2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

PanopticSplatting: End-to-End Panoptic Gaussian Splatting

Yuxuan Xie[1], Xuan Yu[1], Changjian Jiang[1], Sitong Mao[2], Shunbo Zhou[2], Rui Fan[3], Rong Xiong[1], Yue Wang[1]

***Abstract*— Open-vocabulary panoptic reconstruction is a challenging task for simultaneous scene reconstruction and understanding. Recently, methods have been proposed for 3D scene understanding based on Gaussian splatting. However, these methods are multi-staged, suffering from the accumulated errors and the dependence of hand-designed components. To streamline the pipeline and achieve global optimization, we propose PanopticSplatting, an end-to-end system for open-vocabulary panoptic reconstruction. Our method introduces query-guided Gaussian segmentation with local cross attention, lifting 2D instance masks without cross-frame association in an end-to-end way. The local cross attention within view frustum effectively reduces the training memory, making our model more accessible to large scenes with more Gaussians and objects. In addition, to address the challenge of noisy labels in 2D pseudo masks, we propose label blending to promote consistent 3D segmentation with less noisy floaters, as well as label warping on 2D predictions which enhances multi-view coherence and segmentation accuracy. Our method demonstrates strong performances in 3D scene panoptic reconstruction on the ScanNet-V2 and ScanNet++ datasets, compared with both NeRF-based and Gaussian-based panoptic reconstruction methods. Moreover, PanopticSplatting can be easily generalized to numerous variants of Gaussian splatting, and we demonstrate its robustness on different Gaussian base models.**

I. INTRODUCTION

Open-world panoptic reconstruction is a significant task in 3D scene understanding for robotics. Since the high cost of 3D annotation and the remarkable progress in 2D open-vocabulary segmentation [1], [2], most of the existing methods [3]–[5] lift the ability of 2D VLMs [1], [2], [6], [7] to 3D for 3D open-vocabulary segmentation.

3D panoptic reconstruction methods of this type are initially based on NeRF. As a representative work, Panoptic Lifting [8] proposes a multi-view consistent label lifting scheme with linear assignment. Further, PanopticRecon++ [5] uses cross attention to introduce 3D spatial priors, improving the performance of end-to-end panoptic reconstruction. However, these methods suffer from intensive computation and slow rendering speed due to the implicit representation and random sampling. Besides, the simultaneous reconstruction and segmentation enables 3D scene editing applications, while NeRF-based methods cannot be easily applied to these tasks, since objects are implicitly encoded in weights of neural networks [9].

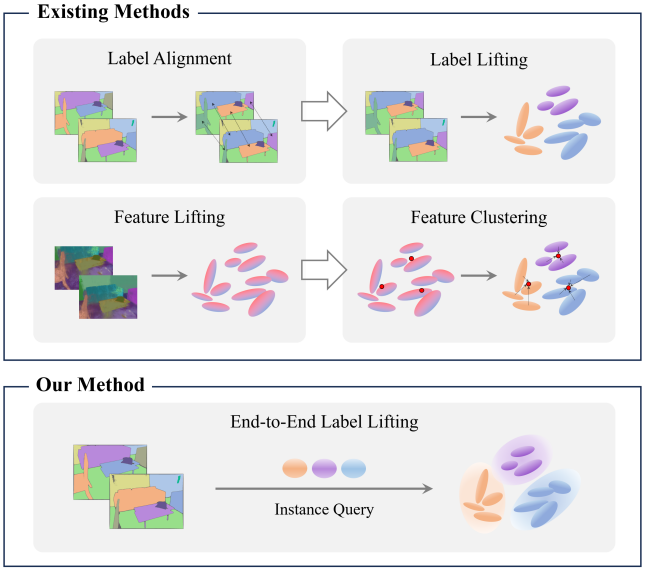

Fig. 1. Existing methods for 3D panoptic reconstruction based on Gaussian splatting have multi-staged procedures. They can be broadly divided into two categories: label-lifting methods decompose the task into 2D mask alignment and label lifting; feature-lifting methods require hand-designed post-processing after feature distillation. Our method introduces instance queries to guide end-to-end panoptic reconstruction.

In comparison, 3D Gaussian Splatting has emerged as an explicit representation which shows a significant advantage in rendering speed, and allows editing in a simple way. Existing methods [10]–[15] extend it to 3D scene understanding by adding feature attributes to Gaussians. Some of them [13], [15] propose effective solutions to align and correct 2D machine-generated masks, and lift them to 3D. Others [10]–[12], [14] focus on consistent and efficient 2D feature distillation to achieve language-aligned 3D scene understanding. However, they are all multi-staged, which causes the decrease in performance, including inconsistent label lifting caused by errors in prior stages, and the limited generalization due to the manual components.

To solve this issue, we aim to learn an end-to-end Gaussian-based model for 3D open-vocabulary panoptic reconstruction. First, we focus on how the existing methods optimize the instance segmentation to analyze the limitations restricting them from end-to-end optimization. The Gaussian-based label lifting methods learn fixed instance labels from 2D masks and lack instance optimization in 3D scenes, which prevents them from learning consistent segmentation from misaligned labels. And the feature lifting methods

[1]Yuxuan Xie, Xuan Yu, Changjian Jiang, Rong Xiong, and Yue Wang are with Zhejiang University, Hangzhou, Zhejiang, China. Yue Wang is the corresponding author wangyue@iipc.zju.edu.cn

[2]Sitong Mao and Shunbo Zhou are with Huawei Cloud Computing Technologies Co., Ltd., Shenzhen, China.

[3]Rui Fan is with the College of Electronics & Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: rui.fan@ieee.org).

locate instances based on optimized feature fields, resulting in phased optimization. Based on the above analysis, we believe that the key to end-to-end learning is simultaneous optimization of feature fields and instance segmentation, which is challenging to implement due to the difficulty in optimizing instances that implicitly exist in the feature field. To address this challenge, we introduce learnable queries to explicitly model the instances in 3D and propose query-guided Gaussian segmentation for end-to-end panoptic reconstruction.

In this paper, we propose PanopticSplatting, an open-vocabulary panoptic reconstruction method based on Gaussian Splatting, deploying end-to-end training under the supervision of 2D masks from VLMs [2]. Specifically, build upon a set of Gaussians, we add feature attributes to Gaussians to model the semantic and instance fields separately. To achieve end-to-end instance reconstruction, learnable instance queries are introduced to guide 3D Gaussian segmentation, with cross attention to model the similarity between queries and Gaussians, followed by linear assignment between predicted instances and pseudo masks. We save training memory without affecting the performance by constraining the cross attention within view frustum. To mitigate the effects of 2D noisy labels in semantic reconstruction, we perform label blending instead of common feature blending to enhance 3D segmentation consistency, and propose label warping across views for multi-view consistent segmentation. In summary, our main contributions are as follows:

- We propose an end-to-end Gaussian-based method for open-vocabulary panoptic reconstruction by query-guided Gaussian segmentation.
- We introduce label blending and label warping to mitigate the effects of 2D noisy labels, and reduce the memory cost of query-guided Gaussian segmentation by constraining cross-attention within view frustum.
- PanopticSplatting shows strong performances in 3D scene panoptic reconstruction compared with existing methods, and we demonstrate its robustness on different Gaussian base models.

II. RELATED WORKS

A. Nerf-based Panoptic Reconstruction

Neural Radiance Field methods encode the scene into a neural network, offering an implicit representation to various properties of 3D scenes, including appearance [16], [17], geometry [18], semantic [19], [20], etc. Since NeRF effectively connects 2D images and 3D scenes, numerous follow-up [8], [19] works build neural feature fields to understand 3D scenes through 2D vision knowledge. Closer to our work are methods that employ NeRFs to address the problem of 3D panoptic segmentation. Panoptic Lifting [8] proposes a label lifting scheme with a linear assignment between predictions and unaligned instance labels to build a multi-view consistent 3D panoptic representation. Contrastive Lift [21] achieves 3D object segmentation without quantity limitation, through a slow-fast clustering objective function using contrastive learning. PVLFF [3] also uses contrastive learning to build an instance feature field, and achieves open-vocabulary panoptic segmentation by distilling features of 2D VLMs. PanopticRecon [4] proposes a two-stage method with label propagation and 2D instance association to align 2D masks. PanopticRecon++ [5] introduces an end-to-end framework for 3D panoptic reconstruction, using Gaussian-modulated instance tokens to guide 3D instance segmentation.

Although these methods perform well in 3D panoptic reconstruction based on NeRF, they suffer from the drawbacks of this implicit and continuous representation. First, they require a significant cost of computation and time for training and rendering, due to the large neural networks and large number of stochastic samples during volumetric ray-marching. Besides, since the implicit and continuous properties of these representations, there are challenges for them to apply to scene editing tasks, which are popular in many areas such as augmented reality and gaming.

B. Gaussian-based Scene Understanding

3DGS [22] and its numerous variants [23], [24] optimize a set of differentiable Gaussians to create compact and flexible representations of the 3D scene, providing outstanding results in appearance reconstruction. In addition, the tile-based rasterizer guarantees real-time rendering and low memory consumption of Gaussian splatting. Similar to NeRF-based feature fields, existing methods [10]–[14] extend it to 3D scene understanding by adding feature attributes for each Gaussian. LEGaussians [10] proposes a language embedded procedure with the quantization scheme and uncertainty modeling to introduce semantics to Gaussians in an efficient and consistent way. LangSplat [11] designs a scene-wise language autoencoder to reduce feature dimensions and learns hierarchical semantics to address ambiguity in language fields. Feature 3DGS [12] introduces a parallel N-dimensional rasterizer with a speed-up module to distill high-dimensional features. Gaussian Grouping [13] uses a tracking method to associate the 2D mask generated by SAM, and then lift the consistent masks to group Gaussians with a 3D regularization loss. To enhance 3D point-level scene understanding, OpenGaussian [14] proposes a multi-stage pipeline that first learns and clusters the instance feature in 3D, followed by 3D-2D feature association. PLGS [15] introduces semantic anchor points and the self-training approach for robust training and generates consistence instance masks through 3D matching.

For 3D panoptic reconstruction task, these existing methods require multi-staged pipelines. The feature lifting [10]–[12], [14] methods allow open-vocabulary querying and segmentation for a single object through language queries, while when addressing scene panoptic segmentation with objects of uncertain quantity, they need more hand-designed components such as clustering the similar features and setting segmentation thresholds. The label lifting [13], [15] methods suffer from the 2D masks without association across views, thus they need to align the 2D labels first. Our method aims to build a Gaussian-based end-to-end panoptic reconstruction

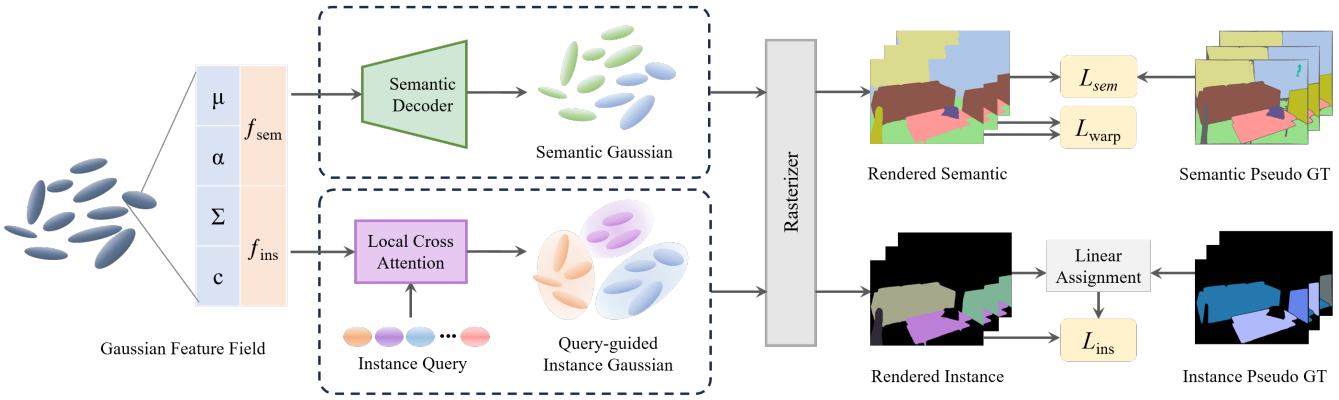

Fig. 2. PanopticSplatting introduces a semantic feature and an instance feature to Gaussians to build semantic and instance feature field. In instance branch, Gaussian-modulated instance queries are introduced to guide Gaussian segmentation through local cross attention. The semantic labels of Gaussians are generated by a simple semantic decoder. Then the labels of Gaussians are rendered to 2D simultaneously. To achieve end-to-end training, linear assignment between 2D pseudo instance masks and predicted labels is performed. We employ the label warping loss on rendered semantic masks.

system, employing 3D consistency to lift 2D labels in a multi-view unified way.

III. METHOD

A. Gaussian-based Feature Representation

Gaussian Splatting [22] uses 3D Gaussians with geometric and appearance attributes to reconstruct 3D scenes. It employs fast differentiable rasterization to render RGB images from any view. During rasterization, 3D Gaussians are first sorted by depth and projected onto the image plane. Then the color of each pixel can be computed via α-blending based on the color attribute:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_i') \tag{1}$$

where α_i' represents the influence intensity of Gaussian on this pixel, determined by the opacity of the 3D Gaussian and the distribution of the projected 2D Gaussian.

To extend 3D Gaussians to panoptic reconstruction, we model the semantic and instance of Gaussians by additionally introducing a semantic feature and an instance feature to them. The features are represented by a learnable embedding, denoted as f_{sem} and f_{ins} respectively. Note that unlike the anisotropic color attribute, the semantics and instances of Gaussians are consistent across rendering views, thus they can be represented in a view-independent way. Similar to rendering RGB images, we can obtain 2D semantics and instances by rasterization.

B. Instance Reconstruction

We lift 2D instance masks to build the 3D instance feature field, while the instance IDs across views are misaligned, which cannot become a consistent supervision for scene-level instance segmentation. Thus, to associate 2D labels by using 3D consistency, we introduce instance queries to guide the instance segmentation on the 3D feature field, ensuring consistent instance labels in 3D.

Query-Guided Gaussian Segmentation. One of the common paradigms for end-to-end instance segmentation is to use object queries to guide object detection [25] and segmentation [26]. Through the interaction between scene features and query features, queries learn object features and guide the scene instance feature clustering. Then, the constituent units of the scene, such as 2D pixels or 3D points, can be segmented through the similarity with queries.

In 3D scene segmentation, previous work [5] uses spatial distance weighted attention map to model this similarity between 3D points and queries which are encoded as 3D Gaussian distribution. We also adopt this similarity representation, using distance-aware cross attention between instance queries and instance feature of Gaussians to segment Gaussians in 3D.

Considering that the scene Gaussians are far smaller in size than the Gaussian-modulated queries, we simplify the scene Gaussians into points when performing cross attention with queries. First, the similarity between query feature f_q and Gaussian instance feature f_{ins} is defined as:

$$S(f_q, f_{ins}) = sigmoid(f_q^T f_{ins}) \tag{2}$$

Besides, since the instance queries are Gaussian distributed, the distance from a 3D point to the query can be described by the probability density. The distance between query q and scene Gaussian g is defined as:

$$D(p_q, p_g) = \frac{\varphi(p_g)}{\varphi(p_q)} \tag{3}$$

where $\varphi(\cdot)$ is the probability density function of query's Gaussian distribution. p_q and p_g are the center point coordinates of the instance query and scene Gaussian, respectively. Thus, the attention map is as follows:

$$A(q, g) = S(f_q, f_{ins}) D(p_q, p_g) \tag{4}$$

The instance label of Gaussian g is defined as:

$$l_{ins}(g) = softmax([A(q_1, g) \dots A(q_i, g) \dots A(q_N, g)]) \tag{5}$$

where N is the number of queries.

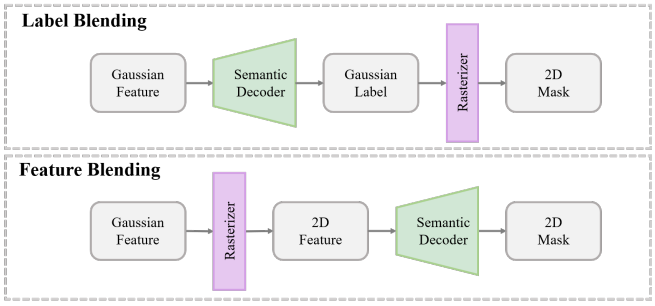

Fig. 3. The pipelines of label blending and feature blending.

Then, similar to (1), the instance label of pixels can be computed via α-blending the Gaussian labels:

$$I = \sum_{i \in \mathcal{N}} l_{ins}(i)\alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_i) \tag{6}$$

Local Cross Attention. Considering the large number of Gaussians in the entire scene, cross attention between Gaussians and queries requires high computational cost, thus we introduce local cross attention to limit the usage of training memory and shorten the training time. For each training view, only Gaussians within the view frustum will update via back propagation during the training process. Thus, we perform local cross attention, which limits the cross attention within view frustum.

Following the previous work [22], Gaussians within the view frustum are defined as Gaussians with a 99% confidence interval intersecting the view frustum. We implement local cross attention in CUDA kernel. To be specific, after filtering out the Gaussians within the view frustum, we launch one thread for each selected Gaussian to compute cross attention in parallel. Thus, the attention map is obtained efficiently as the instance feature of 3D valid Gaussians. Then, as shown in (6), we perform tile-based rasterization to obtain 2D instance maps. This strategy effectively reduces the training memory compared with global cross attention, especially when the number of scene Gaussians increases, making our model more accessible to large scenes with more Gaussians.

C. Semantic Reconstruction

The Gaussian semantic field can be learned through lifting the 2D semantic labels, while we find that the predictions struggle with multi-view semantic inconsistency due to the noisy labels in 2D pseudo masks. To address this issue, we propose label blending to promote consistent 3D segmentation compared with feature blending, and introduce label warping loss on 2D predictions to enhance segmentation accuracy and correct errors in pseudo masks.

Label Blending. During the rasterization, different from feature blending in the previous work [13], we obtain the 2D semantic mask by α-blending the semantic label of Gaussians instead of the semantic feature. Specifically, the Gaussian semantic feature f_{sem} is first decoded by an MLP and a softmax layer to gain the semantic label of the Gaussian, donated as $l_{sem} \in \mathbb{R}^{N_c}$, where N_c is the number of categories. Then, the 2D semantic mask can be computed via label blending:

$$S = \sum_{i \in \mathcal{N}} l_{sem}(i)\alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_i) \tag{7}$$

Fig. 3 shows the pipelines of label blending and feature blending. In label blending, Gaussians are first classified in 3D according to Gaussian semantic features, followed by splatting to 2D and blending. While feature blending pipeline learns a semantic head in 2D, whose strong fitting ability weakens the categories of Gaussians. Thus, label blending emphasizes the 3D Gaussian segmentation which enhances the 3D consistency to ease the impact of 2D noisy labels. Besides, in label blending pipeline, the softmax layer normalizes Gaussian feature before blending, avoiding that the 2D predictions are dominated by Gaussians with considerable values while other Gaussians with wrong labels along this direction may not be punished. In the ablation study, a detailed comparison is provided between semantic label blending and feature blending.

Label Warping Loss. Due to the lack of correlation between discrete 3D Gaussians, the feature field based on Gaussian has weaker smoothness compared with NeRF-based methods, which is manifested as the semantic inconsistency across views in 2D predictions under the impact of noisy labels. Thus, to enhance the association and consistency between multi-view label predictions and correct the few semantic errors in 2D, we propose a per-pixel label warping loss on 2D semantic predictions.

For a pixel in frame m, denoted as r_m, we first unproject it to 3D using depth and intrinsic to get the pointcloud p_m in the coordinate system of frame m:

$$\begin{bmatrix} p_m \\ 1 \end{bmatrix} = D(r_m)K^{-1}\begin{bmatrix} r_m \\ 1 \end{bmatrix} \tag{8}$$

where $D(r_m)$ donates the depth on pixel r_m and K donates the intrinsic of camera. Then we transfer the pointcloud to the coordinate system of the adjacent frame n:

$$\begin{bmatrix} p_{m\to n} \\ 1 \end{bmatrix} = T_n T_m^{-1}\begin{bmatrix} p_m \\ 1 \end{bmatrix} \tag{9}$$

where T_m and T_n denote the extrinsic matrix of frame m and n. Then the projected pixel $r_{m\to n}$ in the nearby frame n can be generated by projecting $p_{m\to n}$ to frame n as the reverse process of (8). The label warping loss is then defined as:

$$\mathcal{L}_{warp}(r_m) = \sum_{n \in \mathcal{K}_m} \|M_{sem}(r_m) - M_{sem}(r_{m\to n})\| \tag{10}$$

where $\mathcal{K}_m$ denotes the adjacent frame list for the current frame m. We mask out the pixels that are projected outside the image boundary of frame n.

D. End-to-End Training

Instance Assignment. To achieve end-to-end 2D instance supervision, we use the Hungarian Algorithm for linear assignment between pseudo instance groundtruth and predicted labels on each frame.

TABLE I

PANOPTIC SEGMENTATION QUALITY USING DIFFERENT METHODS

Method	ScanNet-V2							ScanNet++						
	PQ	SQ	RQ	mIoU	mAcc	mCov	mW-Cov	PQ	SQ	RQ	mIoU	mAcc	mCov	mW-Cov
Panoptic Lifting	57.86	61.96	85.31	67.91	78.59	45.88	59.93	71.14	77.48	88.14	81.34	**89.67**	56.17	68.51
Contrastive Lift	37.35	41.91	57.60	64.77	75.80	13.21	23.26	47.58	57.23	65.81	81.09	89.30	27.39	36.51
PVLFF	30.11	51.71	44.43	55.41	63.96	45.75	48.41	52.24	66.86	65.56	62.53	70.31	67.95	75.47
PanopticRecon	63.70	64.81	81.17	68.62	80.87	66.58	77.84	68.29	77.01	85.05	77.75	87.08	51.34	62.79
Gaussian Grouping	43.75	50.63	72.68	58.05	68.68	52.70	58.10	33.10	40.60	67.27	59.53	68.13	29.83	36.83
OpenGaussian	48.73	51.48	88.10	54.05	68.43	44.43	49.60	51.03	56.93	85.73	61.80	73.97	50.00	51.02
Ours	**74.75**	**74.75**	**100.0**	**74.95**	**83.70**	**73.18**	**79.63**	**77.73**	**82.70**	**93.60**	**81.90**	89.50	**74.73**	**78.03**

Specifically, we first calculate the matching cost between the binary masks in a pair of pseudo GT and predicted masks. Since the number of instance queries is larger than GT, each GT instance will be distributed a predicted mask in Hungarian Algorithm. We get the the optimal assignment that builds the connection between GT instances and predicted ones by minimizing the total assignment cost.

Loss Function. After instance assignment, we use dice loss and binary cross-entropy loss between a pair of instance prediction M_{ins} and aligned instance GT M_{ins}^{gt} to supervise the instance branch. The instance loss $\mathcal{L}_{ins}$ is defined as:

$$\mathcal{L}_{ins} = \mathcal{L}_{dice}(M_{ins}, M_{ins}^{gt}) + \mathcal{L}_{bce}(M_{ins}, M_{ins}^{gt}) \tag{11}$$

The semantic branch is supervised by a cross-entropy loss between rendered semantic M_{sem} and semantic GT M_{sem}^{gt}:

$$\mathcal{L}_{sem} = \mathcal{L}_{ce}(M_{sem}, M_{sem}^{gt}) \tag{12}$$

The image rendering is supervised with the groundtruth images by a combination of L1 loss and SSIM loss. With the rendered image denoted as I and the groundtruth donated as I^{gt}, the rendering loss is defined as:

$$\mathcal{L}_{rgb} = (1 - \lambda_{SSIM})\mathcal{L}_1(I, I^{gt}) + \lambda_{SSIM}\mathcal{L}_{SSIM}(I, I^{gt}) \tag{13}$$

Combined with the label warping loss in (10), our total loss $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{ins} + \mathcal{L}_{sem} + \mathcal{L}_{warp} \tag{14}$$

IV. EXPERIMENTS

We validate the performance of our method in real-world datasets, comparing with related works based on NeRF and Gaussian. We also deploy our method on several representative Gaussian models to illustrate the robustness to Gaussian base models. Besides, we conduct ablation studies to demonstrate the effectiveness of network components.

A. Setup

Datasets. We leverage two real-world datasets for evaluating our method: ScanNet-V2 [27] and ScanNet++ [28]. ScanNet-V2 is an RGB-D dataset that contains diverse real-world indoor scenes with images, labels and reconstructed geometry, which makes it suitable for 3D reconstruction and segmentation. ScanNet++ provides high-resolution 3D scenes with fine annotations, crucial for novel view synthesis and 3D scene understanding tasks. We use 4 scenes in ScanNet-V2 and 3 scenes in ScanNet++ for evaluation.

Baselines. We select several baselines for 3D panoptic segmentation based on Nerf and Gaussian. Panoptic Lifting [8], Contrastive Lift [21], PVLFF [3], and PanopticRecon [4] are representative scene segmentation methods based on NeRF. Gaussian Grouping [13] and OpenGaussian [14] are Gaussian-based methods that lift 2D vision knowledge for scene understanding. We fairly use the same 2D labels with our method to supervise comparison methods except PVLFF and Gaussian Grouping, for which we use their default 2D mask generating methods. For OpenGaussian, we add post-processing on its binary mask predictions to generate panoptic masks.

Evaluation Metrics. We evaluate the scene-level segmentation metrics in 2D predictions. Compared with image-level segmentation, scene-level segmentation requires multi-view consistent identities for the same instance, which ensures the segmentation consistency in 3D. Following previous work [8], we evaluate panoptic quality (PQ), semantic quality (SQ) and recognition quality (RQ) for 2D panoptic segmentation, together with mIoU, mAcc for semantic segmentation and mCov, mWCov for instance segmentation.

Implementation. We implement Grounded-SAM [2] to generate 2D semantic and instance masks as supervision, with the IDs of the same instance in multi-view masks being inconsistent. In comparative studies with baselines and ablation studies, we deploy our method on LI-GS [24], a scene reconstruction model based on Gaussian surfels. The semantic feature of Gaussians has 16 dimensions, and the instance feature of Gaussians and queries has 32 dimensions.

B. Comparative Study

Comparative Study with Baselines. The quantitative results on ScanNet-V2 and ScanNet++ datasets are shown in Tab.I. Fig. 4 and Fig. 5 present the visualization of the comparative results with the NeRF-based and Gaussian-based methods on the two datasets.

Our method significantly outperforms both the NeRF-based and Gaussian-based methods. We attribute the seg-

Fig. 4. Comparison of the quality of panoptic segmentation and semantic segmentation of NeRF-based methods on ScanNet-V2 and ScanNet++.

mentation errors of the comparative methods to the following reasons, and prove the superiority of PanopticSplatting.

First, the methods with multi stages, such as PanopticRecon and Gaussian Grouping, suffer from accumulated errors. As shown in "Scene0628_02" (Fig. 4), PanopticRecon fails to segment the "bag" on the chair, which is caused by the under-segmentation in its first phase. In Gaussian Grouping, 2D consistent labels generated by tracking is not optimistic in indoor scenes, especially in the ScanNet++ dataset where there is a significant view change between adjacent frames. Thus, the errors in label association lead to the failure of segmentation, as shown in "Scene1ada7a0617" (Fig. 5). Compared with them, our method proposes an end-to-end pipeline to effectively avoid accumulated errors and achieve global optimization.

Besides, the non-unique label of 3D instance, manifested as different predictions of the same instance across views, leads to the low scene-level metrics of PanopticLifting and OpenGaussian. Compared with their implicit scene instance representation, we leverage query tokens to explicitly model the 3D instances, learning both the instance feature and 3D spatial priors, which promotes that the masks of one instance on different views correspond to a unique 3D instance.

The feature-lifting methods struggle to accurately separate objects with similar features. For example, Contrastive Lift fails to segment the chairs in the "Scene0088_00" (Fig. 4), since the chairs share similar features. OpenGaussian uses two-level discretization to add 3D spatial priors, while the spatial priors introduced by simple clustering are coarse. In contrast, our method introduces deformable and learnable instance queries to learn more precision 3D spatial priors, and deploys distance-weighted similarity, which effectively guides spatial-aware Gaussian segmentation.

In addition, to address noisy labels, we enhance the multi-view semantic segmentation consistency by label blending and label warping, which effectively improve segmentation accuracy of stuff categories. As shown in "Scene0420_01" (Fig. 5), only our method accurately segment the "door" among the Gaussian-based methods.

Study on Model Robustness. We verify the robustness of our model on typical and novel Gaussian bases, including 3DGS [22], 2DGS [23], and LI-GS [24]. Tab.II shows the results on three bases. As we can see, our method achieves good performance on these base models, which demonstrates that it can adapt to different Gaussian models and handle scene panoptic reconstruction methods on different Gaussian representations. In addition, LI-GS based model achieves the best segmentation performance with the fewest number of Gaussians, mainly benefiting from its excellent geometric reconstruction ability with less floaters. Therefore, we use LI-GS as the base model in other experiments.

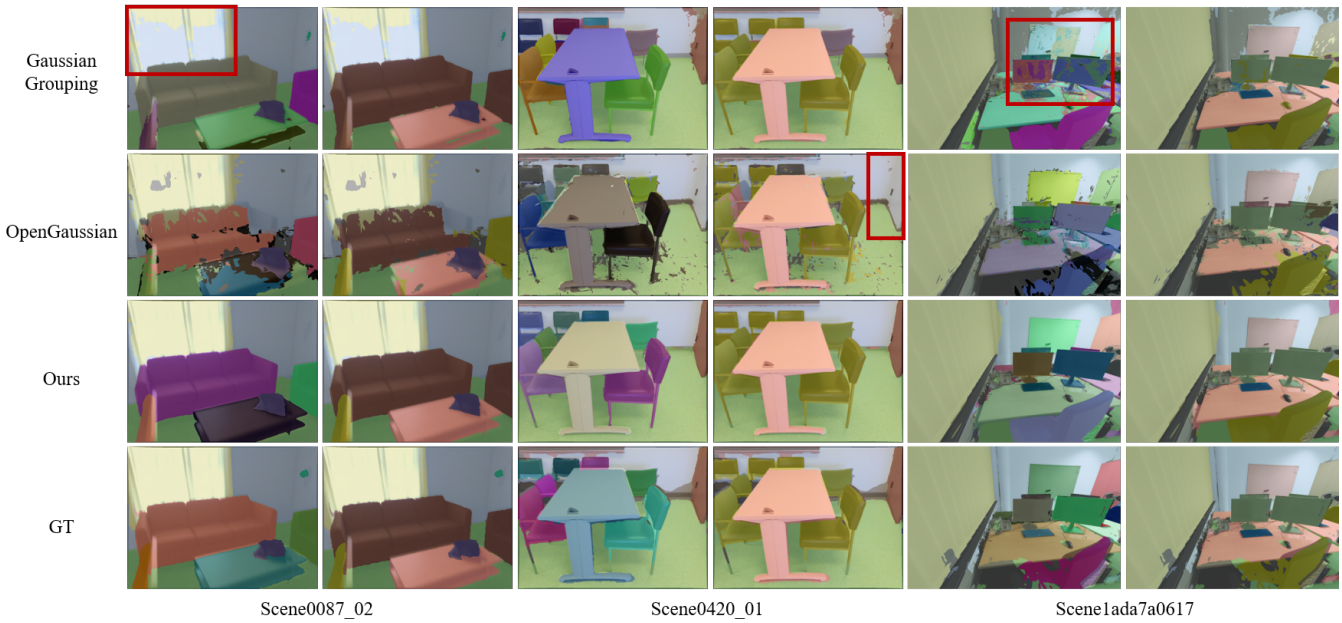

Fig. 5. Comparison of the quality of panoptic segmentation and semantic segmentation of Gaussian-based methods on ScanNet-V2 and ScanNet++.

TABLE II
STUDY ON MODEL ROBUSTNESS ON SCANNET-V2 DATASET

	PQ	SQ	RQ	mIoU	mAcc	mCov	mW-Cov
3DGS	73.83	74.08	99.33	**75.08**	83.68	72.20	76.43
2DGS	73.30	73.45	99.93	73.75	82.58	71.28	75.98
LI-GS	**74.75**	**74.75**	**100.0**	74.95	**83.70**	**73.18**	**79.63**

TABLE III
ABLATION STUDIES ON LABEL BLENDING AND LABEL WARPING

	PQ	SQ	RQ	mIoU	mAcc	mCov	mW-Cov
FB	72.43	73.73	98.60	73.13	82.95	71.03	77.90
LB	73.50	74.60	98.68	74.38	82.73	73.00	79.40
LB+WP	**74.75**	**74.75**	**100.0**	**74.95**	**83.70**	**73.18**	**79.63**

C. Ablation Study

We conduct ablation experiments of the designs in our model to validate the effectiveness of our method. All ablation experiments are performed on the ScanNet-V2 dataset.

Label Blending. We compare the performance of label blending and feature blending in semantic branch to validate the advantage of label blending. As shown in Tab.III, the label blending (LB) outperforms the feature blending (FB) on both semantic and instance segmentation.

Specifically, the panoptic metrics show that when the recognition quality (RQ) is comparable, label blending achieves better semantic quality (SQ) than feature blending, which means label blending predicts more precise masks for objects. This is mainly attributed to the 3D segmentation consistency enhanced by semantic segmentation on 3D Gaussians instead of 2D features, and less noisy floaters with normalization on 3D Gaussians.

Label Warping Loss. We study the influence of the label warping loss on the segmentation accuracy. As shown in Tab.III, adding label warping loss (WP) achieves effective improvement of segmentation metrics, especially on semantic segmentation accuracy, which demonstrates that the model profits from multi-view consistency to alleviate the impact of noisy labels.

TABLE IV
ABLATION STUDY ON LOCAL CROSS ATTENTION

	Base	N_{gs}	PQ	mIoU	mCov	Memory	Train	FPS
Global CA	LI-GS	296K	74.60	74.87	**73.26**	18.59G	6h18m	3.20
	2DGS	569K	73.56	73.80	71.13	28.21G	7h51m	2.93
Local CA	LI-GS	296K	**74.75**	**74.95**	73.18	**12.21G**	**4h25m**	**3.98**
	2DGS	569K	73.30	73.75	71.28	**13.39G**	**4h42m**	**3.95**

Fig. 6 shows the ablation on label blending and label warping loss. As we can see, noisy labels in View 2 of pseudo GT lead to obvious multi-view inconsistency in the predictions of feature blending. The label blending effectively addresses the inconsistency by enhancing 3D consistency of semantic segmentation, and the warping loss further corrects the incorrect labels.

Local Cross Attention. We compare the performance, training memory cost, training and inference times between using local cross attention and global cross attention based on LI-GS and 2DGS. As shown in Tab.IV, the local cross attention between Gaussians and queries effectively decreases the memory cost without affecting the performance. In particular, when the number of scene Gaussians N_{gs} increases, the memory of local cross attention increases significantly less than that of global, which makes our model fit to large scenes with more Gaussians. Besides, the local cross attention reduces training and inference time.

V. CONCLUSION

In this paper, we propose PanopticSplatting, an end-to-end panoptic reconstruction method based on Gaussian splatting. Our method first builds Gaussian semantic and instance

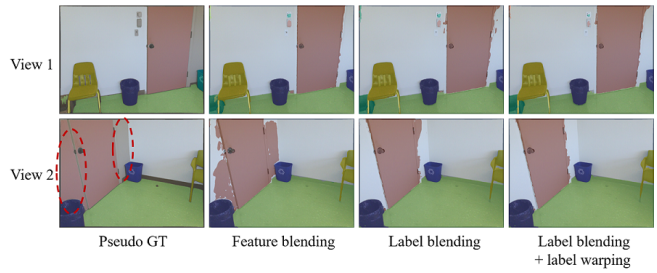

Fig. 6. Ablation on label blending and label warping loss. The noisy labels in View 2 of pseudo GT lead to obvious multi-view inconsistency in feature blending. The label blending effectively addresses this by enhancing 3D segmentation consistency, and the warping loss further correct the incorrect labels through multi-view consistency.

fields by adding features to each Gaussian. Then query-guided 3D Gaussian segmentation and linear assignment between instance predictions and pseudo GT ensure the end-to-end instance reconstruction. To address the challenge of noisy labels in 2D pseudo masks, we further introduce label blending and label warping to promote consistent segmentation and enhance segmentation accuracy. PanopticSplatting demonstrates strong performance in numerous scenes and good generalization on different Gaussian base models.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant 62373322, Grant 62473288 and Grant 62233013, Zhejiang Provincial Natural Science Foundation of China under Grant No. LD25F030001, Research and Development Project of Zhejiang Province under Grant No. 2025C01205(SD2), the Fundamental Research Funds for the Central Universities, NIO University Programme (NIO UP), and the Xiaomi Young Talents Program.

References

[1] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[2] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[3] H. Chen, K. Blomqvist, F. Milano, and R. Siegwart, "Panoptic vision-language feature fields," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2144–2151, 2024.

[4] X. Yu, Y. Liu, C. Han, S. Mao, S. Zhou, R. Xiong, Y. Liao, and Y. Wang, "Panopticrecon: Leverage open-vocabulary instance segmentation for zero-shot panoptic reconstruction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 947–12 954.

[5] X. Yu, Y. Xie, Y. Liu, H. Lu, R. Xiong, Y. Liao, and Y. Wang, "Leverage cross-attention for end-to-end open-vocabulary panoptic reconstruction," *arXiv preprint arXiv:2501.01119*, 2025.

[6] 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 *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[8] Y. Siddiqui, L. Porzi, S. R. Buló, N. Müller, M. Nießner, A. Dai, and P. Kontschieder, "Panoptic lifting for 3d scene understanding with neural fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9043–9052.

[9] S. Kobayashi, E. Matsumoto, and V. Sitzmann, "Decomposing nerf for editing via feature field distillation," *Advances in neural information processing systems*, vol. 35, pp. 23 311–23 330, 2022.

[10] J.-C. Shi, M. Wang, H.-B. Duan, and S.-H. Guan, "Language embedded 3d gaussians for open-vocabulary scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5333–5343.

[11] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "Langsplat: 3d language gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 051–20 060.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 676–21 685.

[13] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *European Conference on Computer Vision*. Springer, 2024, pp. 162–179.

[14] 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.

[15] Y. Wang, X. Wei, M. Lu, and G. Kang, "Plgs: Robust panoptic lifting with 3d gaussian splatting," *arXiv preprint arXiv:2410.17505*, 2024.

[16] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[17] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[18] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[19] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 838–15 847.

[20] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 729–19 739.

[21] Y. Bhalgat, I. Laina, J. F. Henriques, A. Zisserman, and A. Vedaldi, "Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion," *arXiv preprint arXiv:2306.04633*, 2023.

[22] 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. 139–1, 2023.

[23] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2d gaussian splatting for geometrically accurate radiance fields," in *ACM SIGGRAPH 2024 conference papers*, 2024, pp. 1–11.

[24] C. Jiang, R. Gao, K. Shao, Y. Wang, R. Xiong, and Y. Zhang, "Li-gs: Gaussian splatting with lidar incorporated for accurate large-scale reconstruction," *IEEE Robotics and Automation Letters*, 2024.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[26] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.

[27] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[28] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.