

3D Gaussian Splatting in Robotics: A Survey

Siting Zhu, Guangming Wang, Dezhi Kong, Hesheng Wang *Senior Member, IEEE*

Abstract—Dense 3D representations of the environment have been a long-term goal in the robotics field. While previous Neural Radiance Fields (NeRF) representation have been prevalent for its implicit, coordinate-based model, the recent emergence of 3D Gaussian Splatting (3DGS) has demonstrated remarkable potential in its explicit radiance field representation. By leveraging 3D Gaussian primitives for explicit scene representation and enabling differentiable rendering, 3DGS has shown significant advantages over other radiance fields in real-time rendering and photo-realistic performance, which is beneficial for robotic applications. In this survey, we provide a comprehensive understanding of 3DGS in the field of robotics. We divide our discussion of the related works into two main categories: the application of 3DGS and the advancements in 3DGS techniques. In the application section, we explore how 3DGS has been utilized in various robotics tasks from scene understanding and interaction perspectives. The advance of 3DGS section focuses on the improvements of 3DGS own properties in its adaptability and efficiency, aiming to enhance its performance in robotics. We then summarize the most commonly used datasets and evaluation metrics in robotics. Finally, we identify the challenges and limitations of current 3DGS methods and discuss the future development of 3DGS in robotics.

Index Terms—3D Gaussian Splatting, Robotics, Scene Understanding and Interaction, Challenges and Future Directions

1 INTRODUCTION

The advent of Neural Radiance Fields (NeRF) [1] has promoted the development of robotics, particularly enhancing robots' capabilities in perception, scene reconstruction and interaction with their environments. However, this implicit representation suffers from its inefficient per-pixel raycasting rendering for optimization. The emergence of 3D Gaussian Splatting (3DGS) [2] addresses this inefficiency by its explicit representation and achieves high-quality and real-time rendering through splatting. Specifically, 3DGS models the environment using a set of Gaussian primitives with learnable parameters, providing explicit representation of scenes. In rendering process, 3DGS employs splatting [3] to project 3D Gaussians into 2D image space given camera poses and applies tile-based rasterizer for acceleration, enabling real-time performance. Therefore, 3DGS has greater potential to promote the performance and expand the capabilities of robotic systems.

With the debut of 3DGS in 2023, several survey papers [4], [5], [6], [7], [8], [9] have been published to show the developments in this area. Chen *et al.* [4] presented the first survey of 3DGS. This survey describes the recent developments and critical contributions of 3DGS methods. Fei *et al.* [5] introduced a unified framework for categorizing existing works in 3DGS. Wu *et al.* [7] presented a survey including both traditional splatting methods and recent neural-based 3DGS methods, showing the development in

splatting techniques of 3DGS. Bao *et al.* [9] provided a more detailed classification based on technologies of 3DGS. Additionally, Dalal *et al.* [6] focused on the tasks of 3D Reconstruction in 3DGS and Bagdasarian *et al.* [8] summarized 3DGS-based compression methods, which demonstrates the weaknesses and advantages of 3DGS in specific domains.

However, existing 3DGS surveys either provide broad categorizations of 3DGS works or focus on real-time view synthesis of 3DGS, which lacks detailed summaries in the field of robotics. Therefore, in this paper, we provide a comprehensive summary and detailed classification of 3DGS in robotics. We introduce application of 3DGS in robotics and provide a detailed classification of related works regarding 3DGS-based robotic applications. Moreover, we summarize the potential solutions to enhance 3DGS representation for robotic systems. Finally, we show performance evaluation of 3DGS-based works and discuss the future development of 3DGS in robotics. The overall framework of our survey is demonstrated in Fig. 1.

Section 2 provides a brief background on core concepts and mathematical principles of 3DGS. Section 3 categorizes various application directions of 3DGS in robotics and presents a detailed classification of related works in specific directions. Section 4 discusses various advances that improve the representation of 3DGS, aiming to enhance its capability for robotics tasks. Additionally, in Section 5, we summarize the datasets and evaluation metrics that are used in applications of 3DGS in robotics. Moreover, this section also compares the performance of existing methods in different robotic directions. In Section 6, we discuss the challenges and future directions of 3DGS in robotics. Finally, Section 7 presents the conclusion of this survey.

2 BACKGROUND

2.1 3DGS Theory

3DGS [2] introduces an explicit radiance field representation for real-time and high-quality rendering of 3D scenes.

This work was supported in part by the Natural Science Foundation of China under Grant 62225309, 62073222, U21A20480 and 62361166632. (Corresponding author: Hesheng Wang.)

Siting Zhu is with the Department of Automation, Shanghai Jiao Tong University, Shanghai. (email: zhusiting@sjtu.edu.cn)

Guangming Wang is with the University of Cambridge.

Dezhi Kong is with the Department of Naval Architecture and Ocean Engineering (NAOE), Shanghai Jiao Tong University, Shanghai.

Hesheng Wang is with Department of Automation, Key Laboratory of System Control and Information Processing of Ministry of Education, Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai Jiao Tong University, Shanghai, China. (email: wanghesheng@sjtu.edu.cn)

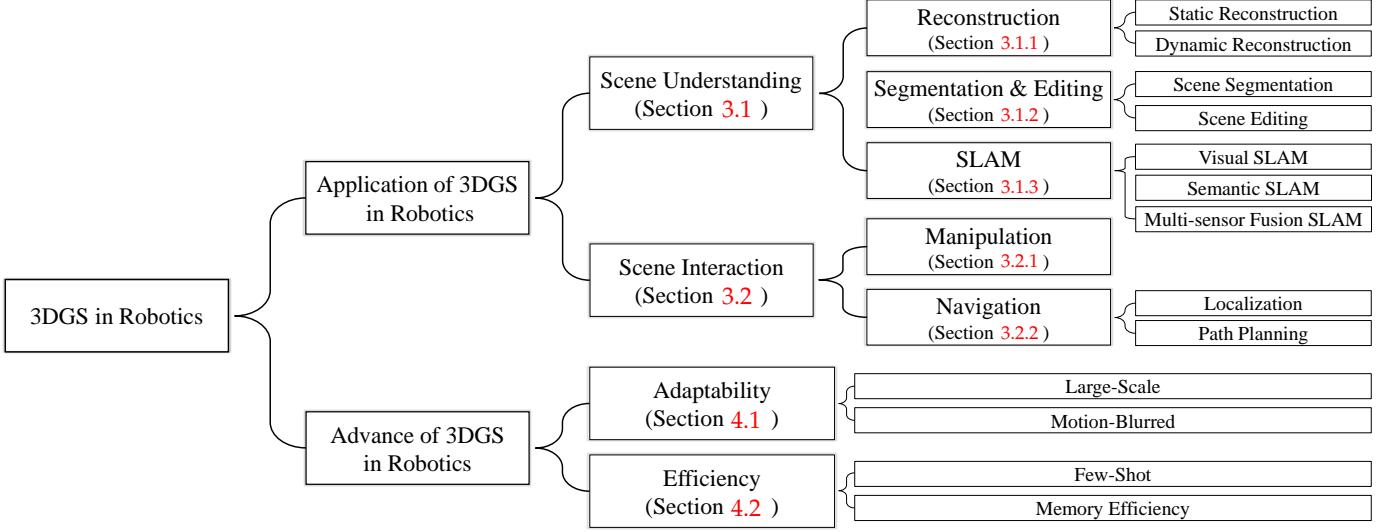


Fig. 1: A taxonomy of 3D Gaussian Splatting (3DGS) in Robotics.

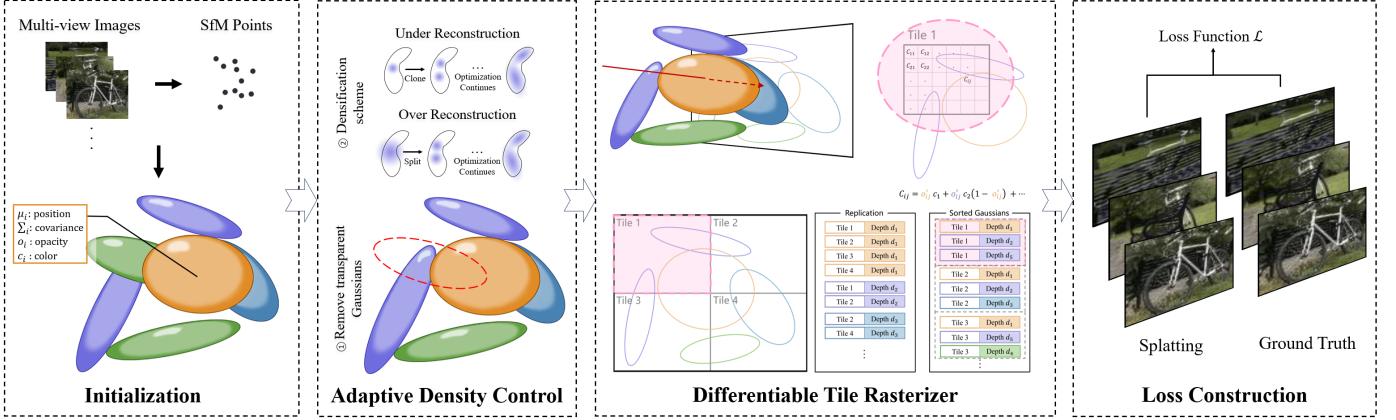


Fig. 2: An illustration of the forward process of 3DGS.

This technique models the environment using a set of 3D Gaussian primitives, denoted as $G = \{g_1, g_2, \dots, g_n\}$. Each 3D Gaussian g_i is parameterized by a set of learnable properties, including the 3D center position $\mu_i \in \mathbb{R}^3$, 3D covariance matrix $\Sigma_i \in \mathbb{R}^{3 \times 3}$, opacity $o_i \in [0, 1]$, and color c_i represented by spherical harmonics (SH) coefficients [10] for view-dependent appearance. These properties allow for the spatial compact representation of an individual Gaussian as:

$$g_i(x) = e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}. \quad (1)$$

Here, covariance matrix Σ of 3D Gaussian is analogous to describing the configuration of an ellipsoid, and is computed as $\Sigma = RSS^T R^T$. S is a scaling matrix and R represents the rotation.

The input to 3DGS is multi-view images which produce a sparse point cloud through Structure-from-Motion (SfM) [11]. This point cloud is then utilized to create initial 3D Gaussians. Then, 3D Gaussian representation is optimized by updating properties of 3D Gaussian primitives via comparing ground-truth images and rendering images. Here, rendering results are generated by splatting process, along with designed differentiable tile-based rasterizer for acceleration. Additionally, 3DGS employs adaptive density control to manage the number of 3D Gaussians used for

scene representation. Fig. 2 shows forward process of 3DGS. In what follows, splatting process in 3DGS is first described, which enables fast rendering speed. Subsequently, the loss function used in 3D Gaussian optimization is then described in detail. Moreover, adaptive density control technique employed in 3DGS is introduced. This technique allows for obtaining a denser set of Gaussians from an initially sparse set, enabling an improved scene representation.

Splatting. In this process, 3D Gaussians are projected to 2D image space for rendering. Given the viewing transformation W , the projected 2D covariance matrix in camera coordinates is computed as $\Sigma' = JW\Sigma W^T J^T$, where J is the Jacobian of the affine approximation of the projective transformation. Therefore, the final pixel color C can be computed by α -blending 3D Gaussian splats that overlap at a given pixel, with the Gaussians sorted in depth order:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

where the final opacity α_i is formulated as:

$$\alpha_i = o_i \exp \left(-\frac{1}{2}(x' - \mu'_i)^T \Sigma'^{-1} (x' - \mu'_i) \right), \quad (3)$$

where x' and μ'_i are coordinates in the projected space.

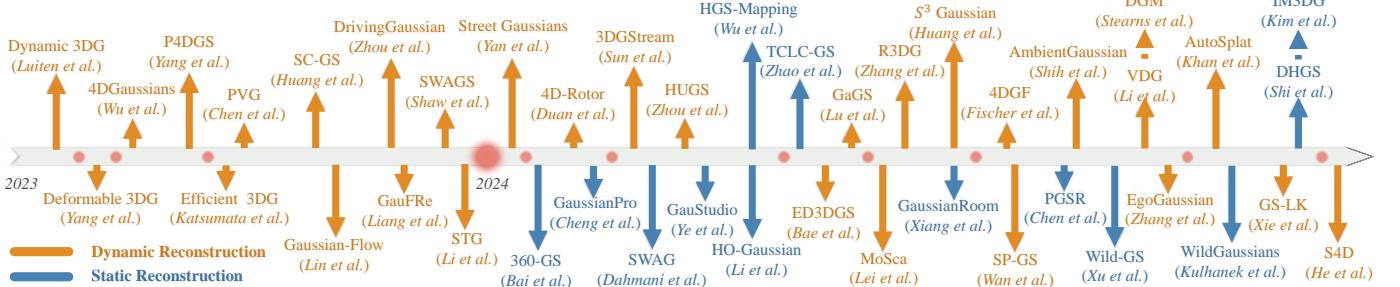


Fig. 3: Chronological: 3DGS for Scene Reconstruction. The red dots in the figure represent months, which also applies to the other figures.

Loss Function. For the optimization of 3D Gaussian properties, the loss is constructed as the difference between splatting image and ground truth image. Specifically, the loss function is a combination of \mathcal{L}_1 loss and a D-SSIM term:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}, \quad (4)$$

where λ is a weighting coefficient and is set to 0.2 in 3DGS.

Adaptive Density Control. In the process of Gaussian optimization, 3DGS employs periodic adaptive densification for detailed reconstruction. This strategy focuses on areas with missing geometric features or regions where Gaussians are over-expanded, both exhibiting large view-space positional gradients. For under-reconstructed areas, small Gaussians are cloned and moved towards the positional gradient direction. In over-reconstructed regions, large Gaussians with high variance are split into two smaller ones. Additionally, 3DGS removes Gaussians that are virtually transparent, with opacity less than a specific threshold.

In conclusion, 3DGS introduces a novel explicit radiance field representation using 3D Gaussian primitives, which offers a compact, efficient, and flexible approach to model 3D scenes. The use of 3D Gaussian representation, combined with the splatting process, differentiable optimization, and adaptive density control, makes 3DGS a powerful tool for real-time and high-quality rendering of complex 3D environments.

3 APPLICATION OF 3DGS IN ROBOTICS

The advantages of 3DGS, including its explicit radiance field representation and fast rendering capability, make it an attractive representation option for robotics applications. These properties are crucial for achieving comprehensive scene understanding in robotics and executing particular tasks through interaction with the environment.

3.1 Scene Understanding

3.1.1 Reconstruction

Scene reconstruction in robotics refers to the process of constructing a 3D representation of the environment using sensor data. 3DGS emerges as a promising scene representation to precisely model geometry and appearance information of the environment, enabling photo-realistic scene reconstruction. 3DGS-based reconstruction can be categorized into *static reconstruction* and *dynamic reconstruction*, depending on whether dynamic objects in the environment are modeled. We present a timeline of related works in Fig. 3.

Static Reconstruction. Static scene reconstruction focuses on environments with time-invariant geometry properties. Considering that indoor and outdoor environments present distinct challenges in reconstruction tasks due to differences in scale, structural complexity, and lighting conditions, we split our discussion into indoor and outdoor scene reconstruction. Moreover, the effectiveness of static reconstruction methods heavily relies on the input sensor data, appearance and geometry modeling of the scenes, and the chosen scene representation. These factors play a crucial role in determining the accuracy, level of detail, and efficiency of the reconstruction process. Therefore, we categorize existing static reconstruction methods based on four criteria: (i) the type of input sensors, (ii) the approach to appearance modeling, (iii) the use of geometric normal constraint, (iv) scene representation methods, as presented in Table 1. Particularly, appearance modeling includes illumination and exposure, as changing illumination conditions of the scenes and camera exposure problem lead to inaccurate appearance modeling of the environment. Scene representation in reconstruction can be classified into using a single 3D Gaussian representation (Single Gaussian), combining multiple Gaussian representations for joint scene modeling (Hybrid Gaussian), and integrating 3D Gaussian with neural network (Neural Field Gaussian).

Indoor scenes exhibit defined spatial layouts and rich textures. 360-GS [12] introduces a layout-guided regularization on 3D Gaussians to reduce floaters caused by under-constrained regions. For modeling input panoramas, it projects 3D Gaussians onto the tangent plane of the unit sphere and then maps them to the spherical projections. GaussianRoom [13] incorporates a learnable neural SDF field to guide the densification and pruning of Gaussian, achieving accurate surface reconstruction. Kim et al. [14] employ a hybrid representation for indoor scene reconstruction, using meshes to represent the room layout and 3D Gaussians for the reconstruction of other objects.

Outdoor scenes present additional challenges, such as varying illumination and scale variations. SWAG [16] models the appearance of scenes using a learnable Multi-Layer Perceptron (MLP) network to address varying lighting conditions in outdoor environments. Wild-GS [22] utilizes hierarchical appearance modeling by extracting high-level 2D appearance information and constructing the positional-aware local appearance embedding to handle the complex appearance variances across different views. WildGaussians [23] enhances 3DGS with a per-Gaussian trainable

TABLE 1: Categorization: 3DGS for Static Reconstruction.

Methods	Input			Appearance Modeling		Geometry	Scene Representation		
	Monocular	Multi-camera	LiDAR	Illumination	Exposure	Normal	Single Gaussian	Hybrid Gaussian	Neural Field Gaussian
Indoor Reconstruction									
360-GS [12]	✓					✓	✓		
GaussianRoom [13]	✓					✓			
IM3DG [14]	✓					✓			
Outdoor Reconstruction									
GaussianPro [15]	✓					✓	✓		
SWAG [16]	✓					✓			
GauStudio [17]	✓					✓		✓	
HGS-Mapping [18]	✓			✓		✓		✓	
HO-Gaussian [19]		✓			✓				✓
TCLC-GS [20]	✓			✓					✓
PGSR [21]	✓				✓	✓			
Wild-GS [22]	✓				✓	✓			✓
WildGaussians [23]	✓				✓	✓			✓
DHGS [24]		✓	✓					✓	

embedding and uses a small MLP to integrate image and appearance embedding, which addresses varying illumination conditions in the wild. PGSR [21] proposes a camera exposure compensation model to address the large illumination variations in outdoor scenes.

Moreover, different parts of the scene exhibit distinct characteristics that necessitate specialized modeling. For instance, sky regions, which lack geometric details and are positioned distantly in images, cannot be effectively represented by vanilla 3DGS in world coordinates. Additionally, as the road surface is flat, applying 3D Gaussian for road modeling can cause substantial redundancy in representation. To address this challenge, HGS-Mapping [18] employs hybrid Gaussian representation, comprised of spherical Gaussian for sky modeling, 2D Gaussian plane for road surface modeling, and 3D Gaussian for roadside landscapes modeling. Similarly, GauStudio [17] utilizes spherical Gaussian map to model the sky separately from the foreground. DHGS [24] employs 2DGS [25] for road representation and applies SDF-based surface constraints for road optimization. In addition, HO-Gaussian [19] combines a grid-based volume with the 3DGS pipeline to facilitate the learning of geometric information in the sky, distant and low-texture areas. TCLC-GS [20] utilizes a hybrid representation, combining colorized 3D mesh and hierarchical octree features, to enrich the properties of 3D Gaussian representation in urban areas. GaussianPro [15] incorporates planar loss to regularize the geometry of the Gaussians in outdoor scenes.

Since real-world scenarios are complex and variable, static reconstruction requires the capability of accurately modeling the appearance and geometry information in the environment. For static scene representation, hybrid Gaussian enables a more accurate modeling of the entire scene compared to single Gaussian. This is because hybrid Gaussian can adapt different modeling approaches based on the geometric structures of different regions in the scene, while single Gaussian applies a uniform modeling method to the entire scene. Moreover, neural field Gaussian demonstrates superiority in illumination modeling compared to single Gaussian and hybrid Gaussian due to its incorporation of an MLP network, which is more conducive to modeling illumination information.

Dynamic Reconstruction. In real-world robotic applications, the presence of dynamic objects is inevitable. Vanilla 3DGS representation, primarily designed for static modeling, struggles with modeling the motions of dynamic objects as they inevitably interfere with the optimization of Gaus-

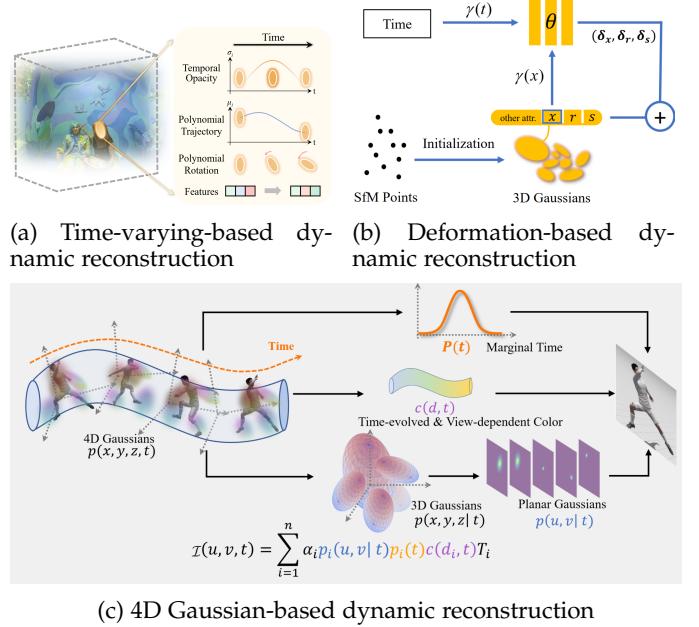


Fig. 4: An illustration of 3DGS for dynamic reconstruction. Fig. 4a, Fig. 4b, and Fig. 4c are originally shown in [45], [37], and [39], respectively.

sian parameters. Therefore, learning 3DGS-based models in dynamic scenes is crucial. For dynamic reconstruction, the key issues are how to differentiate between dynamic and static components, and how to model the dynamic objects.

Therefore, we categorize existing 3DGS-based dynamic reconstruction methods based on three criteria: (i) the type of input sensors, (ii) methods to separate static and dynamic objects in the scenes, (iii) the dynamic modeling methods, as presented in Table 2. Specifically, the separation of static and dynamic components in the scenes can be achieved by applying motion or semantic masks to detect the locations of dynamic objects in images (referred to as 2D prior), using 3D bounding boxes to identify dynamic objects (3D prior), or by simultaneously modeling both static and dynamic elements (No prior). Dynamic modeling methods can be classified into *time-varying*, *deformation-based*, and *4D Gaussian* modeling, as shown in Fig. 4.

In terms of time-varying modeling methods, each 3D Gaussian position is expressed as a function of time to model the temporal change of the position. Dynamic 3DG [36] regularizes Gaussians' motion and rotation with

TABLE 2: Categorization: 3DGS for Dynamic Reconstruction.

Methods	Input			Dynamic-Static Separation			Dynamic Modeling		
	Monocular	Multi-camera	LiDAR	3D Prior	2D Prior	No Prior	Time-varying	Deformation	4D Gaussian
PVG [26]	✓		✓			✓	✓		
DrivingGaussian [27]		✓	✓	✓	✓		✓		
Street Gaussians [28]	✓		✓	✓	✓		✓		
HUGS [29]	✓			✓	✓				
MoSca [30]	✓				✓		✓		
S^3 Gaussian [31]	✓		✓		✓			✓	
4DGF [32]	✓		✓	✓				✓	
VDG [33]		✓			✓		✓		
EgoGaussian [34]	✓				✓			✓	
AutoSplat [35]	✓		✓	✓	✓			✓	
Dynamic 3DG [36]		✓				✓	✓		
Deformable 3DG [37]	✓					✓		✓	
4DGaussians [38]	✓					✓		✓	
P4DGS [39]	✓					✓			✓
Efficient 3DG [40]	✓					✓			
SC-GS [41]	✓					✓		✓	
Gaussian-Flow [42]	✓					✓		✓	
GauFRe [43]	✓					✓		✓	
SWAGS [44]	✓					✓		✓	
STG [45]	✓					✓			
4D-Rotor [46]	✓					✓			✓
3DGStream [47]	✓					✓		✓	
ED3DGS [48]	✓					✓		✓	
GaGS [49]	✓					✓		✓	
R3DG [50]	✓					✓		✓	
SP-GS [51]	✓					✓		✓	
AmbientGaussian [52]	✓					✓		✓	
DGM [53]	✓					✓			
GS-LK [54]	✓					✓		✓	
S4D [55]	✓					✓		✓	

local-rigidity constraints. PVG [26] and VDG [33] introduce periodic vibration-based temporal dynamics by modifying the mean and opacity of vanilla 3DGS to be time-dependent functions centered around the life peak. DrivingGaussian [27] and Street Gaussians [28] leverage a composite dynamic Gaussian graph to represent multiple moving objects across time. HUGS [29] extends 3DGS to model camera exposure on dynamic scenes and utilizes optical flow prediction for dynamic separation. Efficient 3DG [40] models the temporal changes in position dynamics using Fourier approximation. STG [45] stores features instead of SH coefficients in each Gaussian to accurately encode view- and time-dependent radiance.

Deformation-based methods model motions as deformations relative to the canonical space of previous observations that represent a static field. S^3 Gaussian [31] and 4DGaussians [38] employ a multi-resolution spatial-temporal field network to represent dynamic 3D scenes following Hexplane [56] and an lightweight MLP to decode deformation. 4DGF [32] utilizes neural fields to represent sequence- and object-specific appearance and geometry variations. AutoSplat [35] estimates residual SH for each foreground Gaussian to model the dynamic appearance of foreground objects. SC-GS [41] introduces sparse control points together with an MLP for modeling scene motion. Gaussian-Flow [42] models time-dependent residuals of each Gaussian attribute by a polynomial fitting in the time domain, and a Fourier series fitting in the frequency domain. SWAGS [44] partitions the sequence into windows and trains a separate dynamic 3DGS model for each window, allowing the canonical representation to change. 3DGStream [47] employs neural transformation cache (NTC) to model the translations and rotations of 3DGS. ED3DGS [48] introduces coarse and fine temporal

embeddings to represent the slow and fast state of the dynamic scene. GaGS [49] extracts 3D geometry features and integrates them in learning the 3D deformation. GS-LK [54] introduces an analytical regularization of the warp canonical field in dynamic 3DGS by deriving a Lucas-Kanade style velocity field. S4D [55] employs partially learnable control points for local 6-DoF motion representation.

4D Gaussian methods consider the spacetime as an entirety and incorporate a time dimension into 3D Gaussian primitives, forming 4D Gaussian primitives for dynamic modeling. P4DGS [39] employs a 4D Gaussian parameterized by anisotropic ellipses that can rotate arbitrarily in space and time, as well as view-dependent and time-evolved appearance represented by 4D SH coefficients. Instead of using isoclinic quaternions for 4D rotation representation as in P4DGS [39], 4D-Rotor [46] introduces 4D rotors to characterize the 4D rotations motivated by [57], achieving decoupled spatial and temporal rotations.

Current 3DGS dynamic modeling methods have demonstrated their capability to reconstruct dynamic scenes. In small-scale scenes, these methods can reconstruct both dynamic and static components in a unified manner. However, when dealing with larger-scale autonomous driving scenarios, the computational burden of unified reconstruction becomes increasingly high. To address this issue, additional prior information is required to distinguish between dynamic and static regions, which are then reconstructed separately. The limitation of these methods lies in their reliance on additional prior information, such as 3D bounding boxes, which are not easily accessible. Consequently, the future development of dynamic reconstruction, besides enhancing the precision of dynamic modeling, is to leverage the geometry modeling of 3DGS for larger-scale dynamic reconstruction with minimal prior knowledge.

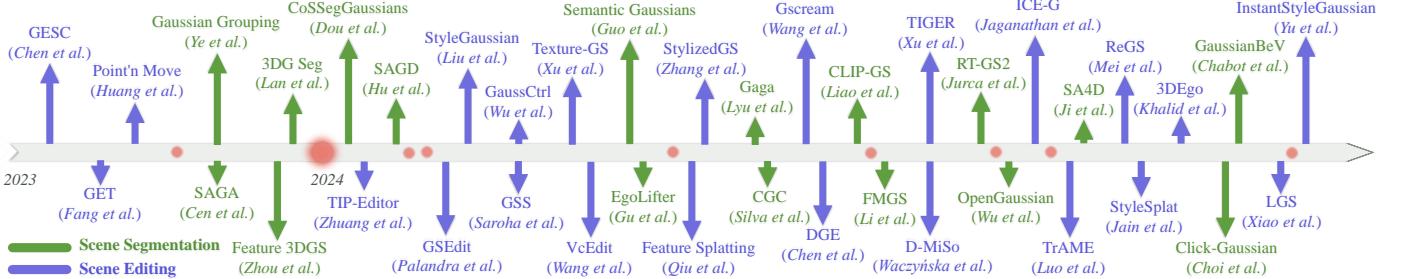


Fig. 5: Chronological: 3DGS for Scene Segmentation and Editing.

TABLE 3: Categorization: 3DGS for Scene Segmentation.

Methods	Semantic representation			Multi-view Semantic Consistency		
	Semantic Labels	Feature Embedding	MLP	Prior	Contrastive Learning	3D Spatial Association
Gaussian Grouping [58]		✓		✓		
SAGA [59]		✓				✓
Feature 3DGS [60]		✓				
3DG Seg [61]	✓		✓	✓	✓	
CoSSegGaussians [62]						
SAGD [63]	✓			✓		
Semantic Gaussians [64]	✓					✓
EgoLifter [65]		✓			✓	
Gaga [66]	✓					✓
CGC [67]	✓				✓	
CLIP-GS [68]	✓			✓		
FMGS [69]		✓				✓
RT-GS2 [70]	✓				✓	
OpenGaussian [71]	✓			✓		
Click-Gaussian [72]	✓			✓		
SA4D [73]	✓			✓		
GaussianBeV [74]	✓			✓		

3.1.2 Segmentation & Editing

The timeline of the scene segmentation and editing is illustrated in Fig. 5.

Scene Segmentation. Scene segmentation refers to the division of an observed scene into distinct components, where each component represents a different semantic category. Compared to 2D segmentation, 3D segmentation meets the operational and navigational demands of robots in real-world environments more effectively. 3DGS offers an innovative scene representation to achieve 3D semantic segmentation from 2D images. We categorize existing 3DGS segmentation methods based on two criteria: (i) the modeling method of 3DGS semantic representation, (ii) the method to maintain consistent semantic labels across multiple 2D image inputs, as shown in Table 3. Specifically, semantic representation modeling methods can be classified into three types: semantic labels, feature embedding, and MLP. The first two methods incorporate semantic labels and feature embeddings into 3D Gaussian primitives as their additional attributes, while the third method employs an MLP network for semantic modeling. Moreover, current 3DGS segmentation methods rely on SAM model [75] to provide 2D semantic labels for open vocabulary segmentation. Since the SAM model can only ensure semantic consistency within a single image, corresponding semantic regions across different images may not be consistently labeled, which leads to ambiguities in 3D semantic modeling. To address this issue, existing 3DGS works employ prior, contrastive learning, or 3D spatial association to achieve multi-view semantic consistency for 3D semantic segmentation.

Prior-based methods employ pretrained tracker [76] or input multi-view consistent 2D masks to propagate and associate masks across views. Gaussian Grouping [58] utilizes a 2D identity loss and a 3D regularization loss for Gaussian

optimization, leveraging the coherent segmentation across views. SAGD [63] incorporates a Gaussian decomposition module to address boundary roughness in 3D segmentation. CoSSegGaussians [62] designs a multi-scale spatial and semantic Gaussian features fusion module to achieve compact segmentation. CLIP-GS [68] introduces coherent semantic regularization by incorporating the semantics of adjacent views to eliminate semantic ambiguity within the same object. SA4D [73] introduces a temporal identity feature field to learn Gaussians' identity information across time for 4D segmentation. GaussianBeV [74] transforms image features into bird's-eye view (BEV) Gaussian representation for BEV segmentation in an end-to-end manner.

Contrastive learning works align features of corresponding masks while separating features of different masks in the embedding space to achieve cross-view semantic consistency. For semantic gaussian optimization, SAGA [59] introduces the correspondence loss based on the principle that pixels with a higher intersection over union (IoU) should have more similar features. OpenGaussian [71] proposes an intra-mask smoothing loss and inter-mask contrastive loss to promote feature diversity among different instances. CGC [67] presents a spatial-similarity regularization loss to enforce the spatial continuity of feature vectors, addressing misclassification in regions where the scene is not well observed. Moreover, RT-GS2 [70] fuses view-dependent features extracted from images and view-independent 3D Gaussian features obtained from contrastive learning to enhance semantic consistency across different views. Click-Gaussian [72] incorporates coarse-to-fine level features into Gaussian and clusters global features from noisy 2D segments across views to enhance global semantic consistency.

3D spatial association methods employ 3D semantic feature extraction network or label voting in spatial coordinates to achieve semantic consistency. Semantic Gaussians [64] constructs loss between 3D features obtained from 3D semantic network and mapped 3D Gaussian features from 2D features projection to optimize 3DGS semantic representation. Gaga [66] employs 3D-aware mask association process, where a 3D-aware memory bank is used to assign a consistent group ID based on the overlap ratio of shared 3D Gaussians between each mask and existing groups in the memory bank. FMGS [69] integrates semantic and language representation for scene understanding by distilling vision-language features from foundation models CLIP [77] and DINO [78] into 3DGS. Additionally, this work enhances spatial precision through a pixel alignment loss leveraging DINO features.

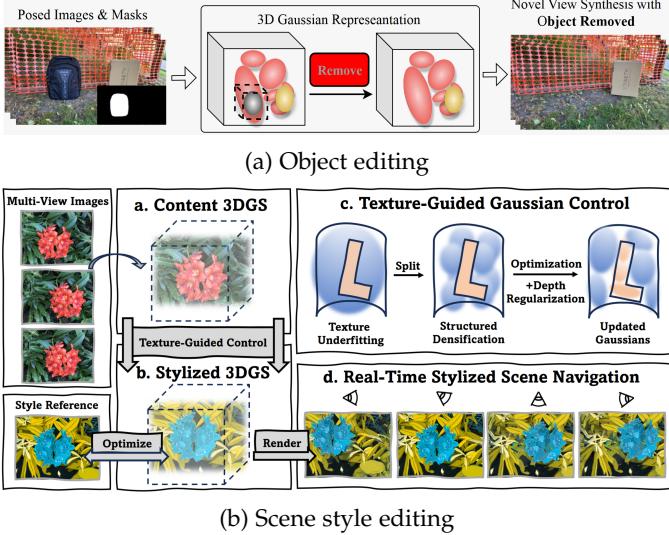


Fig. 6: An illustration of 3DGS for scene editing. Fig. 6a and Fig. 6b are originally shown in [79] and [80], respectively.

In summary, 3DGS achieves faster and more accurate 3D semantic modeling for 3D semantic segmentation compared with other scene representations. This improvement can be attributed to the Gaussian radiance field representation of 3DGS, which enables detailed scene modeling, and 3DGS's efficient rendering capability, which allows for faster optimization. In addition, by leveraging the explicit structure of Gaussians and 3DGS-based semantic segmentation results, it becomes straightforward to perform editing and manipulation of semantic objects from scene representation. This capability facilitates the application of 3DGS in various downstream robotic tasks, such as manipulation and autonomous navigation, where understanding and interacting with semantic entities in the environment is crucial.

Scene Editing. Scene editing refers to the modification of scene elements following user prompts to achieve desired effects. Edited scenes serve as valuable training resources for robots, providing a practical alternative when real-world data collection is challenging or time-consuming. The use of 3DGS representation simplifies the scene editing process by leveraging the explicit structure of Gaussians, which enables direct relocation of Gaussian elements to achieve editing. We classify related works into *object editing* and *scene style editing*, as shown in Fig. 6.

Object editing methods includes *object insertion and removal* [79], [81], [82], [83] as well as *appearance and texture editing* [84]. The challenges of object insertion and removal editing are the preservation of geometric consistency and the maintenance of texture coherence, due to the discrete properties of Gaussian primitives. GScream [79] employs cross-attention feature regularization to propagate the accurate texture in the surrounding region into the in-painted region, ensuring texture coherence of the Gaussian representation. Feature Splatting [83] incorporates visual-language embedding into 3D Gaussian for text-prompt editing. For appearance and texture editing, Texture-GS [84] establishes a connection between geometry (3D Gaussians) and appearance (2D texture map) to facilitate appearance editing.

Text-driven object editing methods [85], [86], [87], [88],

[89], [90], [91], [92], [93], [94], [95] all leverage pretrained diffusion model [96], which enables efficient use of text prompts for precise editing and incorporates rich prior knowledge to enhance the quality and coherence of the edited objects. GaussCtrl [87] employs naturally consistent depth maps obtained from ControlNet [97] to guide geometric consistency. Moreover, this work introduces self and cross-view attention of latent codes from different views for appearance consistency. VcEdit [88] and TrAME [93] employ cross-attention across maps obtained from stages of diffusion.

Scene style editing methods [80], [98], [99], [100], [101], [102], [103] generate diverse stylistic scene data based on style prompts, which can shorten data collection periods and improve the robustness of the systems trained with this data for robotic applications.

3.1.3 SLAM

We categorize 3DGS-based SLAM into *visual SLAM*, *multi-sensor fusion SLAM*, *semantic SLAM* based on the sensor used and the level of environmental understanding. 3DGS-based SLAM process is illustrated in Fig. 8. We also present timeline of the related works in Fig. 7. Compared with NeRF-based SLAM, 3DGS-based SLAM methods leverage the explicit representation of 3D Gaussian for unbounded and photo-realistic mapping.

Visual SLAM. Visual SLAM refers to the process of simultaneously performing dense color mapping and localization of a camera. Considering that accurate depth information is crucial for providing geometric supervision in constructing dense Gaussian maps, we categorize our discussion of visual SLAM into RGB-D SLAM and RGB SLAM, based on whether the input camera data contains depth information.

RGB-D SLAM methods [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118] have accurate depth input, which can provide good geometric constraints for dense mapping and pose estimation. For SLAM process, it is crucial to determine when to expand the map and how to achieve precise tracking results. SplaTAM [104] employs silhouette rendering of the existing Gaussian map to identify which portions of the scene are new content, thereby guiding the expansion of the map and the optimization of camera poses. Gaussian-SLAM [105] and NGM-SLAM [106] focus on building Gaussian submaps progressively to achieve local map optimization. MG-SLAM [107] incorporates structure prior of the scene to generate constraints that address gaps and imperfections of the reconstructed map. In addition, CG-SLAM [109] proposes a novel depth uncertainty model to build up a consistent and stable 3D Gaussian map. I²-SLAM [116] integrates the image formation process into SLAM to overcome motion blur and varying appearances. LoopSplat [117] detects loop closure online and computes relative loop edge constraints between submaps via 3DGS registration for global map consistency.

For accurate tracking, GS-SLAM [110] employs a coarse-to-fine approach to avoid drifted camera tracking caused by artifacts in images. GS-ICP SLAM [111] leverages 3D explicit representation of 3DGS for tracking, achieved by utilizing G-ICP [119] for Gaussian map matching to directly regress estimated camera poses. TAMBRIDGE [112] jointly

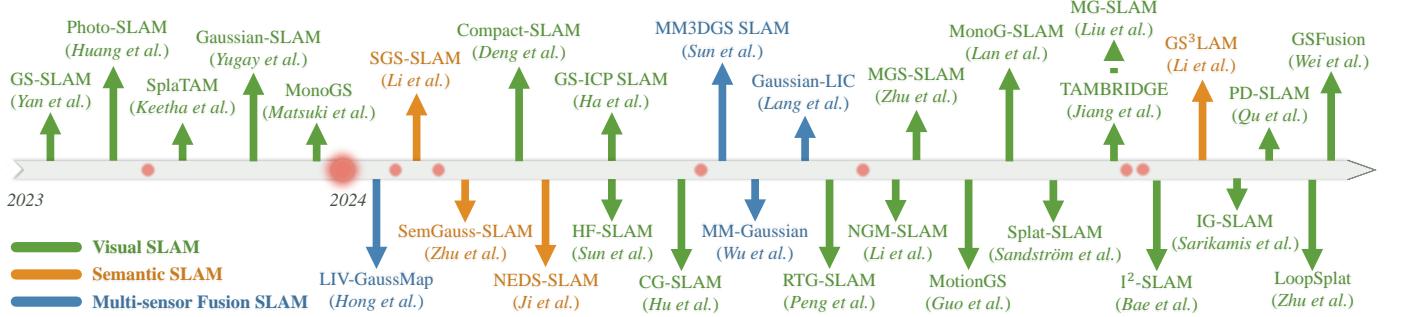


Fig. 7: Chronological: 3DGS for SLAM.

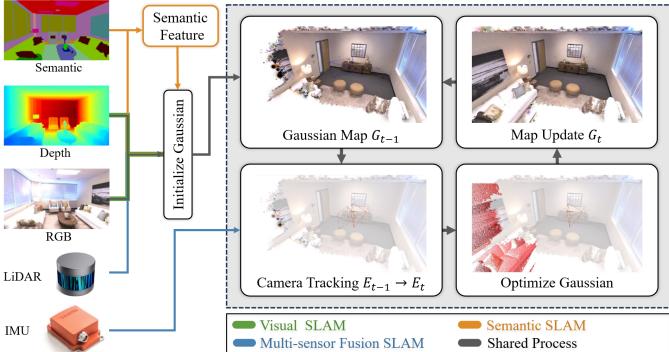


Fig. 8: An illustration of 3DGS for SLAM. Part of images are taken from SplatTAM [104].

optimizes sparse re-projection and dense rendering errors, resulting in reduced cumulative errors in tracking. While these advancements improve SLAM accuracy, detailed 3D Gaussian maps generated in the SLAM process require significant memory resources. To address this, Compact-SLAM [113] proposes a sliding window-based online masking method to remove redundant Gaussian ellipsoids for compact representation. RTG-SLAM [114] uses a single opaque Gaussian instead of multiple overlapping Gaussians to fit a local region of the surface for reducing memory consumption in mapping.

RGB SLAM methods [120], [121], [122], [123], [124], [125], [126] lack accurate depth input, so it requires additional information such as multi-view constraints or depth estimation to recover the 3D geometry of the scene. MonoGS [120] and MotionGS [121] employ a regularization term and multi-view optimization to constrain scene geometry information in scenes where depth is unknown. Photo-SLAM [122] minimizes the reprojection error between matched 2D geometric keypoints in the frame and corresponding 3D points to achieve geometric consistency. Moreover, MGS-SLAM [123] utilizes depth estimation and DPVO [127] to recover 3D geometry of the scene. MonoG-SLAM [124] employs patch graph based on DPVO [127] and CLIP [77]-based loop closure optimization to guide the estimation of the scene geometry. Splat-SLAM [125] introduces dense optical flow estimation and DSPO (disparity, scale and pose optimization) for accurate depth recovery.

The accuracy of existing 3DGS-based visual SLAM is highly dependent on precise depth information. RGB SLAM methods, due to the absence of accurate depth informa-

tion, often exhibit errors in 3DGS geometric reconstruction. Although depth estimation techniques can provide depth information for RGB SLAM, their limited accuracy leads to decreased SLAM performance. Moreover, current 3DGS-based visual SLAM systems are tested and evaluated mostly in indoor environments, as depth measurements are unreliable in outdoor scenes. Therefore, the unsolved key problem of visual SLAM is to enhance geometric reconstruction accuracy despite inaccurate depth information, thereby achieving high-precision SLAM in various environments.

Multi-sensor Fusion SLAM. Multi-sensor Fusion SLAM integrates data from different sensors to achieve accurate mapping and robust tracking in SLAM system. We classify multi-sensor fusion SLAM methods into *LiDAR-based* and *image depth estimation-based*, depending on the primary sensor modality used for 3DGS geometry reconstruction.

LiDAR-based methods use accurate point cloud obtained from LiDAR as the initial input for 3D Gaussian geometry representation. LIV-GaussMap [128] utilizes point cloud obtained from an IESKF-based LiDAR-inertial system to provide an initial Gaussian structure for the scene. Gaussian-LIC [129] performs LiDAR-Inertial-Camera odometry for tracking. LiDAR points are projected onto the corresponding image, where they are colored by querying the pixel values, and then used to initialize 3D Gaussian. MM-Gaussian [130] performs tracking using point cloud registration algorithm [131] and utilizes multi-frame camera constraints for relocalization when the tracking module fails. Image depth estimation-based methods employ monocular dense depth estimation results as geometric supervision for constructing Gaussian map. MM3DGS SLAM [132] utilizes IMU (Inertial Measurement Unit) pre-integration for initial pose estimation and constructs loss between rendering and observations for pose optimization. As depth estimation outputs a relative depth, Pearson correlation coefficient is used to compute the depth loss between estimated and rendered depth maps with actual scale information.

Semantic SLAM. Semantic SLAM incorporates semantic understanding of the environment into map construction and estimates camera pose simultaneously. Compared with visual SLAM and multi-sensor fusion SLAM, semantic SLAM enables dense semantic mapping of the scene, which is essential for downstream tasks such as navigation and manipulation. As the original 3DGS representation lacks semantic information, two methods have been developed in 3DGS-based semantic SLAM to incorporate semantics: color-based and feature-based. For color-based

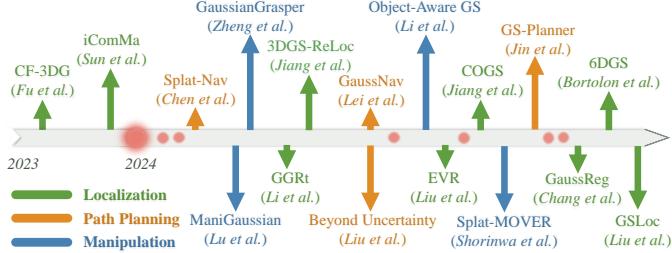


Fig. 9: Chronological: 3DGS for Scene Interaction.

method, SGS-SLAM [133] utilizes semantic color associated with the Gaussian for semantic representation. However, this color-based semantic modeling approach overlooks the higher-level information inherent in semantics. For feature-based semantic integration, SemGauss-SLAM [134] incorporates semantic embedding into 3D Gaussian for semantic representation and performs semantic-informed bundle adjustment using multi-view constraints to achieve high-accuracy semantic SLAM. NEDS-SLAM [135] proposes a fusion module that combines semantic features with appearance features to address the spatial inconsistency of semantic features, inspired by [136]. GS³LAM [137] introduces depth-adaptive scale regularization to reduce the blurring of geometric surfaces induced by irregular Gaussian scales within semantic gaussian.

3.2 Scene Interaction

Navigation and manipulation are fundamental aspects of robotic interaction with the environment. The timeline for related works is shown in Fig. 9.

3.2.1 Manipulation

Manipulation refers to using robotic arms or grippers to perform various tasks, substituting for human hands. Unlike NeRF-based manipulation approaches, 3DGS employs an explicit radiance field representation which constructs the scene and provides direct access to the positional information of objects within it. Consequently, additional object pose estimation is unnecessary in 3DGS-based manipulation. Manipulation tasks can be categorized into *single-stage* and *multi-stage* based on whether the task requires consideration of dynamic environmental changes. These changes are caused by object movement resulting from the manipulation itself, as illustrated in Fig. 10.

In single-stage manipulation, the grasping task is completed through a single, continuous motion, so the environment is considered static during the process. GaussianGrasper [138] reconstructs a 3D Gaussian feature field via efficient feature distillation to support language-guided manipulation tasks and uses rendered normals to filter out unfeasible grasp poses.

For multi-stage manipulation, the task is accomplished through a sequence of actions where the object's movement during each stage results in dynamic changes to the environment. ManiGaussian [140] proposes dynamic 3DGS framework to model the scene-level spatiotemporal dynamics and builds a Gaussian world model to parameterize distributions in dynamic 3DGS model for multi-task robotic manipulation. Object-Aware GS [141] models the

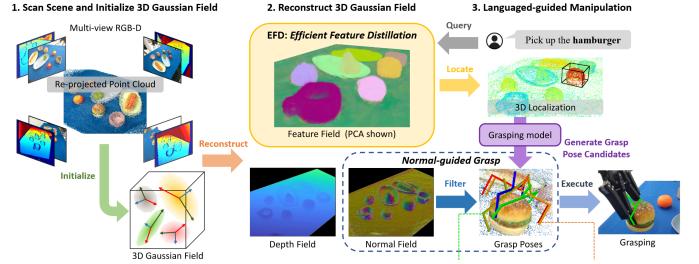


Fig. 10: An illustration of 3DGS for manipulation. Fig. 10a and Fig. 10b are originally shown in [138] and [139], respectively.

Gaussian representation as time-variant and employs object-centric dynamic updates for dynamic modeling of multi-stage manipulation. Instead of modeling dynamic changes, Splat-MOVER [139] employs a scene-editing module using 3D semantic masking and infilling to visualize the motions of the objects that result from the robot's interactions with the environment. Moreover, this work introduces GSplat representation which distills latent codes for language semantics and grasp affordance into the 3D scene for scene understanding.

3.2.2 Navigation

Navigation in robotics involves two essential and interconnected components: *localization* and *path planning*. Localization deals with the challenge of locating the robot's own position within the environment. Based on the localization result, the robot performs path planning, which refers to the process of determining an optimal route to reach the destination. 3DGS serves as a detailed scene representation for high-precision navigation tasks.

Localization. Localization refers to estimating a 6 Degree-Of-Freedom (DoF) pose (position and orientation) through the processing of sensor data. We categorize 3DGS-based localization into *known map-based localization* and *relative pose regression* based on whether the prior global map is available, as shown in Fig. 12.

For known map-based localization, iComMa [142] computes the residuals between the query image and the rendered image obtained from a prebuilt Gaussian map to optimize the camera pose. 3DGS-ReLoc [144] stores 3DGS map as a 2D voxel map with a KD-tree for efficient spatial queries and achieves relocalization by brute-force search to match the query image within the global map. Liu et al. [145] generate pseudo scene coordinates from 3DGS for initializing and enhancing scene coordinate regression. 6DGS [146] estimates the camera location by selecting a bundle of rays projected from the ellipsoid surface of Gaussian map and learning an attention map to output ray pixel correspondences for pose optimization.

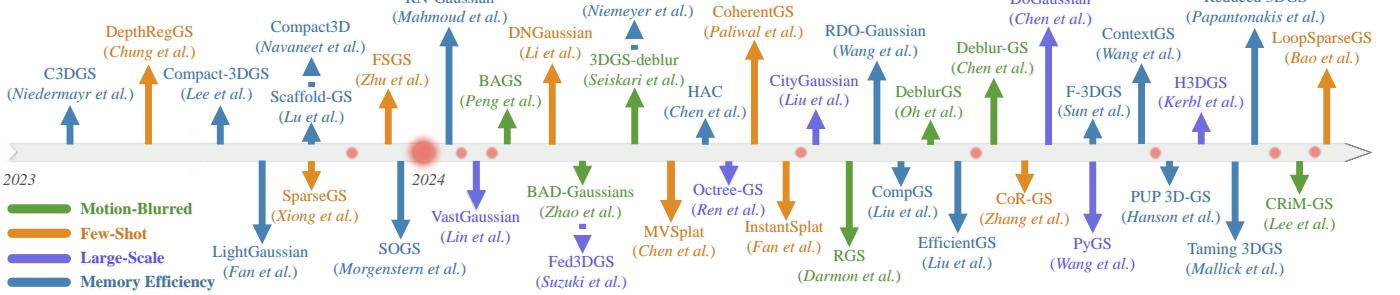
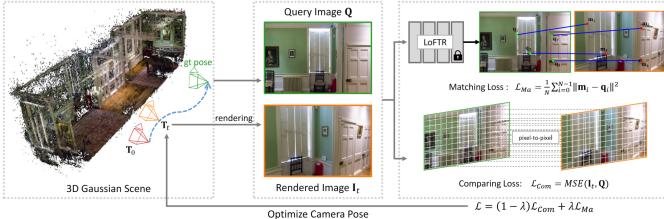
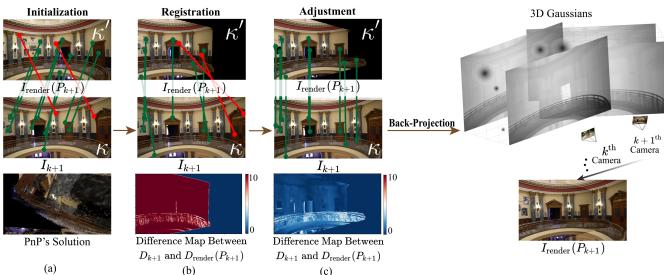


Fig. 11: Chronological Advance of 3DGS.



(a) Known map-based localization



(b) Relative Pose Regression

Fig. 12: An illustration of 3DGS for localization. Fig. 12a and Fig. 12b are originally shown in [142] and [143], respectively.

In terms of relative pose regression, CF-3DG [147] estimates the relative camera pose that can transform the local 3D Gaussian map of the last frame to render the pixels that align with the current frame. GGRt [148] designs a joint learning framework that consists of an iterative pose optimization network that estimates relative poses and a generalizable 3D-Gaussians model that predicts Gaussians. COGS [143] detects 2D correspondences between training views and the corresponding rendered images to regress relative pose. GaussReg [149] explores the registration of 3D scenes using 3DGS representation to estimate the relative pose of the scenes. GSLoc [150] incorporates an exposure-adaptive module into the 3DGS model to improve the robustness of matching between two images under domain shift, achieving accurate pose regression of images.

Path planning. We categorize path planning into active planning and non-active planning based on whether the robot actively explores the environment for reconstruction as well as planning.

In terms of active planning, GS-Planner [151] maintains a voxel map to represent unobserved volumes. This voxel map is integrated into the splatting process, enabling the exploration of unobserved space. Moreover, this work leverages physical concept of opacity to formulate a chance

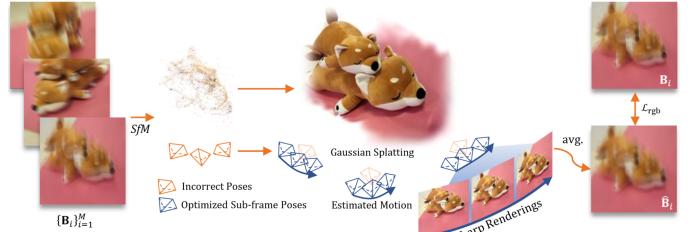


Fig. 13: An illustration of 3DGS for deblurring motion blur. Figure is originally shown in [155].

constraint for safe active planning in 3DGS maps. For non-active planning, GaussNav [152] constructs a semantic Gaussian map and transfers this map into 2D BEV grids for navigation. Furthermore, a set of descriptive images is generated by rendering the object from multiple viewpoints. These images are then used to match and identify the target position for path planning. Splat-Nav [153] generates a safe path by discretizing free space on a Gaussian map for safe navigation. This discretization is achieved through intersection tests between Gaussian ellipsoids. Beyond Uncertainty [154] leverages 3DGS map to dynamically assess collision risks at each waypoint and guides risk-aware next-best-view selection for efficient and safe robot navigation.

4 ADVANCE OF 3DGS IN ROBOTICS

As vanilla 3DGS [2] exhibits limitations, various novel variant models have made adaptive modifications to the 3DGS representation, enhancing its effectiveness in robotics applications. Fig. 11 presents a timeline of related works that focus on enhancing 3DGS properties for robotic use.

4.1 Adaptability

Enhancing the adaptability of vanilla 3DGS involves improving its performance in *large-scale* environment and its resilience to *motion blur*, enabling more effectiveness and reliability across diverse and unpredictable scenarios.

4.1.1 Motion-Blurred

Motion blur of captured images is a common challenge in robotics, primarily caused by high-speed robot movements and slow shutter speed, resulting in degraded image quality. Consequently, deblurring is essential to restore image quality and enhance visual perception for robotic systems. Deblurring methods can be divided into two categories:

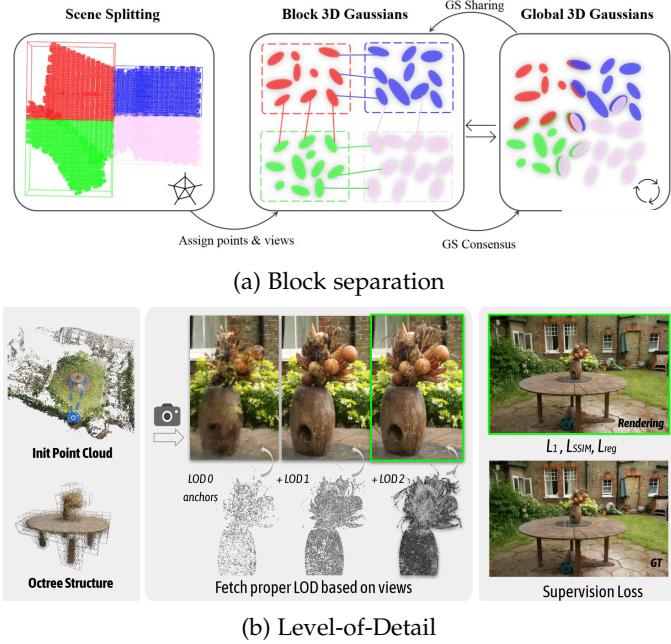


Fig. 14: An illustration of 3DGS for large-scale reconstruction. Fig. 14a and Fig. 14b are originally shown in [163] and [164], respectively.

physical modeling, which focuses on understanding and simulating the blurring process based on the formation of motion blur, and *implicit modeling*, which leverages MLP networks to directly learn the mapping between blurred and sharp images without explicitly modeling the underlying physical process, as illustrated in Fig. 13.

Physical modeling methods [155], [156], [157] simulate motion-blurred images by averaging virtual sharp images captured during the camera exposure time. These methods then construct loss between simulated and observed blurry images for optimization of Gaussian representation and camera pose, ensuring that the constructed scene is sharp and free from motion blur. RGS [158] models motion blur as a Gaussian distribution over camera poses to obtain the expected image at the given noise for optimization. 3DGS-deblur [159] simplifies motion blur modeling by approximating motion in pixel coordinates and adjusting the Gaussian means to reflect this movement. CRiM-GS [160] applies neural ODEs [161] to model the continuous camera motion trajectory during the exposure time and achieves deblurred images by rendering and averaging images from multiple poses sampled along this trajectory. In terms of implicit modeling, BAGS [162] introduces a blur proposal network to model image blur and produces a quality-assessing mask that indicates regions where blur occur.

4.1.2 Large-Scale

Vanilla 3DGS representation requires millions of 3D Gaussians for large-scale reconstruction, leading to high GPU memory demands for training, as well as long training and rendering time. To address large-scale reconstruction, existing methods [163], [164], [165], [166], [167], [168], [169] divide the scene into independent blocks for separate training (Block separation), and model the scene with hierarchical resolution levels (Level-of-Detail), as shown in Fig. 14.

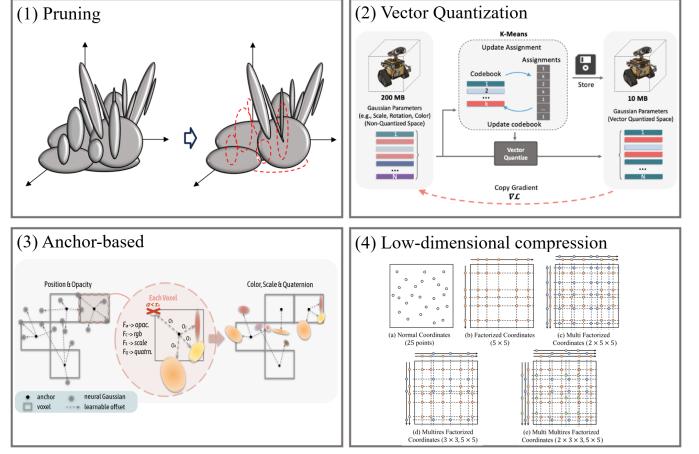


Fig. 15: An illustration of 3DGS for memory efficiency. Subplot(1) to subplot(4) are extracted from [170], [171], [172], [173], respectively.

DoGaussian [163], H3DGS [166], and CityGS [165] employs a divide-and-conquer strategy that partitions the scene into spatially adjacent blocks, which are optimized independently and in parallel. Fed3DGS [167] utilizes decentralized processing of large-scale data across millions of clients to reduce the computational load on a central server. Octree-GS [164] incorporates an octree structure for multiple-level scene representation and employs level-of-detail decomposition for efficient rendering. PyGS [168] employs a hierarchical structure of 3D Gaussians organized into pyramid levels to represent scenes at varying levels of detail for large-scale scene representation. VastGaussian [169] introduces a progressive partitioning strategy to divide a large scene into multiple cells for parallel optimization.

4.2 Efficiency

Vanilla 3DGS suffers from significant storage demands due to its explicit representation and requires extensive multi-view information for scene modeling. However, in real-world robotic applications, this substantial storage requirement is inefficient for storing global maps. Additionally, the need to obtain extensive multi-view data for reconstruction leads to data inefficiency. Therefore, improving the efficiency of 3DGS in robotics focuses on two main aspects: *memory efficiency*, which aims to reduce the large storage demands, and *few-shot*, which enhances data utilization by enabling effective reconstruction from minimal multi-view information.

4.2.1 Memory Efficiency

Explicit representation of 3D Gaussians requires a significant amount of storage space. To address this issue, existing methods primarily adopt four approaches: *pruning*, *vector quantization*, *anchor-based*, and *low-dimensional compression*, to reduce the storage space required for 3D Gaussian scene representation in robotic applications, as shown in Fig. 15.

Intuitively, reducing the number of 3D Gaussian primitives that contribute little to the scene can decrease the overall storage requirements, which refers to as pruning [170],

[174], [175], [176], [177], [178], [179], [180], [181]. Compact3DGS [170] and RDO-Gaussian [174] remove Gaussians with small scales and low opacities. LightGaussian [175], RadSplat [176], and EfficientGS [177] prune Gaussians that have low ray-contribution to the pixels across all training views. PUP 3D-GS [178] employs a sensitivity score derived from Hessian matrix of reconstruction errors, pruning Gaussians with higher spatial uncertainty and lesser contributions to reconstruction quality. Reduced 3DGS [179] identifies regions that are densely populated with Gaussian primitives and prunes Gaussians that overlap significantly with other Gaussians. RN-Gaussian [180] and Taming 3DGS [181] restrict the cloning and splitting of Gaussians based on position and correlation with existing Gaussians to reduce the number of Gaussian. Moreover, Vector Quantization methods [170], [171], [179], [182] convert Gaussian parameters into codebooks to represent Gaussian attributes compactly.

Anchor-based methods [172], [183], [184], [185] employ a set of anchor primitives to cluster related nearby 3D Gaussians and predict their attributes from anchors' attributes, resulting in memory-efficient 3D scene representation. HAC [183] leverages relations between the unorganized anchors and the structured hash grid for context modeling of anchor attributes. ContextGS [185] divides anchors into multiple levels for compact representation. Moreover, low-dimensional compression methods [173], [186] represent 3D Gaussian in 1D or 2D space for compact storage. SOGS [186] maps originally unstructured 3D Gaussians to 2D grids for reducing memory. F-3DGS [173] proposes a factorized coordinate scheme that maintains 1D or 2D coordinates in each axis or plane and generates 3D coordinates by a tensor product.

4.2.2 Few-Shot

Novel view rendering from few-shot reconstruction is challenging due to the sparse information available. In scenarios where only limited observations are present, vanilla 3DGS faces two challenges: (1) Sparse SfM points derived from few-shot images fail to represent the global geometry for Gaussian initialization. (2) Leveraging a few images for optimization can lead to overfitting. To address these challenges, few-shot methods [187], [188], [189], [190], [191], [192], [193], [194], [195] incorporates additional constraints and prior into SfM initialization and Gaussian optimization.

FSGS [187] inserts new Gaussians near existing ones based on their proximity to deal with sparse initial point sets and generates unseen viewpoints for training to address overfitting. DepthRegGS [188] employs depth-guided optimization to mitigate overfitting. CoherentGS [189] proposes depth-based initialization to achieve dense SfM points. MVSSplat [190] extracts multi-view image features to construct per-view cost volumes and predicts per-view depth maps for few-shot reconstruction. SparseGS [191] integrates depth priors with generative and explicit constraints to enhance consistency from unseen viewpoints. LoopSparseGS [192] incorporates iteratively rendered images with training images into SfM to densify the initialized point cloud. Additionally, this work leverages window-based dense monocular depth to provide precise geometric supervision for Gaussian optimization.

5 DATASETS AND PERFORMANCE EVALUATION

In this section, we present a detailed summary of the datasets commonly used in robotics. Moreover, we provide performance comparison and evaluation of key modules in robotics, including mapping and localization, perception, manipulation and navigation. These modules are fundamental and crucial to the functionality of intelligent robotic systems. By comparing the performance of these modules, we aim to offer an overview of the current state of research in robotics based on 3DGS representation. In terms of perception, we only report on scene segmentation, as there are currently no other 3DGS-based perception methods available at the time of this paper's completion.

5.1 SLAM Evaluation

Datasets. Existing 3DGS-based visual SLAM and semantic SLAM methods use indoor scene datasets for evaluation, including Replica [196], ScanNet [197], and TUM RGB-D [198]. As TUM RGB-D dataset [198] lacks semantic annotations, it is generally not used for benchmarking semantic SLAM methods. For multi-sensor fusion SLAM methods, evaluations are conducted on different outdoor scene datasets based on the specific sensors used, rather than a unified dataset. Detailed descriptions of the SLAM datasets are provided below.

- *Replica* [196] is an RGB-D dataset that contains 18 indoor scene reconstructions at room and building scales. Each scene consists of dense mesh, high-dynamic-range (HDR) textures, per-primitive semantic class and instance information.
- *ScanNet dataset* [197] contains RGB-D frames and IMU data of 2.5M views in 1513 scenes, each annotated with 3D camera poses, surface reconstructions, textured meshes, and semantic segmentations.
- *TUM RGB-D* [198] is an RGB-D dataset that consists of 39 sequences recorded in an office environment and an industrial hall. RGB-D data are captured from the Kinect. Ground truth pose estimates are obtained from the motion capture system.

Metrics. Reconstruction and tracking metrics are employed for evaluating SLAM accuracy. Specifically, *Depth L1 (cm)* metric is used for reconstruction evaluation, which is the average absolute error between ground truth depth and reconstructed depth. *ATE RMSE (cm)* [198] is tracking metric that quantifies the error between the estimated trajectory and the ground truth trajectory. For SLAM systems, real-time performance is crucial, and metric *FPS* is utilized for evaluating the time consumption of SLAM process. In addition, *mIoU (%)* is used for semantic segmentation evaluation in semantic SLAM that measures the average percentage of overlap between the predicted and ground truth areas across different classes. Moreover, as most 3DGS-based SLAM methods report the rendering performance of SLAM mapping results, we also include rendering metrics in the SLAM evaluation. Peak signal-to-noise ratio (*PSNR[dB]*), structural similarity (*SSIM*) [199], and learned perceptual image patch similarity (*LPIPS*) [200] are rendering metrics.

Results. As visual SLAM and semantic SLAM methods are mostly tested on Replica dataset [196], we present performance comparison on Replica regarding SLAM metrics, as

TABLE 4: SLAM performance comparison. The results are average of 8 scenes on Replica dataset [196]. Best results are highlighted as **first**, **second**. These notes also apply to the other tables.

	Methods	Reconstruction		Rendering		Tracking RMSE↓	Segmentation mIoU↑
		Depth L1↓	PSNR↑	SSIM↑	LPIPS↓		
Visual SLAM	GS-SLAM [100]	1.16	34.27	0.98	0.08	0.50	—
	Photo-SLAM [122]	—	34.96	0.94	0.06	0.60	—
	SplatAM [104]	0.72	34.11	0.97	0.10	0.36	—
	Gaussian-SLAM [105]	1.56	38.90	0.99	0.07	0.48	—
	MonoGo [120]	—	37.50	0.96	0.07	0.32	—
	Compact-SLAM [113]	—	34.44	0.98	0.09	0.33	—
	GS-ICP SLAM [111]	—	38.83	0.98	0.04	0.16	—
	HF-SLAM [108]	0.52	36.19	0.98	0.05	0.25	—
	NGM-SLAM [106]	—	37.43	0.98	0.08	0.51	—
	MotionGS [121]	—	39.60	0.98	0.04	0.49	—
RGB-D	RTG-SLAM [114]	—	35.43	0.98	0.11	0.18	—
	CG-SLAM [109]	—	—	—	—	0.27	—
	LoopSplat [117]	0.51	36.63	0.99	0.11	0.26	—
	Photo-SLAM(RGB) [122]	—	33.30	0.93	0.08	1.09	—
	NGM-SLAM(RGB) [106]	—	35.02	0.96	0.13	8.51	—
	MGS-SLAM [123]	—	29.90	0.88	0.09	0.32	—
	MonoGo-SLAM [124]	—	33.59	0.93	0.22	0.32	—
	Splat-SLAM [125]	2.41	36.45	0.95	0.06	0.34	—
	IG-SLAM [126]	4.33	36.21	0.96	0.05	0.44	—
	SCS-SLAM [133]	0.36	34.66	0.97	0.10	0.41	92.72
Semantic SLAM	SemGause-SLAM [134]	0.50	35.03	0.98	0.06	0.33	96.34
	NEDS-SLAM [135]	0.47	34.76	0.96	0.09	0.35	90.78
	GS ³ SLAM [137]	—	36.26	0.99	0.05	0.37	96.63

shown in Table 4. In addition, time consumption results reported in 3DGS SLAM works are obtained from experiments conducted on different GPUs. Since the computational capabilities of GPUs vary significantly, it is unfair to directly compare the real-time performance of SLAM methods based on these results. Therefore, FPS metric is not reported.

5.2 Scene Reconstruction Evaluation

Datasets. Indoor reconstruction typically utilizes datasets similar to those used in SLAM. For outdoor scene reconstruction, commonly used datasets include KITTI [201], Waymo [202], nuScenes [203], and Argoverse [204] dataset. Detailed descriptions of outdoor reconstruction datasets are shown below.

- *KITTI dataset* [201] contains data from 4 cameras and 1 LiDAR, consisting of 22 stereo sequences with a total length of 39.2km. This dataset also comprises 200k 3D object annotations and 389 optical flow image pairs.
- *Waymo dataset* [202] contains data from 5 cameras and 5 LiDAR sensors, consisting of 1150 scenes across a range of urban and suburban geographies. This dataset is annotated with 2D (camera image) and 3D (LiDAR) bounding boxes.
- *nuScenes dataset* [203] contains data from 6 cameras, 5 radars, 1 lidar sensor, and IMU, comprising 1000 scenes, each fully annotated with 3D bounding boxes for 23 classes and 8 attributes.
- *Argoverse dataset* [204] contains sequences of LiDAR measurements, 360° images from 7 cameras, and forward-facing stereo imagery. This dataset also includes annotations for 290km of mapped lanes and 300k interesting vehicle trajectories.

Metrics. Rendering metrics *PSNR*, *SSIM*, and *LPIPS* are employed for evaluation.

Results. In terms of static reconstruction, various methods use different scenes from these datasets for evaluation rather than using specific scenes. For dynamic scene reconstruction, most methods employ consistent training and testing splits on the KITTI dataset [201] for evaluation. We present performance comparison on KITTI [201] regarding rendering metrics, as shown in Table 5. Some dynamic scene

TABLE 5: Dynamic scene reconstruction performance comparison. The results are novel view rendering performance on KITTI dataset [201] using 75% and 50% of full data for training respectively.

Methods	KITTI [75%]			KITTI [50%]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR [dB]↑	SSIM↑	LPIPS↓
PVG [26]	27.43	0.896	0.114	—	—	—
VDG [33]	25.29	0.851	0.152	—	—	—
Street Gaussians [28]	25.79	0.844	0.081	25.52	0.841	0.084
AutoSplat [35]	26.59	0.913	0.204	26.22	0.907	0.207
4DGF [32]	31.34	0.945	0.026	30.55	0.931	0.028

reconstruction methods that do not report KITTI results are not displayed in the table.

5.3 Localization Evaluation

Datasets. Evaluation of localization performance typically uses *Tanks&Temples* [205], *7 Scenes* [206] datasets, and autonomous driving datasets that are similar to those used in outdoor scene reconstruction. Detailed descriptions of localization datasets are as follows.

- *Tanks&Temples dataset* [205] contains high-resolution video sequences and ground truth poses of both indoor and outdoor scenes.
- *7 Scenes* [206] is an RGB-D dataset with ground truth poses and dense 3D models. The sequences were recorded for each scene by different users and split into distinct training and testing sets for localization evaluation.

Metrics. Pose metrics *Absolute Trajectory Error (ATE)* and *Relative Pose Error (RPE)* are used for localization evaluation. ATE quantifies the difference between the estimated camera positions and the ground truth positions. RPE measures the relative pose errors between pairs of images. In 3DGS-based localization, rendering metrics *PSNR*, *SSIM*, and *LPIPS* are also employed for evaluating the accuracy of images rendered based on the estimated poses.

Results. Existing 3DGS-based localization methods employ various datasets and sequences. Meanwhile, these methods are not open source, so it is not possible to compare the performance of existing localization methods under the same conditions. Generally, existing 3DGS-based localization methods can achieve a positional accuracy of 5cm and an angular accuracy of 2°.

5.4 Segmentation Evaluation

Datasets. Datasets with ground truth semantic annotations are used for segmentation evaluation, including LERF-Mask [58], SPIn-NeRF [207], as well as indoor datasets Replica [196] and ScanNet [197], which are also used in SLAM evaluations. Specifically, LERF-Mask [58] is used to assess text-query semantic segmentation, a crucial task in robotic manipulation and navigation that requires robots to identify specific objects in the environment following text prompts. Detailed descriptions of segmentation datasets are given below.

- *LERF-Mask dataset* [58] contains semantic annotations of three scenes from LERF-Localization dataset [208]. This dataset contains a total of 23 prompts.
- *SPIn-NeRF dataset* [207] contains 10 real-world forward-facing scenes with annotated object masks. Each scene

TABLE 6: Segmentation performance comparison.

Methods	LERF-Mask	SPIn-NeRF	
	mIoU↑	mIoU↑	Acc.↑
SAGA [59]	–	88.0	98.5
SAGD [63]	–	89.9	98.7
Gaussian Grouping [58]	72.8	–	–
Feature 3DGs [60]	65.6	–	–
Gaga [66]	74.7	–	–
CGC [67]	80.3	–	–
Click-Gaussian [72]	89.1	94.0	–

includes 60 training images with the object and 40 test images without the object.

Metrics. Although existing 3DGs-based semantic segmentation methods can achieve 3D segmentation, their performance is still evaluated by measuring the discrepancy between the rendered semantics and the 2D ground truth labels. Mean intersection over union ($mIoU(\%)$), mean boundary intersection over union ($mBIOU(\%)$), and pixel accuracy ($Acc.(\%)$) are used for semantic evaluation. Specifically, mIoU measures the overlap of the ground truth and rendered masks. mBIOU quantifies contour alignment between predicted and ground truth masks.

Results. Performance comparison on LERF-Mask [58] and SPIn-NeRF [207] datasets are shown in Table 6.

5.5 Manipulation Evaluation

Datasets. The performance of manipulation methods is evaluated on both simulated datasets and real-world scenes. Simulated datasets include RLBench [209] and Robomimic [210]. In addition, real-world robotic arm trials are conducted to measure the grasping accuracy in practical scenarios. Detailed descriptions of the simulated datasets are shown below.

- *RLBench dataset* [209] is a large-scale learning environment featuring 100 unique, hand-designed tasks. These tasks range from simple target reaching and door opening, to longer multi-stage tasks, such as opening an oven and placing a tray in it.
- *Robomimic dataset* [210] consists of data collected from three different sources, including Machine-Generated (MG), Proficient-Human (PH), and Multi-Human (MH). MG contains 300 rollout trajectories from agent checkpoints that are saved regularly during training of a reinforcement learning (RL) method [211]. PH consists of 200 demonstrations collected from a single proficient human operator. MH consists of 300 demonstrations collected from six human operators of varied proficiency.

Metrics. *Success rate (%)* metric is used to evaluate the performance of manipulation, representing the percentage of successful executions of grasping or placement tasks by a robot arm over multiple tries.

Results. In existing 3DGs-based manipulation methods, grippers are employed to perform operations. Due to variations in real-world test scenarios and differences in the robot arms used, it is not feasible to conduct comparisons under the same conditions. Generally, for single-stage tasks, the success rate can reach up to 80%, whereas for multi-stage tasks, the success rate is under 50%.

5.6 Path Planning Evaluation

Datasets. 3DGs-based path planning evaluation uses simulated dataset Matterport3D [212], Habitat-Matterport 3D (HM3D) [213] or custom-built simulation environments via the Unity engine. Detailed descriptions of path planning datasets are given below.

- *Matterport3D* [212] is a large-scale RGB-D dataset containing 10,800 panoramic views from 90 building-scale scenes, with surface reconstructions, camera poses, 2D and 3D semantic segmentation annotations. This dataset is typically constructed with the Habitat simulator [214] for evaluating path planning.
- *HM3D* [213], which is designed for the object goal navigation task, consists of 142,646 object instance annotations across 216 3D spaces and 3,100 rooms within those spaces.

Metrics. Planning accuracy metrics *success rate (%)* and *SPL* [215], planning safety metric *Wasserstein distance* ($W_2(P, \hat{P})$) [216], are used to evaluate path planning performance. Specifically, success rate is the percentage of trials in which the agent successfully invokes the STOP action within a predefined Euclidean distance from the goal object. SPL is success rate weighted by normalized inverse path length that considers both the success of reaching the goal and the efficiency (path length) of getting there. $W_2(P, \hat{P})$ quantifies the dissimilarity between the robot’s estimated risk distribution \hat{P} and the true risk distribution P , which indicates the accuracy of the robot’s risk assessment of the environment that represents the safety of robots.

Results. Existing 3DGs-based path planning methods are designed for various tasks, such as Instance ImageGoal Navigation (IIN) and safe navigation. Therefore, these 3DGs-based methods can only be compared with corresponding traditional navigation methods under the same conditions. Typically, for IIN task, existing methods reach a 72% success rate of path planning. In save navigation, existing works can achieve an average of 0.68 in $W_2(P, \hat{P})$ metric.

6 FUTURE RESEARCH

Although 3DGs has been widely used in robotics tasks, there are still many challenges that remain unsolved in such tasks. In this section, we present some valuable research directions as references for future research.

6.1 Robust Tracking

Existing 3DGs-based SLAM methods, although demonstrating high accuracy in dense mapping, typically fail to achieve accurate and robust tracking, especially in complex real-world scenarios. This limitation in current 3DGs-based SLAM systems is due to their reliance on directly using RGB information of image for pose optimization. Such reliance heavily depends on the quality and texture information of the images. However, in real-world robotic applications, image quality is prone to camera motion blur, degrading the performance of 3DGs-based SLAM. Moreover, there are some scenes with limited texture information, such as sky or walls, leading to insufficient constraints for pose estimation. The following presents corresponding directions to improve the robustness of tracking.

Camera motion blur. Camera motion blur is primarily caused by rapid movements of the robot and slow shutter speed of the camera, leading to blurry images. Although deblurring methods have been researched (Section 4.1.1) and used in SLAM [116], these methods fail to directly convert captured blurry images into sharp ones. Instead, they simulate motion blur by averaging virtual sharp images captured during the camera exposure time to synthesize blurry images. These synthesized blurry images are then used to construct loss with the observed blurry images for Gaussian optimization, ensuring that the constructed scene is deblurred. However, such methods fail to address the degradation of image quality in the observed images caused by motion blur, which adversely affects tracking performance that relies on high-quality images for pose optimization. A suitable research direction is to leverage the advantages of 3DGS representation, such as geometric information and spatial distribution, to perform tracking. This method can reduce the reliance on image quality.

Limited texture information. In real-world scenes, there are some corner cases where the environmental texture information is limited, leading to insufficient constraints for pose optimization that solely relies on image quality. Although some 3DGS-based SLAM methods [128], [129] have utilized multi-sensor fusion traditional SLAM as odometry for tracking, these methods fail when traditional SLAM is unable to handle complex corner cases. A potential research direction is to incorporate original sensor data of multiple sensors, such as IMU, wheel encoders, and LiDAR, with 3D Gaussian representation to provide sufficient constraints for pose optimization. This approach not only leverages the spatial structural information and dense scene representation offered by 3DGS, but also exploits the various constraints from multi-sensor information.

6.2 Lifelong Mapping and Localization

Current 3DGS methods primarily focus on short-term reconstruction and localization. However, in most real-world scenarios, the environment undergoes constant changes over time. A prebuilt map that fails to consider these changes may quickly become outdated and unreliable. Consequently, it is crucial to maintain an up-to-date model of the environment to facilitate the long-term operation or navigation of robots. Although some traditional methods [217], [218] have achieved long-term mapping, these approaches focus on constructing and updating sparse maps, which are insufficient for downstream robotic tasks. Therefore, a promising research direction is lifelong 3DGS-based dense mapping and localization. Since 3DGS is an explicit and dense representation, the dynamic update and refinement of the Gaussian map can be achieved through explicit editing of Gaussian primitives. Additionally, we believe that the inconsistencies in the Gaussian map caused by long-term dynamic changes can be optimized by leveraging the inner constraints between Gaussian primitives. Therefore, by harnessing the explicit representation and inherent constraints among Gaussian primitives, lifelong mapping and localization can be achieved.

6.3 Large-scale Relocalization

In robotic applications, it is necessary for robots to relocate their current poses upon entering a pre-established map. However, existing 3DGS-based relocalization methods [144], [145] either require a coarse initial pose or are only capable of achieving relocalization in small indoor scenes. These methods struggle to perform relocalization in large-scale outdoor scenes without an initial pose. Unfortunately, it is challenging to obtain a coarse initial pose for relocalization in practical robotic applications. Therefore, a meaningful research direction is large-scale relocalization without prior poses. We believe that constructing a submap index library or descriptor based on 3DGS representation facilitates coarse pose regression. In addition, the coarse pose can be refined through a registration process that leverages geometric and appearance features within the 3DGS representation.

6.4 Sim-to-Real Manipulation

Collecting real-world manipulation datasets is challenging, leading to a scarcity of data for training effective grasping in real scenarios. Therefore, grasping methods often require initial training in simulation environments before being transferred to real-world settings. Although 3DGS-based sim-to-real method [219] has been explored, it has limitation in generalization. Specifically, this approach heavily depends on scene-specific training, which hinders its ability to generalize and transfer learned knowledge between similar task scenarios. Consequently, this method still requires a substantial amount of real-world datasets for training. Furthermore, the discrepancies in material and physical properties between simulation and reality environments can lead to significant differences in training data distributions for manipulation tasks. These discrepancies may potentially result in entirely different operation strategies. However, existing method [220] only enables modeling the physical properties of real-world scenarios. Therefore, a promising research direction involves directly incorporating uncertainty and environmental features into the 3DGS representation to enhance generalization and property modeling.

7 CONCLUSION

As a powerful radiance field for dense scene representation, 3DGS provides new options in the field of robotics for scene understanding and interaction. Specifically, 3DGS offers a dependable selection for many applications in robotics, such as reconstruction, scene segmentation, scene editing, SLAM, manipulation, and navigation. Moreover, the capability of 3DGS to enhance its performance in large-scale environments, motion-blurred conditions, few-shot scenarios, etc., remains mostly unexploited. Exploring these areas could significantly deepen the integration between 3DGS and robotics. In addition, we conduct a thorough performance evaluation of current 3DGS applications in robotics, helping readers choose their preferred approach. Finally, we discuss in detail the challenges and future development directions of 3DGS in robotics. Therefore, as our survey provides a comprehensive summary of the field’s outstanding work

and emphasizes its potential, we hope this survey encourages more researchers to explore new possibilities and successfully implement them on real robotic platforms.

REFERENCES

- [1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.
- [2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.
- [3] L. A. Westover, *Splatting: a parallel, feed-forward volume rendering algorithm*. The University of North Carolina at Chapel Hill, 1991.
- [4] G. Chen and W. Wang, "A survey on 3d gaussian splatting," *arXiv preprint arXiv:2401.03890*, 2024.
- [5] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3d gaussian as a new vision era: A survey," *arXiv preprint arXiv:2402.07181*, 2024.
- [6] A. Dalal, D. Hagen, K. G. Robbersmyr, and K. M. Knausgård, "Gaussian splatting: 3d reconstruction and novel view synthesis, a review," *IEEE Access*, vol. 12, pp. 96797–96820, 2024.
- [7] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao, "Recent advances in 3d gaussian splatting," *Comp. Visual Media*, pp. 1–30, 2024.
- [8] M. T. Bagdasarian, P. Knoll, F. Barthel, A. Hilsmann, P. Eisert, and W. Morgenstern, "3dgs. zip: A survey on 3d gaussian splatting compression methods," *arXiv preprint arXiv:2407.09510*, 2024.
- [9] Y. Bao, T. Ding, J. Huo, Y. Liu, Y. Li, W. Li, Y. Gao, and J. Luo, "3d gaussian splatting: Survey, technologies, challenges, and opportunities," *arXiv preprint arXiv:2407.17418*, 2024.
- [10] B. Cabral, N. Max, and R. Springmeyer, "Bidirectional reflection functions from surface bump maps," in *Proc. of the 14th annual conf. on Computer graphics and interactive techniques*, 1987, pp. 273–281.
- [11] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016, pp. 4104–4113.
- [12] J. Bai, L. Huang, J. Guo, W. Gong, Y. Li, and Y. Guo, "360-gs: Layout-guided panoramic gaussian splatting for indoor roaming," *arXiv preprint arXiv:2402.00763*, 2024.
- [13] H. Xiang, X. Li, X. Lai, W. Zhang, Z. Liao, K. Cheng, and X. Liu, "Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction," *arXiv preprint arXiv:2405.19671*, 2024.
- [14] J. Kim and J. Lim, "Integrating meshes and 3d gaussians for indoor scene reconstruction with sam mask guidance," *arXiv preprint arXiv:2407.16173*, 2024.
- [15] K. Cheng, X. Long, K. Yang, Y. Yao, W. Yin, Y. Ma, W. Wang, and X. Chen, "Gaussianpro: 3d gaussian splatting with progressive propagation," in *Forty-first Int. conf. on Machine Learning*, 2024.
- [16] H. Dahmani, M. Bennehar, N. Piasco, L. Roldao, and D. Tsishkou, "Swag: Splatting in the wild images with appearance-conditioned gaussians," *arXiv preprint arXiv:2403.10427*, 2024.
- [17] C. Ye, Y. Nie, J. Chang, Y. Chen, Y. Zhi, and X. Han, "Gaustudio: A modular framework for 3d gaussian splatting and beyond," *arXiv preprint arXiv:2403.19632*, 2024.
- [18] K. Wu, K. Zhang, Z. Zhang, S. Yuan, M. Tie, J. Wei, Z. Xu, J. Zhao, Z. Gan, and W. Ding, "Hgs-mapping: Online dense mapping using hybrid gaussian representation in urban scenes," *RA-L*, vol. 9, no. 11, pp. 9573–9580, 2024.
- [19] Z. Li, Y. Zhang, C. Wu, J. Zhu, and L. Zhang, "Ho-gaussian: Hybrid optimization of 3d gaussian splatting for urban scenes," *arXiv preprint arXiv:2403.20032*, 2024.
- [20] C. Zhao, S. Sun, R. Wang, Y. Guo, J.-J. Wan, Z. Huang, X. Huang, Y. V. Chen, and L. Ren, "Tclc-gs: Tightly coupled lidar-camera gaussian splatting for surrounding autonomous driving scenes," *arXiv preprint arXiv:2404.02410*, 2024.
- [21] D. Chen, H. Li, W. Ye, Y. Wang, W. Xie, S. Zhai, N. Wang, H. Liu, H. Bao, and G. Zhang, "Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction," *arXiv preprint arXiv:2406.06521*, 2024.
- [22] J. Xu, Y. Mei, and V. M. Patel, "Wild-gs: Real-time novel view synthesis from unconstrained photo collections," *arXiv preprint arXiv:2406.10373*, 2024.
- [23] J. Kulhanek, S. Peng, Z. Kukelova, M. Pollefeys, and T. Sattler, "Wildgaussians: 3d gaussian splatting in the wild," *NeurIPS*, 2024.
- [24] X. Shi, L. Chen, P. Wei, X. Wu, T. Jiang, Y. Luo, and L. Xie, "Dhgs: Decoupled hybrid gaussian splatting for driving scene," *arXiv preprint arXiv:2407.16600*, 2024.
- [25] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2d gaussian splatting for geometrically accurate radiance fields," in *ACM SIGGRAPH 2024 Conf.*, 2024, pp. 1–11.
- [26] Y. Chen, C. Gu, J. Jiang, X. Zhu, and L. Zhang, "Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering," *arXiv preprint arXiv:2311.18561*, 2023.
- [27] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," in *CVPR*, 2024, pp. 21634–21643.
- [28] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians for modeling dynamic urban scenes," *arXiv preprint arXiv:2401.01339*, 2024.
- [29] H. Zhou, J. Shao, L. Xu, D. Bai, W. Qiu, B. Liu, Y. Wang, A. Geiger, and Y. Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *CVPR*, 2024, pp. 21336–21345.
- [30] J. Lei, Y. Weng, A. Harley, L. Guibas, and K. Daniilidis, "Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds," *arXiv preprint arXiv:2405.17421*, 2024.
- [31] N. Huang, X. Wei, W. Zheng, P. An, M. Lu, W. Zhan, M. Tomizuka, K. Keutzer, and S. Zhang, "S³ gaussian: Self-supervised street gaussians for autonomous driving," *arXiv preprint arXiv:2405.20323*, 2024.
- [32] T. Fischer, J. Kulhanek, S. R. Bulò, L. Porzi, M. Pollefeys, and P. Kontschieder, "Dynamic 3d gaussian fields for urban areas," *arXiv preprint arXiv:2406.03175*, 2024.
- [33] H. Li, J. Li, D. Zhang, C. Wu, J. Shi, C. Zhao, H. Feng, E. Ding, J. Wang, and J. Han, "Vdg: Vision-only dynamic gaussian for driving simulation," *arXiv preprint arXiv:2406.18198*, 2024.
- [34] D. Zhang, G. Li, J. Li, M. Bressieux, O. Hilliges, M. Pollefeys, L. Van Gool, and X. Wang, "Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting," *arXiv preprint arXiv:2406.19811*, 2024.
- [35] M. Khan, H. Fazlali, D. Sharma, T. Cao, D. Bai, Y. Ren, and B. Liu, "Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction," *arXiv preprint arXiv:2407.02598*, 2024.
- [36] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, "Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis," *arXiv preprint arXiv:2308.09713*, 2023.
- [37] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction," in *CVPR*, 2024, pp. 20331–20341.
- [38] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *CVPR*, 2024, pp. 20310–20320.
- [39] Z. Yang, H. Yang, Z. Pan, X. Zhu, and L. Zhang, "Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting," *arXiv preprint arXiv:2310.10642*, 2023.
- [40] K. Katsumata, D. M. Vo, and H. Nakayama, "An efficient 3d gaussian representation for monocular/multi-view dynamic scenes," *arXiv preprint arXiv:2311.12897*, 2023.
- [41] Y.-H. Huang, Y.-T. Sun, Z. Yang, X. Lyu, Y.-P. Cao, and X. Qi, "Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes," in *CVPR*, 2024, pp. 4220–4230.
- [42] Y. Lin, Z. Dai, S. Zhu, and Y. Yao, "Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle," in *CVPR*, 2024, pp. 21136–21145.
- [43] Y. Liang, N. Khan, Z. Li, T. Nguyen-Phuoc, D. Lanman, J. Tompkin, and L. Xiao, "Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis," *arXiv preprint arXiv:2312.11458*, 2023.
- [44] R. Shaw, J. Song, A. Moreau, M. Nazarczuk, S. Catley-Chandar, H. Dhamo, and E. Perez-Pellitero, "Swags: Sampling windows adaptively for dynamic 3d gaussian splatting," *arXiv preprint arXiv:2312.13308*, 2023.
- [45] Z. Li, Z. Chen, Z. Li, and Y. Xu, "Spacetime gaussian feature splatting for real-time dynamic view synthesis," in *CVPR*, 2024, pp. 8508–8520.
- [46] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen, "4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes," *arXiv preprint arXiv:2402.03307*, 2024.
- [47] J. Sun, H. Jiao, G. Li, Z. Zhang, L. Zhao, and W. Xing, "3dgstream: On-the-fly training of 3d gaussians for efficient streaming of

- photo-realistic free-viewpoint videos," in *CVPR*, 2024, pp. 20 675–20 685.
- [48] J. Bae, S. Kim, Y. Yun, H. Lee, G. Bang, and Y. Uh, "Per-gaussian embedding-based deformation for deformable 3d gaussian splatting," *arXiv preprint arXiv:2404.03613*, 2024.
- [49] Z. Lu, X. Guo, L. Hui, T. Chen, M. Yang, X. Tang, F. Zhu, and Y. Dai, "3d geometry-aware deformable gaussian splatting for dynamic view synthesis," in *CVPR*, 2024, pp. 8900–8910.
- [50] B. Zhang, B. Zeng, and Z. Peng, "A refined 3d gaussian representation for high-quality dynamic scene reconstruction," *arXiv preprint arXiv:2405.17891*, 2024.
- [51] D. Wan, R. Lu, and G. Zeng, "Superpoint gaussian splatting for real-time high-fidelity dynamic scene reconstruction," *arXiv preprint arXiv:2406.03697*, 2024.
- [52] M.-L. Shih, J.-B. Huang, C. Kim, R. Shah, J. Kopf, and C. Gao, "Modeling ambient scene dynamics for free-view synthesis," *arXiv preprint arXiv:2406.09395*, 2024.
- [53] C. Stearns, A. Harley, M. Uy, F. Dubost, F. Tombari, G. Wetzstein, and L. Guibas, "Dynamic gaussian marbles for novel view synthesis of casual monocular videos," *arXiv preprint arXiv:2406.18717*, 2024.
- [54] L. Xie, J. Julin, K. Niinuma, and L. A. Jeni, "Gaussian splatting lk," *arXiv preprint arXiv:2407.11309*, 2024.
- [55] B. He, Y. Chen, G. Lu, L. Song, and W. Zhang, "S4d: Streaming 4d real-world reconstruction with gaussians and 3d control points," *arXiv preprint arXiv:2408.13036*, 2024.
- [56] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *CVPR*, 2023, pp. 130–141.
- [57] M. T. Bosch, "N-dimensional rigid body dynamics," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 55–1, 2020.
- [58] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," *arXiv preprint arXiv:2312.00732*, 2023.
- [59] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Segment any 3d gaussians," *arXiv preprint arXiv:2312.00860*, 2023.
- [60] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in *CVPR*, 2024, pp. 21 676–21 685.
- [61] K. Lan, H. Li, H. Shi, W. Wu, Y. Liao, L. Wang, and P. Zhou, "2d-guided 3d gaussian segmentation," *arXiv preprint arXiv:2312.16047*, 2023.
- [62] B. Dou, T. Zhang, Y. Ma, Z. Wang, and Z. Yuan, "Cosseggaussians: Compact and swift scene segmenting 3d gaussians," *arXiv preprint arXiv:2401.05925*, 2024.
- [63] X. Hu, Y. Wang, L. Fan, J. Fan, J. Peng, Z. Lei, Q. Li, and Z. Zhang, "Semantic anything in 3d gaussians," *arXiv preprint arXiv:2401.17857*, 2024.
- [64] J. Guo, X. Ma, Y. Fan, H. Liu, and Q. Li, "Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting," *arXiv preprint arXiv:2403.15624*, 2024.
- [65] Q. Gu, Z. Lv, D. Frost, S. Green, J. Straub, and C. Sweeney, "Ego-lifter: Open-world 3d segmentation for egocentric perception," *arXiv preprint arXiv:2403.18118*, 2024.
- [66] W. Lyu, X. Li, A. Kundu, Y.-H. Tsai, and M.-H. Yang, "Gaga: Group any gaussians via 3d-aware memory bank," *arXiv preprint arXiv:2404.07977*, 2024.
- [67] M. C. Silva, M. Dahaghin, M. Toso, and A. Del Bue, "Contrastive gaussian clustering: Weakly supervised 3d scene segmentation," *arXiv preprint arXiv:2404.12784*, 2024.
- [68] G. Liao, J. Li, Z. Bao, X. Ye, J. Wang, Q. Li, and K. Liu, "Clip-gs: Clip-informed gaussian splatting for real-time and view-consistent 3d semantic understanding," *arXiv preprint arXiv:2404.14249*, 2024.
- [69] X. Zuo, P. Samangouei, Y. Zhou, Y. Di, and M. Li, "Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding," *arXiv preprint arXiv:2401.01970*, 2024.
- [70] M.-B. Jurca, R. Royen, I. Giosan, and A. Munteanu, "Rt-gs2: Real-time generalizable semantic segmentation for 3d gaussian representations of radiance fields," *arXiv preprint arXiv:2405.18033*, 2024.
- [71] Y. Wu, J. Meng, H. Li, C. Wu, Y. Shi, X. Cheng, C. Zhao, H. Feng, E. Ding, J. Wang *et al.*, "Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding," *arXiv preprint arXiv:2406.02058*, 2024.
- [72] S. Choi, H. Song, J. Kim, T. Kim, and H. Do, "Click-gaussian: Interactive segmentation to any 3d gaussians," *arXiv preprint arXiv:2407.11793*, 2024.
- [73] S. Ji, G. Wu, J. Fang, J. Cen, T. Yi, W. Liu, Q. Tian, and X. Wang, "Segment any 4d gaussians," *arXiv preprint arXiv:2407.04504*, 2024.
- [74] F. Chabot, N. Granger, and G. Lapouge, "Gaussianbev: 3d gaussian representation meets perception models for bev segmentation," *arXiv preprint arXiv:2407.14108*, 2024.
- [75] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.
- [76] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *ICCV*, 2023, pp. 1316–1326.
- [77] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Int. conf. on machine learning*. PMLR, 2021, pp. 8748–8763.
- [78] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021, pp. 9650–9660.
- [79] Y. Wang, Q. Wu, G. Zhang, and D. Xu, "Gscream: Learning 3d geometry and feature consistent gaussian splatting for object removal," *arXiv preprint arXiv:2404.13679*, 2024.
- [80] Y. Mei, J. Xu, and V. M. Patel, "Reference-based controllable scene stylization with gaussian splatting," *arXiv preprint arXiv:2407.07220*, 2024.
- [81] J. Huang and H. Yu, "Point'n move: Interactive scene object manipulation on gaussian splatting radiance fields," *arXiv preprint arXiv:2311.16737*, 2023.
- [82] G. Luo, T.-X. Xu, Y.-T. Liu, X.-X. Fan, F.-L. Zhang, and S.-H. Zhang, "3d gaussian editing with a single image," *ACM Multimedia*, 2024.
- [83] R.-Z. Qiu, G. Yang, W. Zeng, and X. Wang, "Feature splatting: Language-driven physics-based scene synthesis and editing," *arXiv preprint arXiv:2404.01223*, 2024.
- [84] T.-X. Xu, W. Hu, Y.-K. Lai, Y. Shan, and S.-H. Zhang, "Texture-gs: Disentangling the geometry and texture for 3d gaussian splatting editing," *arXiv preprint arXiv:2403.10050*, 2024.
- [85] J. Zhuang, D. Kang, Y.-P. Cao, G. Li, L. Lin, and Y. Shan, "Tipeditor: An accurate 3d editor following both text-prompts and image-prompts," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–12, 2024.
- [86] F. Palandra, A. Sanchietti, D. Baieri, and E. Rodolà, "Gsedit: Efficient text-guided editing of 3d objects via gaussian splatting," *arXiv preprint arXiv:2403.05154*, 2024.
- [87] J. Wu, J.-W. Bian, X. Li, G. Wang, I. Reid, P. Torr, and V. A. Prisacariu, "Gaussctrl: multi-view consistent text-driven 3d gaussian splatting editing," *arXiv preprint arXiv:2403.08733*, 2024.
- [88] Y. Wang, X. Yi, Z. Wu, N. Zhao, L. Chen, and H. Zhang, "View-consistent 3d editing with gaussian splatting," *arXiv preprint arXiv:2403.11868*, 2024.
- [89] J. Wang, J. Fang, X. Zhang, L. Xie, and Q. Tian, "Gaussianeditor: Editing 3d gaussians delicately with text instructions," in *CVPR*, 2024, pp. 20 902–20 911.
- [90] Y. Chen, Z. Chen, C. Zhang, F. Wang, X. Yang, Y. Wang, Z. Cai, L. Yang, H. Liu, and G. Lin, "Gaussianeditor: Swift and controllable 3d editing with gaussian splatting," in *CVPR*, 2024, pp. 21 476–21 485.
- [91] M. Chen, I. Laina, and A. Vedaldi, "Dge: Direct gaussian 3d editing by consistent multi-view editing," 2024.
- [92] T. Xu, J. Chen, P. Chen, Y. Zhang, J. Yu, and W. Yang, "Tiger: Text-instructed 3d gaussian retrieval and coherent editing," *arXiv preprint arXiv:2405.14455*, 2024.
- [93] C. Luo, D. Di, Y. Ma, Z. Xue, C. Wei, X. Yang, and Y. Liu, "Trame: Trajectory-anchored multi-view editing for text-guided 3d gaussian splatting manipulation," *arXiv preprint arXiv:2407.02034*, 2024.
- [94] U. Khalid, H. Iqbal, A. Farooq, J. Hua, and C. Chen, "3dgeo: 3d editing on the go!" in *ECCV*, 2024.
- [95] H. Xiao, Y. Chen, H. Huang, H. Xiong, J. Yang, P. Prasad, and Y. Zhao, "Localized gaussian splatting editing with contextual awareness," *arXiv preprint arXiv:2408.00083*, 2024.
- [96] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

- [97] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.
- [98] K. Liu, F. Zhan, M. Xu, C. Theobalt, L. Shao, and S. Lu, "Style-gaussian: Instant 3d style transfer with gaussian splatting," *arXiv preprint arXiv:2403.07807*, 2024.
- [99] D. Zhang, Z. Chen, Y.-J. Yuan, F.-L. Zhang, Z. He, S. Shan, and L. Gao, "Stylizedgs: Controllable stylization for 3d gaussian splatting," *arXiv preprint arXiv:2404.05220*, 2024.
- [100] A. Saroha, M. Gladkova, C. Curreli, T. Yenamandra, and D. Cremers, "Gaussian splatting in style," *arXiv preprint arXiv:2403.08498*, 2024.
- [101] V. Jaganathan, H. H. Huang, M. Z. Irshad, V. Jampani, A. Raj, and Z. Kira, "Ice-g: Image conditional editing of 3d gaussian splats," *arXiv preprint arXiv:2406.08488*, 2024.
- [102] S. Jain, A. Kuthiala, P. S. Sethi, and P. Saxena, "Stylesplat: 3d object style transfer with gaussian splatting," *arXiv preprint arXiv:2407.09473*, 2024.
- [103] X.-Y. Yu, J.-X. Yu, L.-B. Zhou, Y. Wei, and L.-L. Ou, "Instantstyle-gaussian: Efficient art style transfer with 3d gaussian splatting," *SIGGRAPH Asia*, 2024.
- [104] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, "Splatam: Splat track & map 3d gaussians for dense rgb-d slam," in *CVPR*, 2024, pp. 21357–21366.
- [105] V. Yugay, Y. Li, T. Gevers, and M. R. Oswald, "Gaussian-slam: Photo-realistic dense slam with gaussian splatting," *arXiv preprint arXiv:2312.10070*, 2023.
- [106] M. Li, J. Huang, L. Sun, A. X. Tian, T. Deng, and H. Wang, "Ngm-slam: Gaussian splatting slam with radiance field submap," *arXiv preprint arXiv:2405.05702*, 2024.
- [107] S. Liu, H. Zhou, L. Li, Y. Liu, T. Deng, Y. Zhou, and M. Li, "Structure gaussian slam with manhattan world hypothesis," *arXiv preprint arXiv:2405.20031*, 2024.
- [108] S. Sun, M. Mielle, A. J. Lilienthal, and M. Magnusson, "High-fidelity slam using gaussian splatting with rendering-guided densification and regularized optimization," in *IROS*, 2024.
- [109] J. Hu, X. Chen, B. Feng, G. Li, L. Yang, H. Bao, G. Zhang, and Z. Cui, "Cg-slam: Efficient dense rgb-d slam in a consistent uncertainty-aware 3d gaussian field," in *ECCV*, 2024.
- [110] C. Yan, D. Qu, D. Xu, B. Zhao, Z. Wang, D. Wang, and X. Li, "Gs-slam: Dense visual slam with 3d gaussian splatting," in *CVPR*, 2024, pp. 19 595–19 604.
- [111] S. Ha, J. Yeon, and H. Yu, "Rgbd gs-icp slam," in *ECCV*, 2024.
- [112] P. Jiang, H. Liu, X. Li, T. Wang, F. Zhang, and J. M. Buhmann, "Tamberge: Bridging frame-centered tracking and 3d gaussian splatting for enhanced slam," *arXiv preprint arXiv:2405.19614*, 2024.
- [113] T. Deng, Y. Chen, L. Zhang, J. Yang, S. Yuan, D. Wang, and W. Chen, "Compact 3d gaussian splatting for dense visual slam," *arXiv preprint arXiv:2403.11247*, 2024.
- [114] Z. Peng, T. Shao, Y. Liu, J. Zhou, Y. Yang, J. Wang, and K. Zhou, "Rtg-slam: Real-time 3d reconstruction at scale using gaussian splatting," in *SIGGRAPH*, 2024.
- [115] Z. Qu, Z. Zhang, C. Liu, and J. Yin, "Visual slam with 3d gaussian primitives and depth priors enabling novel view synthesis," *arXiv preprint arXiv:2408.05635*, 2024.
- [116] G. Bae, C. Choi, H. Heo, S. M. Kim, and Y. M. Kim, "I²-slam: Inverting imaging process for robust photorealistic dense slam," in *ECCV*, 2024.
- [117] L. Zhu, Y. Li, E. Sandström, K. Schindler, and I. Armeni, "Loopsplat: Loop closure by registering 3d gaussian splats," *arXiv preprint arXiv:2408.10154*, 2024.
- [118] J. Wei and S. Leutenegger, "Gsfusion: Online rgb-d mapping where gaussian splatting meets tsdf fusion," *arXiv preprint arXiv:2408.12677*, 2024.
- [119] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp." in *Robotics: science and systems*, vol. 2, no. 4. Seattle, WA, 2009, p. 435.
- [120] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," in *CVPR*, 2024, pp. 18 039–18 048.
- [121] X. Guo, P. Han, W. Zhang, and H. Chen, "Motions: Compact gaussian splatting slam by motion filter," *arXiv preprint arXiv:2405.11129*, 2024.
- [122] H. Huang, L. Li, H. Cheng, and S.-K. Yeung, "Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras," in *CVPR*, 2024, pp. 21 584–21 593.
- [123] P. Zhu, Y. Zhuang, B. Chen, L. Li, C. Wu, and Z. Liu, "Mgs-slam: Monocular sparse tracking and gaussian mapping with depth smooth regularization," *arXiv preprint arXiv:2405.06241*, 2024.
- [124] T. Lan, Q. Lin, and H. Wang, "Monocular gaussian slam with language extended loop closure," *arXiv preprint arXiv:2405.13748*, 2024.
- [125] E. Sandström, K. Tateno, M. Oechsle, M. Niemeyer, L. Van Gool, M. R. Oswald, and F. Tombari, "Splat-slam: Globally optimized rgb-only slam with 3d gaussians," *arXiv preprint arXiv:2405.16544*, 2024.
- [126] F. A. Sarikamis and A. A. Alatan, "Ig-slam: Instant gaussian slam," *arXiv preprint arXiv:2408.01126*, 2024.
- [127] Z. Teed, L. Lipson, and J. Deng, "Deep patch visual odometry," *NeurIPS*, vol. 36, 2024.
- [128] S. Hong, J. He, X. Zheng, C. Zheng, and S. Shen, "Liv-gaussmap: Lidar-inertial-visual fusion for real-time 3d radiance field map rendering," *RA-L*, 2024.
- [129] X. Lang, L. Li, H. Zhang, F. Xiong, M. Xu, Y. Liu, X. Zuo, and J. Lv, "Gaussian-lic: Photo-realistic lidar-inertial-camera slam with 3d gaussian splatting," *arXiv preprint arXiv:2404.06926*, 2024.
- [130] C. Wu, Y. Duan, X. Zhang, Y. Sheng, J. Ji, and Y. Zhang, "Mmgaussian: 3d gaussian-based multi-modal fusion for localization and reconstruction in unbounded scenes," in *IROS*, 2024.
- [131] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss, "Kiss-icp: In defense of point-to-point icp—simple, accurate, and robust registration if done the right way," *RA-L*, vol. 8, no. 2, pp. 1029–1036, 2023.
- [132] L. C. Sun, N. P. Bhatt, J. C. Liu, Z. Fan, Z. Wang, T. E. Humphreys, and U. Topcu, "Mm3dgs slam: Multi-modal 3d gaussian splatting for slam using vision, depth, and inertial measurements," in *IROS*, 2024.
- [133] M. Li, S. Liu, and H. Zhou, "Sgs-slam: Semantic gaussian splatting for neural dense slam," in *ECCV*, 2024.
- [134] S. Zhu, R. Qin, G. Wang, J. Liu, and H. Wang, "Semgauss-slam: Dense semantic gaussian splatting slam," *arXiv preprint arXiv:2403.07494*, 2024.
- [135] Y. Ji, Y. Liu, G. Xie, B. Ma, and Z. Xie, "Neds-slam: A novel neural explicit dense semantic slam framework using 3d gaussian splatting," *RA-L*, vol. 9, no. 10, pp. 8778–8785, 2024.
- [136] S. Zhu, G. Wang, H. Blum, J. Liu, L. Song, M. Pollefeys, and H. Wang, "Sni-slam: Semantic neural implicit slam," in *CVPR*, 2024, pp. 21 167–21 177.
- [137] L. Li, L. Zhang, Z. Wang, and Y. Shen, "Gs³lam: Gaussian semantic splatting slam," in *ACM Multimedia* 2024.
- [138] Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu et al., "Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping," *arXiv preprint arXiv:2403.09637*, 2024.
- [139] O. Shorinwa, J. Tucker, A. Smith, A. Swann, T. Chen, R. Firooz, M. Kennedy III, and M. Schwager, "Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting," *arXiv preprint arXiv:2405.04378*, 2024.
- [140] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation," *arXiv preprint arXiv:2403.08321*, 2024.
- [141] Y. Li and D. Pathak, "Object-aware gaussian splatting for robotic manipulation," in *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*.
- [142] Y. Sun, X. Wang, Y. Zhang, J. Zhang, C. Jiang, Y. Guo, and F. Wang, "Icommma: Inverting 3d gaussians splatting for camera pose estimation via comparing and matching," *arXiv preprint arXiv:2312.09031*, 2023.
- [143] K. Jiang, Y. Fu, M. Varma T, Y. Belhe, X. Wang, H. Su, and R. Ramamoorthi, "A construct-optimize approach to sparse view synthesis without camera pose," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.
- [144] P. Jiang, G. Pandey, and S. Saripalli, "3dgs-reloc: 3d gaussian splatting for map representation and visual relocalization," *arXiv preprint arXiv:2403.11367*, 2024.
- [145] J.-M. Liu, H.-K. Yang, T.-C. Chiang, T.-R. Liu, C.-W. Huang, Q. Kong, N. Kobori, and C.-Y. Lee, "Enhancing visual relocalization with dense scene coordinates derived from 3d gaussian splatting," in *RoboNerF: 1st Workshop On Neural Fields In Robotics at ICRA 2024*.
- [146] M. Bortolon, T. Tssemelis, S. James, F. Poiesi, and A. Del Bue, "6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model," *arXiv preprint arXiv:2407.15484*, 2024.

- [147] Y. Fu, S. Liu, A. Kulkarni, J. Kautz, A. A. Efros, and X. Wang, "Colmap-free 3d gaussian splatting," *arXiv preprint arXiv:2312.07504*, 2023.
- [148] H. Li, Y. Gao, D. Zhang, C. Wu, Y. Dai, C. Zhao, H. Feng, E. Ding, J. Wang, and J. Han, "Ggrt: Towards generalizable 3d gaussians without pose priors in real-time," *arXiv preprint arXiv:2403.10147*, 2024.
- [149] J. Chang, Y. Xu, Y. Li, Y. Chen, and X. Han, "Gaussreg: Fast 3d registration with gaussian splatting," *arXiv preprint arXiv:2407.05254*, 2024.
- [150] C. Liu, S. Chen, Y. Bhalgat, S. Hu, Z. Wang, M. Cheng, V. A. Prisacariu, and T. Braud, "Gsloc: Efficient camera pose refinement via 3d gaussian splatting," *arXiv preprint arXiv:2408.11085*, 2024.
- [151] R. Jin, Y. Gao, H. Lu, and F. Gao, "Gs-planner: A gaussian-splatting-based planning framework for active high-fidelity reconstruction," *arXiv preprint arXiv:2405.10142*, 2024.
- [152] X. Lei, M. Wang, W. Zhou, and H. Li, "Gaussnav: Gaussian splatting for visual navigation," *arXiv preprint arXiv:2403.11625*, 2024.
- [153] T. Chen, O. Shorinwa, W. Zeng, J. Bruno, P. Dames, and M. Schwager, "Splat-nav: Safe real-time robot navigation in gaussian splatting maps," *arXiv preprint arXiv:2403.02751*, 2024.
- [154] G. Liu, W. Jiang, B. Lei, V. Pandey, K. Daniilidis, and N. Motee, "Beyond uncertainty: Risk-aware active view acquisition for safe robot navigation and 3d scene understanding with fisherrf," *arXiv preprint arXiv:2403.11396*, 2024.
- [155] J. Oh, J. Chung, D. Lee, and K. M. Lee, "Deblurgs: Gaussian splatting for camera motion blur," *arXiv preprint arXiv:2404.11358*, 2024.
- [156] L. Zhao, P. Wang, and P. Liu, "Bad-gaussians: Bundle adjusted deblur gaussian splatting," *arXiv preprint arXiv:2403.11831*, 2024.
- [157] W. Chen and L. Liu, "Deblur-gs: 3d gaussian splatting from camera motion blurred images," *Proc. of the ACM on Computer Graphics and Interactive Techniques*, vol. 7, no. 1, pp. 1–15, 2024.
- [158] F. Darmon, L. Porzi, S. Rota-Bulò, and P. Kortscheder, "Robust gaussian splatting," *arXiv preprint arXiv:2404.04211*, 2024.
- [159] O. Seiskari, J. Ylälimmi, V. Kaatrasalo, P. Rantalaikila, M. Turkulainen, J. Kannala, E. Rahtu, and A. Solin, "Gaussian splatting on the move: Blur and rolling shutter compensation for natural camera motion," *arXiv preprint arXiv:2403.13327*, 2024.
- [160] J. Lee, D. Kim, D. Lee, S. Cho, and S. Lee, "Crim-gs: Continuous rigid motion-aware gaussian splatting from motion blur images," *arXiv preprint arXiv:2407.03923*, 2024.
- [161] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *NeurIPS*, vol. 31, 2018.
- [162] C. Peng, Y. Tang, Y. Zhou, N. Wang, X. Liu, D. Li, and R. Chellappa, "Bags: Blur agnostic gaussian splatting through multi-scale kernel modeling," *arXiv preprint arXiv:2403.04926*, 2024.
- [163] Y. Chen and G. H. Lee, "Dogaussian: Distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus," *arXiv preprint arXiv:2405.13943*, 2024.
- [164] K. Ren, L. Jiang, T. Lu, M. Yu, L. Xu, Z. Ni, and B. Dai, "Octreegs: Towards consistent real-time rendering with lod-structured 3d gaussians," *arXiv preprint arXiv:2403.17898*, 2024.
- [165] Y. Liu, H. Guan, C. Luo, L. Fan, J. Peng, and Z. Zhang, "Citygaussian: Real-time high-quality large-scale scene rendering with gaussians," *arXiv preprint arXiv:2404.01133*, 2024.
- [166] B. Kerbl, A. Meuleman, G. Kopanas, M. Wimmer, A. Lanvin, and G. Drettakis, "A hierarchical 3d gaussian representation for real-time rendering of very large datasets," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–15, 2024.
- [167] T. Suzuki, "Fed3dgs: Scalable 3d gaussian splatting with federated learning," *arXiv preprint arXiv:2403.11460*, 2024.
- [168] Z. Wang and D. Xu, "Pygs: Large-scale scene representation with pyramidal 3d gaussian splatting," *arXiv preprint arXiv:2405.16829*, 2024.
- [169] J. Lin, Z. Li, X. Tang, J. Liu, S. Liu, J. Liu, Y. Lu, X. Wu, S. Xu, Y. Yan *et al.*, "Vastgaussian: Vast 3d gaussians for large scene reconstruction," in *CVPR*, 2024, pp. 5166–5175.
- [170] J. C. Lee, D. Rho, X. Sun, J. H. Ko, and E. Park, "Compact 3d gaussian representation for radiance field," in *CVPR*, 2024, pp. 21719–21728.
- [171] K. Navaneet, K. P. Meibodi, S. A. Koohpayegani, and H. Pirsiavash, "Compact3d: Smaller and faster gaussian splatting with vector quantization," *arXiv preprint arXiv:2311.18159*, 2023.
- [172] T. Lu, M. Yu, L. Xu, Y. Xiangli, L. Wang, D. Lin, and B. Dai, "Scaffold-gs: Structured 3d gaussians for view-adaptive rendering," in *CVPR*, 2024, pp. 20654–20664.
- [173] X. Sun, J. C. Lee, D. Rho, J. H. Ko, U. Ali, and E. Park, "F-3dgs: Factorized coordinates and representations for 3d gaussian splatting," *arXiv preprint arXiv:2405.17083*, 2024.
- [174] H. Wang, H. Zhu, T. He, R. Feng, J. Deng, J. Bian, and Z. Chen, "End-to-end rate-distortion optimized 3d gaussian representation," *arXiv preprint arXiv:2406.01597*, 2024.
- [175] Z. Fan, K. Wang, K. Wen, Z. Zhu, D. Xu, and Z. Wang, "Light-gaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps," *NeurIPS*, 2023.
- [176] M. Niemeyer, F. Manhardt, M.-J. Rakotosaona, M. Oechsle, D. Duckworth, R. Gosula, K. Tateno, J. Bates, D. Kaeser, and F. Tombari, "Radsplat: Radiance field-informed gaussian splatting for robust real-time rendering with 900+ fps," *arXiv preprint arXiv:2403.13806*, 2024.
- [177] W. Liu, T. Guan, B. Zhu, L. Ju, Z. Song, D. Li, Y. Wang, and W. Yang, "Efficients: Streamlining gaussian splatting for large-scale high-resolution scene representation," *arXiv preprint arXiv:2404.12777*, 2024.
- [178] A. Hanson, A. Tu, V. Singla, M. Jayawardhana, M. Zwicker, and T. Goldstein, "Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting," *arXiv preprint arXiv:2406.10219*, 2024.
- [179] P. Papantonakis, G. Kopanas, B. Kerbl, A. Lanvin, and G. Drettakis, "Reducing the memory footprint of 3d gaussian splatting," *Proc. of the ACM on Computer Graphics and Interactive Techniques*, vol. 7, no. 1, pp. 1–17, 2024.
- [180] O. Mahmoud and M. Gendrin, "On reducing the number of gaussians for radiance field real-time rendering," in *2024 ECTI DAMT & NCON*. IEEE, 2024, pp. 259–264.
- [181] S. Subhajyoti Mallick, R. Goel, B. Kerbl, F. V. Carrasco, M. Steinberger, and F. De La Torre, "Taming 3dgs: High-quality radiance fields with limited resources," *arXiv e-prints*, pp. arXiv-2406, 2024.
- [182] S. Niedermayr, J. Stumpfegger, and R. Westermann, "Compressed 3d gaussian splatting for accelerated novel view synthesis," in *CVPR*, 2024, pp. 10349–10358.
- [183] Y. Chen, Q. Wu, J. Cai, M. Harandi, and W. Lin, "Hac: Hash-grid assisted context for 3d gaussian splatting compression," *arXiv preprint arXiv:2403.14530*, 2024.
- [184] X. Liu, X. Wu, P. Zhang, S. Wang, Z. Li, and S. Kwong, "Compgs: Efficient 3d scene representation via compressed gaussian splatting," *arXiv preprint arXiv:2404.09458*, 2024.
- [185] Y. Wang, Z. Li, L. Guo, W. Yang, A. C. Kot, and B. Wen, "Contextgs: Compact 3d gaussian splatting with anchor level context model," *arXiv preprint arXiv:2405.20721*, 2024.
- [186] W. Morgenstern, F. Barthel, A. Hilsmann, and P. Eisert, "Compact 3d scene representation via self-organizing gaussian grids," *arXiv preprint arXiv:2312.13299*, 2023.
- [187] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, "Fsgs: Real-time few-shot view synthesis using gaussian splatting," *arXiv preprint arXiv:2312.00451*, 2023.
- [188] J. Chung, J. Oh, and K. M. Lee, "Depth-regularized optimization for 3d gaussian splatting in few-shot images," in *CVPR*, 2024, pp. 811–820.
- [189] A. Paliwal, W. Ye, J. Xiong, D. Kotovenko, R. Ranjan, V. Chandra, and N. K. Kalantari, "Coherentgs: Sparse novel view synthesis with coherent 3d gaussians," *arXiv preprint arXiv:2403.19495*, 2024.
- [190] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefey, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," *arXiv preprint arXiv:2403.14627*, 2024.
- [191] H. Xiong, S. Mutukuru, R. Upadhyay, P. Chari, and A. Kadambi, "Sparsegs: Real-time 360° sparse view synthesis using gaussian splatting," *arXiv preprint arXiv:2312.00206*, 2023.
- [192] Z. Bao, G. Liao, K. Zhou, K. Liu, Q. Li, and G. Qiu, "Loopsparsesgs: Loop based sparse-view friendly gaussian splatting," *arXiv preprint arXiv:2408.00254*, 2024.
- [193] J. Zhang, J. Li, X. Yu, L. Huang, L. Gu, J. Zheng, and X. Bai, "Corgs: Sparse-view 3d gaussian splatting via co-regularization," *arXiv preprint arXiv:2405.12110*, 2024.
- [194] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, "Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization," in *CVPR*, 2024, pp. 20775–20785.

- [195] Z. Fan, W. Cong, K. Wen, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, "Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds," *arXiv preprint arXiv:2403.20309*, 2024.
- [196] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv:1906.05797*, 2019.
- [197] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017, pp. 5828–5839.
- [198] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ int. conf. on intelligent robots and systems*, 2012, pp. 573–580.
- [199] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE trans. on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [200] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
- [201] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.
- [202] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020, pp. 2446–2454.
- [203] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020, pp. 11621–11631.
- [204] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, 2019, pp. 8748–8757.
- [205] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.
- [206] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *CVPR*, 2013, pp. 2930–2937.
- [207] A. Mirzaei, T. Aumentado-Armstrong, K. G. Derpanis, J. Kelly, M. A. Brubaker, I. Gilitschenski, and A. Levinstein, "Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields," in *CVPR*, 2023, pp. 20669–20679.
- [208] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *ICCV*, 2023, pp. 19729–19739.
- [209] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *RA-L*, vol. 5, no. 2, pp. 3019–3026, 2020.
- [210] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," *arXiv preprint arXiv:2108.03298*, 2021.
- [211] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. conf. on machine learning*. PMLR, 2018, pp. 1861–1870.
- [212] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.
- [213] K. Yadav, R. Ramrakhyta, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva *et al.*, "Habitat-matterport 3d semantics dataset," in *CVPR*, 2023, pp. 4927–4936.
- [214] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *ICCV*, 2019, pp. 9339–9347.
- [215] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.
- [216] C. R. Givens and R. M. Shortt, "A class of wasserstein metrics for probability distributions." *Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.
- [217] A. Adkins, T. Chen, and J. Biswas, "Obvi-slam: Long-term object-visual slam," *RA-L*, 2024.
- [218] M. Zhao, X. Guo, L. Song, B. Qin, X. Shi, G. H. Lee, and G. Sun, "A general framework for lifelong localization and mapping in changing environment," in *IROS*, 2021.
- [219] Y. Wu, L. Pan, W. Wu, G. Wang, Y. Miao, and H. Wang, "Rlisbridge: 3d gaussian splatting based real2sim2real method for robotic manipulation learning," *arXiv:2409.20291*, 2024.
- [220] T. Zhang, H.-X. Yu, R. Wu, B. Y. Feng, C. Zheng, N. Snavely, J. Wu, and W. T. Freeman, "Physdreamer: Physics-based interaction with 3d objects via video generation," *arXiv preprint arXiv:2404.13026*, 2024.