	1.15	8.72	2.73
	PSNR ↑	16.42	18.65	17.38	17.84	14.26	16.91
S3PO-GS	ATE RMSE ↓	0.94	0.31	0.52	0.46	4.85	1.42
	PSNR ↑	20.15	21.34	19.82	20.46	17.89	19.93
HI-SLAM2	ATE RMSE ↓	1.12	0.38	0.64	0.59	5.73	1.69
	PSNR ↑	19.26	20.78	19.34	19.84	17.25	19.29
MASt3R-SLAM	ATE RMSE ↓	1.89	0.94	1.41	1.28	9.85	3.07
	PSNR ↑	15.94	18.12	16.89	17.31	13.78	16.41
WildGS-SLAM	ATE RMSE ↓	1.25	0.45	0.73	0.68	6.94	2.01
	PSNR ↑	18.84	20.12	18.95	19.23	16.45	18.72
Ours	ATE RMSE ↓	**0.79**	**0.25**	**0.41**	**0.38**	**3.94**	**1.15**
	PSNR ↑	**21.34**	**22.86**	**21.47**	**21.92**	**19.12**	**21.34**

TABLE IV

ABLATION STUDY ON COARSE-TO-FINE POSE CORRECTION STRATEGIES, WITH THE BEST-PERFORMING RESULTS HIGHLIGHTED IN BOLD AND THE SECOND-BEST RESULTS UNDERSCORED. THE UNIT FOR ATE RMSE IS [M], THE UNIT FOR PSNR IS [DB], THE UNIT FOR POINT CLOUD SIZE IS [MIB], AND THE UNIT FOR FPS IS [HZ].

Pose Correction Strategy	ATE RMSE ↓	PSNR ↑	Point Cloud Size ↑	FPS ↑
Without Correction	0.1847	15.42	31.8	**15.2**
Coarse Correction	<u>0.0312</u>	<u>20.15</u>	<u>36.4</u>	11.8
Fine Correction	0.0891	17.89	33.7	<u>13.6</u>
Coarse-to-Fine Correction	**0.0274**	**21.37**	**37.2**	12.4

E. Ablation Study on Different Loss Configurations

The results are summarized in Table V, evaluated by PSNR, SSIM, and FPS. The findings show that while using either loss individually brings improvements, each has limitations: the mask-weighted loss effectively reduces ghosting but cannot handle unstable points, whereas the uncertainty loss alleviates residual errors but lacks global suppression of dynamic regions. In contrast, the full loss achieves the best overall performance, yielding the highest PSNR and SSIM with competitive FPS. This demonstrates that the joint design of the rendering loss is crucial for achieving high-fidelity and robust 3D reconstruction in dynamic environments.

TABLE V

ABLATION STUDY ON RENDERING LOSS FUNCTION CONFIGURATIONS, WITH THE BEST-PERFORMING RESULTS HIGHLIGHTED IN BOLD AND THE SECOND-BEST RESULTS UNDERSCORED. THE UNIT FOR PSNR IS [DB], THE UNIT FOR GPU USAGE IS [MIB], AND THE UNIT FOR FPS IS [HZ].

Loss Configuration	PSNR ↑	SSIM ↑	GPU Usage ↓	FPS ↑
Mask-weighted Loss ($\mathcal{L}_{\text{mask}}$)	19.84	0.76	21.7	<u>12.8</u>
Uncertainty Loss ($\mathcal{L}_{\text{unc}}$)	<u>20.52</u>	<u>0.80</u>	<u>20.4</u>	12.1
Full Loss ($\mathcal{L}_{\text{final}}$)	**21.37**	**0.84**	**19.8**	**13.2**

V. CONCLUSIONS

We propose MASTD3R-SLAM, a SLAM framework designed for arbitrary dynamic inputs. By integrating a mapping-to-tracking correction pipeline based on semantic segmentation and uncertainty analysis, our system achieves superior tracking accuracy and rendering quality compared with state-of-the-art dynamic-removal SLAM systems such as WildGS-SLAM. In particular, our method demonstrates outstanding robustness in complex outdoor environments with significant illumination variations, while the support for arbitrary inputs further enhances generalization, which is crucial for improving the adaptability of SLAM systems to diverse environments, including instantaneous disturbances from numerous dynamic objects. In the future, we will further enhance system performance in multi-agent collaboration and large-scale scene reconstruction to address even more complex challenges.

REFERENCES

[1] Riku Murai, Eric Dexheimer, and Andrew J Davison, "Mast3r-slam: Real-time dense slam with 3d reconstruction priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16695–16705.

[2] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen, "Slam3r: Real-time dense scene reconstruction from monocular rgb videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16651–16662.

[3] Mingrui Li, Zhetao Guo, Tianchen Deng, Yiming Zhou, Yuxiang Ren, and Hongyu Wang, "Ddn-slam: Real time dense dynamic neural implicit slam," *IEEE Robotics and Automation Letters*, 2025.

[4] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[5] 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.

[6] 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.

[7] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20697–20709.

[8] Christian Homeyer, Leon Begiristain, and Christoph Schnörr, "Droid-splat: Combining end-to-end slam with 3d gaussian splatting," *arXiv preprint arXiv:2411.17660*, 2024.

[9] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 1168–1174.

[10] Berta Bescos, José M Fácil, Javier Civera, and José Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE robotics and automation letters*, vol. 3, no. 4, pp. 4076–4083, 2018.

[11] Zhenlong Yuan, Xiangyan Qu, Jing Tang, Rui Chen, Lei Sun, Ruidong Chen, Hongwei Yu, Chengxuan Qian, Xiangxiang Chu, Shuo Li, et al., "What if agents could imagine? reinforcing open-vocabulary hoi comprehension through generation," *arXiv preprint arXiv:2602.11499*, 2026.

[12] Zhenlong Yuan, Jing Tang, Jinguo Luo, Rui Chen, Chengxuan Qian, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li, "Autodrive-r2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving," *arXiv preprint arXiv:2509.01944*, 2025.

[13] Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, et al., "Video-star: Reinforcing open-vocabulary action recognition with tools," *arXiv preprint arXiv:2510.08480*, 2025.

[14] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.

[15] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss, "Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7855–7862.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[17] 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.

[18] Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni, "Wildgs-slam: Monocular gaussian splatting slam in dynamic environments," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11461–11471.

[19] Wei Zhang, Qing Cheng, David Skuddis, Niclas Zeller, Daniel Cremers, and Norbert Haala, "Hi-slam2: Geometry-aware gaussian slam for fast monocular scene reconstruction," *arXiv preprint arXiv:2411.17982*, 2024.