IQGS: Instance Query-based Gaussian Segmentation

Yichao Gao[1, 2], **Xinyuan Liu** [1, 2], **Yike Ma**[1], **Yucheng Zhang**[1], **Feng Dai**[1*]

[1]Institute of Computing Technology, Chinese Academy of Science
[2]University of Chinese Academy of Sciences
{gaoyichao23s, liuxinyuan21s, ykma, zhangyucheng, fdai}@ict.ac.cn

Abstract

In recent years, Gaussian scene representations have achieved a series of promising results in 3D reconstruction. Compared to the previous 3DGS paradigm, the latest reconstruction approach 2DGS can achieve more accurate geometric representation using fewer Gaussian points. Accordingly, developing a panoramic segmentation algorithm suitable for 2DGS-reconstructed scenes is of significant importance. However, existing segmentation methods are primarily designed for 3DGS. They either fail to account for all objects in complex segmentation scenes or suffer from significant performance degradation when applied to 2D Gaussian scenes. Moreover, these methods consistently exhibit poor cross-dataset generalization. To address these issues, we propose IQGS, a segmentation framework applicable to 2DGS representations. Specifically, IQGS employs per-instance query and relaxed object-level supervision instead of strict pixel-level ID supervision, effectively mitigating the segmentation performance degradation that occurs when applied to 2DGS. At the same time, by learning features independent of specific object ID assignments, IQGS enhances its ability to generalize across diverse datasets. Our method achieves impressive panoramic segmentation results across multiple datasets, with an average mIoU of 66.6%, surpassing the state-of-the-art method Gaussian Grouping, which achieves 57.17%.

Code — https://github.com/tom-gao-gyc/IQGS

Introduction

Recently proposed Gaussian Splatting methods, including 3DGS (Kerbl et al. 2023) and 2DGS (Huang et al. 2024), reconstruct scenes by fitting object surfaces with 3D Gaussian spheres and 2D Gaussian splats. The newer reconstruction paradigm, 2DGS, can reduce surface noise in the 3DGS approach and thus achieve more accurate geometric representations with fewer Gaussian points. Downstream tasks after reconstruction—such as scene editing, object interaction, and semantic understanding—require disentangling individual objects within the reconstructed scene. Therefore, it is crucial to develop segmentation algorithms that are applicable to 2DGS-reconstructed scenes.

*Corresponding Author

To distinguish which instance each Gaussian point belongs to, an additional learnable instance attribute is embedded into each Gaussian point, as shown in Figure 1(a). However, given the general scarcity of 3D annotations for objects in reconstructed scenes, it is necessary to leverage rendered 2D images and masks from multiple viewpoints for 3D scene segmentation. Existing methods for performing panoptic segmentation on Gaussian-reconstructed scenes typically lift per-view 2D masks into the 3D scene space. These approaches can be broadly categorized into two paradigms: clustering-based methods and classification-based methods.

Early clustering-based methods optimize rendered multi-view instance feature maps using contrastive learning, as shown in Figure 1(b). This learning paradigm encourages intra-mask similarity and inter-mask separation based on masks generated by SAM (Kirillov et al. 2023), and retrieves instances via feature clustering. However, these methods often underperform in panoptic segmentation, as it is challenging to find a single clustering threshold that can simultaneously accommodate all instances—largely due to the significant variability in their discriminability across different categories. Additionally, these methods only learn instance attributes embedded in the Gaussian points, without acquiring any transferable modules that can represent object features across different scenes. As a result, they lack the ability to generalize to other scenes.

Another line of work adopts a classification-based Gaussian segmentation paradigm, such as Gaussian Grouping (Ye et al. 2024). As shown in Figure 1(c), it first leverages the external model DEVA (Cheng et al. 2023) to reassign a globally unique ID to each object across SAM-generated multi-view masks. Subsequently, per-view rendered instance attributes are classified into ID maps and supervised via a per-pixel classification loss. A key advantage of this approach lies in providing more explicit optimization signals: DEVA enforces cross-view consistent object IDs, directly enabling precise supervision. This clarity facilitates learning discriminative and stable instance attributes within Gaussian points.

However, such classification-based Gaussian segmentation models also have some limitations. **1) When applied to 2DGS, the segmentation accuracy degrades dramatically.** As shown in Figure 1(d), Gaussian Grouping achieves 60.99% mIoU on the *counter* scene reconstructed via 3DGS,

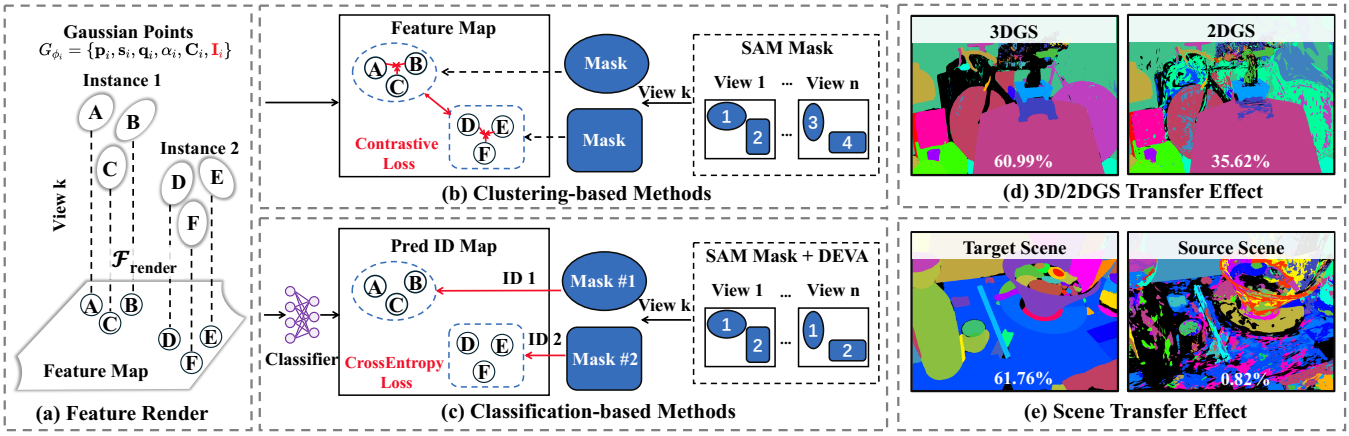


Figure 1: Overview of existing methods. (a) To identify the instance of each Gaussian point, an additional learnable instance attribute is embedded into each point and rendered to each view. Existing Gaussian segmentation methods fall into two categories: (b) clustering-based methods use GT masks to cluster features on the rendered feature map; (c) classification-based methods introduce a classifier to explicitly predict each pixel's instance ID, requiring the external model DEVA for global instance IDs. However, even SOTA methods have notable limitations, as illustrated by Gaussian-Group on *bonsai* dataset: (d) severe accuracy degradation (60.99% $\rightarrow$ 35.62% on mIoU) when transferring from 3DGS to 2DGS; (f) significant performance drop (61.76% $\rightarrow$ 0.82% on mIoU) when transferring from source to target scenes.

but drops sharply to 35.62% under 2DGS. This degradation stems from view-dependent depth ordering in 2DGS, which easily causes inconsistent rendered features for the same object across views (See Appendix.A.1 for detailed analysis). While such variations are expected and even desired for photorealistic reconstruction due to lighting or occlusion, they impair segmentation performance. This is particularly problematic for classification-based methods, as excessive feature discrepancies of the same object across multiple views make it difficult to predict consistent IDs (See Appendix.A.2 for detailed analysis). **2) They have weak generalization ability across different scenes.** As shown in Figure 1(e), the original model trained on the *ramen* scene achieves 61.76% mIoU, but it drops sharply to 0.82% when the classifier is replaced by one trained on *room* scene, revealing poor cross-scene generalization. This is because the instance IDs obtained via classification are only valid within the original scene, meaning IDs from new scenes are completely unfamiliar to the model. To put it simply, ID#1 in a new scene is not the same object as ID#1 in the original scene whatsoever.

To address the aforementioned limitations, we propose IQGS, a instance query-based framework for Gaussian scene segmentation tailored to 2DGS. In IQGS, a set of learnable queries is introduced to model scene objects. These queries decode objects in both the original 3D space and the rendered 2D space, enabling efficient learning through supervision from 2D ground-truth masks, complemented by 2D-3D consistency constraints and 3D regularization. Despite its simple architecture, two key designs enable 2DGS to achieve superior segmentation performance and generalization capability: **(1) ID-agnostic supervision signals:** The object decoded by each query only needs to align with the ground-truth masks of that object in the corresponding views, without the need to predict object ID number. Obviously, this relaxed, view-specific prediction objective is more compatible with the rendered features under unstable depth ordering in 2DGS, and thus improves segmentation performance. Moreover, the learned features are decoupled from object IDs themselves, and this in turn enhances generalizability. **(2) ID-aware sample matching:** Unlike conventional methods that use Hungarian matching to associate queries with ground-truth objects, here queries are directly linked to objects by global ID, i.e., the i-th query specializes in predicting the object with ID#i. This design fully leverages the cross-view consistency priors introduced by DEVA, preventing fluctuations in the optimization target of individual queries during training and allowing them to focus on learning object-specific features. Incidentally, when prioritizing generalization, using Hungarian matching remains a viable option, as it allows queries to learn more general object features without being confined to a single object. Furthermore, Hungarian matching enables implicit ID unification, thereby eliminating the need for the external DEVA model. Our contributions are summarized as follows:

- We are the first to identify that SOTA 3DGS scene segmentation methods suffer severe performance degradation when transferred to 2DGS, while these methods also exhibit poor cross-scene generalization.
- We propose an instance query-based scene segmentation framework tailored for 2DGS, which significantly alleviates the issues via two key designs: ID-agnostic supervision signals and ID-aware sample matching.
- Extensive experiments on multiple classic datasets validate our method: it not only outperforms SOTA 2DGS-adapted versions but also surpasses original 3DGS methods under the 2DGS setting. Moreover, cross-scene generalization experiments show that our model effectively learns object-level features across diverse environments.

Related Work

Nerf-based Methods Neural Radiance Fields (NeRF) reconstructs 3D scenes by learning a continuous volumetric representation from multi-view images. Numerous segmentation methods have been proposed based on NeRF reconstructions (Wang, Chen, and Yang 2022; Chen et al. 2022; Fan et al. 2022; Niemeyer et al. 2022; Fu et al. 2022; Goel et al. 2023; Chen et al. 2023a; Fang et al. 2022; Li et al. 2023c; MIRZAEI et al. 2024; Lin et al. 2023a; Li et al. 2023b; Hao et al. 2023; Wu et al. 2022; Azizi et al. 2023; Chen et al. 2025a,b). However, NeRF suffers from several inherent limitations which have made it increasingly obsolete compared to more recent Gaussian-based reconstruction approaches. Therefore, we focus our subsequent discussion on segmentation methods based on Gaussian reconstruction.

Feature Distilling Methods They transfer high-dimensional features from 2D vision foundation models (Caron et al. 2021; Radford et al. 2021; Li et al. 2023a; Wang et al. 2022; Li et al. 2022) into 3D representations (Oquab et al. 2023; Fan et al. 2023; Xu et al. 2023; Chen et al. 2023b; Lin et al. 2023b; Radford et al. 2021; Zhou et al. 2023; Liu et al. 2023; Lu et al. 2023; Cho et al. 2023). However, since these semantic features are not specifically designed for segmentation tasks, they are not well-suited for fine-grained instance-level 3D understanding. Therefore, we focus our subsequent discussion on the 2D mask-lifting paradigm.

Contrastive Learning-Based Methods Several recent works explore segmentation using clustering-based contrastive learning. SA3D (Cen et al. 2023) utilizes user prompts for single-object segmentation via contrastive supervision. OmniSeg3D (Ying et al. 2024) constructs a feature field from 2D masks with hierarchical contrastive learning and manually tuned similarity thresholds. SAGA (Cen et al. 2025) aligns SAM-generated masks across views using per-Gaussian contrastive loss. LangSplat (Qin et al. 2024) clusters Gaussians in SAM mask space as pseudo-instances. Click-Gaussian (Choi et al. 2024) enhances this with hierarchical contrastive supervision and global feature aggregation. However, all these methods depend on similarity-based clustering, which suffers from the lack of globally optimal thresholds across diverse object types, limiting segmentation completeness and accuracy.

Classifier-based Methods Another approaches perform per-pixel classification over rendered Gaussians. For example, Gaussian Grouping introduces learnable classifiers for each object and applies classification loss directly on the rendered outputs. However, its performance degrades significantly under 2D Gaussian Splatting (2DGS), mainly due to depth ordering ambiguities. Similarly, OMEGAS (Wang et al. 2025) follows a classification-based strategy and additionally enforces cross-view feature consistency using cosine similarity losses, but still struggles with segmentation quality in the 2DGS setting.

Method

Preliminaries: 2D Gaussian Splatting

Scene Representation To represent the scene, each 2D Gaussian's property is characterized by a centroid $\mathbf{p} = \{x, y, z\} \in \mathbb{R}^3$, a 2D size $\mathbf{s} \in \mathbb{R}^2$ in standard deviations, and a rotational quaternion $\mathbf{q} \in \mathbb{R}^4$. To enable efficient α-blending during rendering, the opacity value $\alpha \in \mathbb{R}$ and the color vector $\mathbf{c}$ are parameterized using spherical harmonics (SH) up to degree 3. These adjustable attributes are collectively symbolized by G_{Φ_i}, where $G_{\Phi_i} = \{\mathbf{p}_i, \mathbf{s}_i, \mathbf{q}_i, \alpha_i, \mathbf{c}_i\}$ represents the set of attributes for the i-th 2D Gaussian.

Rendering Process The 2D per-view feature map is computed by rendering per-Gaussian attributes (e.g., SH-encoded color) to the image plane using the differentiable rasterization in 2DGS. At each pixel, attributes are aggregated via kernel-weighted projection:

$$\widetilde{\mathbf{V}}_{2D}^{(k)}(x,y) = \sum_{i=1}^{N} w_i^{(k)}(x,y) \cdot \mathbf{a}^{(i)}, \quad \widetilde{\mathbf{V}}_{2D}^{(k)} \in \mathbb{R}^{D_{\text{feat}} \times H \times W} \tag{1}$$

Here, $\mathbf{a}^{(i)}$ is the attributes of the i-th Gaussian, and $w_i^{(k)}(x,y)$ is its contribution to pixel (x,y) under view k. This produces view-consistent feature maps that reflect the scene's appearance from each camera perspective.

IQGS Architecture

We retain all inherent attributes of the 2D Gaussian primitives. At the same time, to enable semantic reasoning and instance segmentation, we augment each Gaussian with an additional learnable attribute called instance. The architecture and loss of IQGS is illustrated in Figure 2.

Feature Extraction Our model supports both 3D point cloud segmentation and 2D per-view instance segmentation, requiring features from both domains.

For the 3D case, each Gaussian's instance attribute $\mathbf{I}^{(i)} \in \mathbb{R}^{D_{\text{obj}}}$ is projected into a lower-dimensional feature space using a learnable linear projection, yielding the 3D feature tensor:

$$\mathbf{F}_{3D}(i) = \mathcal{P}(\mathbf{I}^{(i)}), \mathbf{F}_{3D} \in \mathbb{R}^{N \times D} \tag{2}$$

For 2D per-view features, the rendered feature map from view v is defined as:

$$\mathbf{F}_{2D}^{(k)}(x,y) = \mathcal{P}\left(\texttt{Render}_k\left(\{\mathbf{I}^{(i)}\}_{i=1}^{N}\right)\right), \quad \mathbf{F}_{2D}^{(k)} \in \mathbb{R}^{D \times H \times W} \tag{3}$$

Here, the function $\texttt{Render}_k(\cdot)$ renders these embeddings into a 2D feature map under view k.

Query Design and Decoding We introduce a learnable query set:

$$\mathbf{Q}' \in \mathbb{R}^{N_q \times D} \tag{4}$$

where N_q is the number of queries, and D is the query embedding dimension, which is also equal to the dimension of the final feature representation.

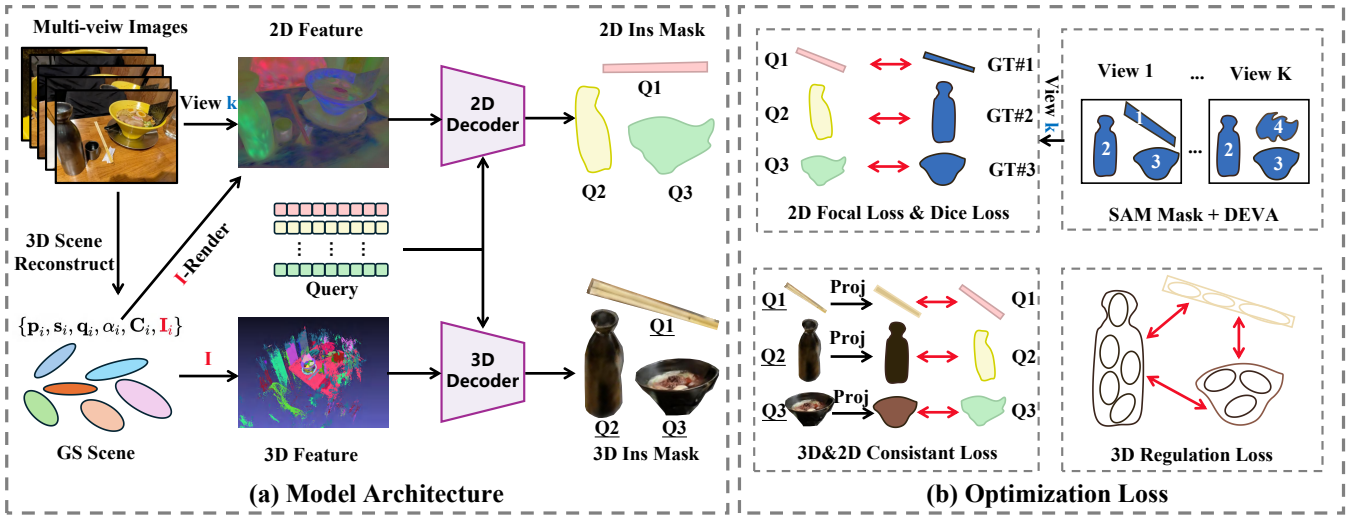


Figure 2: Overview of the proposed IQGS framework. (a) IQGS introduces learnable queries to model scene objects, which decode objects in both the original 3D space and rendered 2D space. (b) This enables efficient learning via supervision from 2D ground-truth masks, supplemented by 2D-3D consistency constraints and 3D regularization. Note that the instance ID is only used for to match queries with GT masks, without direct involvement in loss calculation. This ultimately enables the model to avoid the accuracy degradation and weak generalization under 2DGS.

Before decoding, we enrich the queries via a combination of cross-attention (Vaswani et al. 2017) and self-attention over the 2D feature maps:

$$\mathbf{Q} = \text{SelfAttn}(\text{CrossAttn}(\mathbf{Q}', \mathbf{F}_{2D})) \tag{5}$$

where each query $\mathbf{Q}_i \in \mathbb{R}^D$ targets a potential object instance.

For 3D instance segmentation over the Gaussian set:

$$\begin{aligned} \mathbf{S}_{3D}(i,j) &= \mathbf{Q}_i \cdot \mathbf{F}_{3\mathrm{D}j}, \\ \mathbf{S}_{3D} &\in \mathbb{R}^{B \times N_q \times N_{\text{pts}}} \end{aligned} \tag{6}$$

where $\mathbf{S}_{3D}(i,j)$ represents the predicted assignment score between query i and Gaussian j.

For 2D instance segmentation over the Gaussian set, we compute the per-query segmentation activation map as:

$$\begin{aligned} \mathbf{S}_{2D}^{(k)}(i,x,y) &= \text{reshape}\left(\mathbf{Q}_i \cdot \mathbf{F}_{2\mathrm{D}}^{(k)}\right), \\ \mathbf{S}_{2D}^{(k)} &\in \mathbb{R}^{N_q \times H \times W} \end{aligned} \tag{7}$$

Here, the dot product produces a 1D activation map of length $H \times W$, which is reshaped into the spatial map $\mathbf{S}_{2D}(i,x,y)$ for instance i at view k.

Matching Strategy We introduce two alternative matching strategies, respectively tailored for achieving higher panoptic segmentation accuracy and better cross-dataset generalization.

ID Matching We directly match each query to a segmentation mask with the same ID assigned by DEVA. Formally, given a set of queries $\{q_i\}$ and DEVA-generated masks with corresponding IDs $\{M_j\}$, we establish a one-to-one alignment by :

$$\text{indices} = \{(i,j) \mid \text{ID}(q_i) = \text{ID}(M_j)\} \tag{8}$$

Hungarian Matching For each view, we match predicted query masks to SAM-provided ground-truth masks by computing a pairwise cost:

$$\begin{aligned} \mathbf{C}(i,j) = &\lambda_{\text{focal}} \cdot \mathcal{L}_{\text{focal}}(\mathbf{S}_{2D}(i), \mathbf{M}_j) \\ &+ \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}}(\mathbf{S}_{2D}(i), \mathbf{M}_j) \end{aligned} \tag{9}$$

$$\text{indices} = \arg\min_{\pi} \sum_{(i,j)\in\pi} \mathbf{C}(i,j) \tag{10}$$

where $\mathbf{S}_{2D}(i) \in \{0,1\}^{H \times W}$ is the i-th predicted binary mask and $\mathbf{M}_j$ is the j-th SAM-generated ground-truth mask. $\mathbf{C}(i,j)$ denotes their matching cost, and the Hungarian algorithm is applied per view to find the optimal assignment π minimizing the total cost.

Loss Function

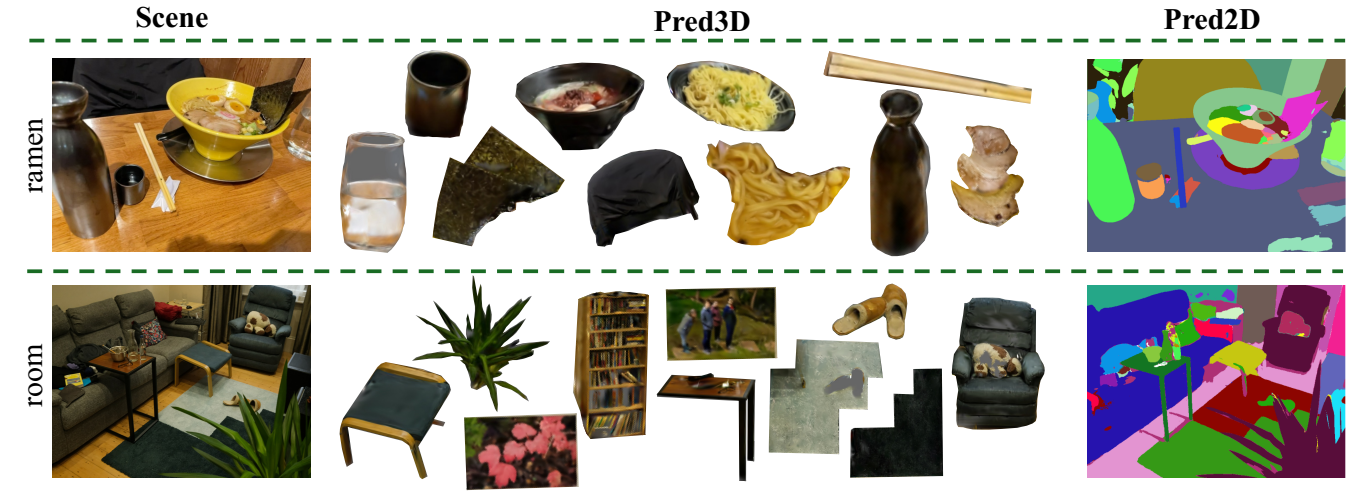
2D Loss To supervise the predicted segmentation masks at the 2D level, we apply several per-view loss functions (Lin et al. 2017; Milletari, Navab, and Ahmadi 2016) that guide both mask accuracy and feature discriminability throughout training.

Focal Loss :

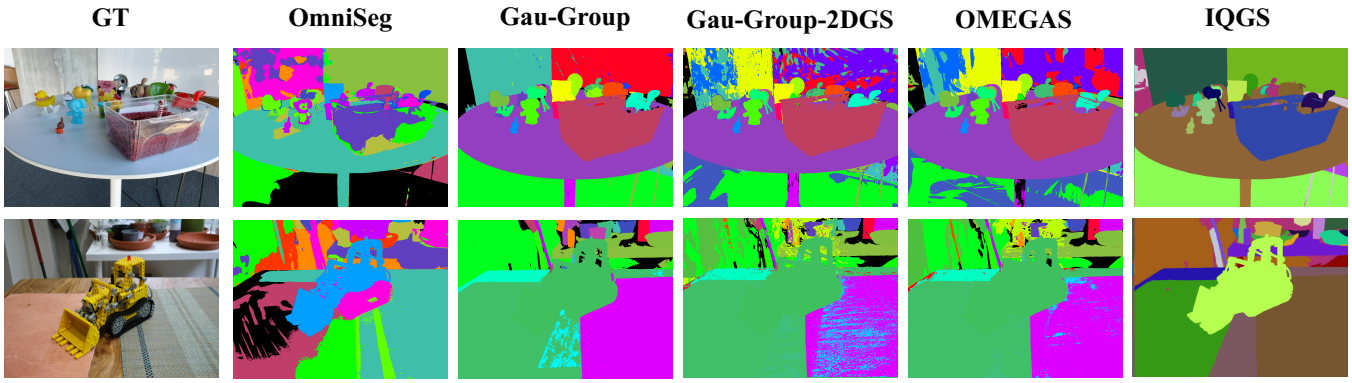
$$\mathcal{L}_{\text{mask}} = \frac{1}{N_{\text{matched}}} \sum_{(i,j)\in\mathcal{M}} \mathcal{L}_{\text{focal}}(\mathbf{S}_{2D}^{(i)}, \mathbf{M}_j) \tag{11}$$

Dice Loss :

$$\mathcal{L}_{\text{dice}} = \frac{1}{N_{\text{matched}}} \sum_{(i,j)\in\mathcal{M}} \mathcal{L}_{\text{dice}}(\mathbf{S}_{2D}^{(i)}, \mathbf{M}_j) \tag{12}$$



(a)



(b)

Figure 3: Visualization of model results (see Appendix.B for more results). (a) provides both 3D and 2D results of IQGS. (b) compares our results with other methods.

3D Loss To enhance global 3D consistency and structure in the reconstructed Gaussian scene, we apply the following two 3D-level losses:

3D Feature Regularization Loss :

$$\mathcal{L}_{\text{reg3D}} = \frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} \left\| \mathbf{f}_{\texttt{mask}}^{(i)} - \mathbf{f}_{\texttt{mask}}^{(j)} \right\|_2^2 \tag{13}$$

where $\mathcal{N}$ is a set of neighboring Gaussian pairs, encouraging semantic smoothness and avoiding over-fragmentation in 3D space.

2D and 3D Consistency Loss :

$$\mathcal{L}_{\text{consist}} = \sum_{k} \sum_{m=1}^{N_q} \mathcal{L}_{\text{overlap}} \left(\Pi_v \left(\mathbf{S}_{3D}(m) \right), \mathbf{S}_{2D}^{(k)}(m) \right) \tag{14}$$

where $\Pi_v(\cdot)$ denotes the projection of 3D instance predictions into view k. This loss aligns the 2D segmentation output with 3D geometry.

Total Loss The final training objective combines all 2D and 3D loss terms with weighting coefficients:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}} + \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}} \\ + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg3D}} + \lambda_{\text{geom}} \cdot \mathcal{L}_{\text{consist}} \end{aligned} \tag{15}$$

where each λ is a hyperparameter controlling the contribution of its corresponding loss.

Experiments

Datasets

We use the same datasets as prior works on Gaussian scene segmentation, including, 360_v2, and LERF reconstruction datasets (Barron et al. 2021). The two datasets consist of seven indoor scenes. All scenes are real-world rather than synthetic, and they feature a wide variety of objects within complex and extensive environments. As such, evaluating models on these datasets effectively demonstrates their robustness and generalization capabilities. Each scene contains multi-view images captured around the environment,

Paradigm	Method	LERF Datasets (%)			360_v2 Datasets (%)			
		ramen	figurines	teatime	bonsai	counter	kitchen	room
3DGS-based	SA3D (NeurIPS 2023)	33.01	27.89	25.03	33.98	28.19	34.09	25.15
	Omniseg3D (CVPR2024)	35.67	29.60	31.59	34.70	30.03	35.14	26.68
	SAGA (AAAI 2025)	43.81	35.47	37.80	44.98	40.25	46.03	49.27
	Gau-Group (ECCV 2024)	<u>61.76</u>	<u>42.81</u>	**53.21**	<u>60.33</u>	<u>60.99</u>	<u>59.79</u>	<u>61.32</u>
2DGS-based	Gau-Group (ECCV 2024)	40.97	31.22	33.65	40.30	35.62	37.42	35.23
	OMEGAS (arXiv 2024)	48.63	40.83	41.51	46.66	46.57	39.45	42.34
	IQGS (Ours)	**63.73**	**61.76**	<u>52.60</u>	**76.50**	**74.96**	**69.71**	**66.91**

Table 1: Per-scene mIoU results on LERF and 360_v2 datasets.

together with SAM-generated segmentation masks, camera intrinsics and extrinsics estimated via SfM algorithms, and an initial sparse point cloud reconstruction.

Implementation Details

All experiments are conducted using PyTorch 2.0.0 on an NVIDIA A100 GPU. The rendering process is based on the Gaussian Render framework. When applying the Segment Anything Model (SAM), we modify its invocation parameters to better suit the Gaussian segmentation task, and apply post-processing steps including mask filtering and merging. The batch size is set to 1 for all experiments. To ensure a fair comparison, training configurations generally follow those used in the baseline Gaussian Grouping and other related works. We use the Adam optimizer with a learning rate of 5×10^{-4} and $\epsilon = 10^{-15}$ for experiments. All models are trained for 30,000 iterations.

Metrics

Due to the absence of 3D ground-truth annotations and standard evaluation metrics for 3D Gaussian segmentation, we evaluate segmentation performance based on the 2D mean Intersection-over-Union (mIoU) across multiple views. Specifically, for each scene, we compute the IoU between each predicted instance in a view and its matched ground-truth instance. Notably, to more accurately reflect the model's ability to predict object instances, we do not limit the evaluation to matched pairs with valid correspondences in both prediction and ground truth. Instead, if a predicted instance has no corresponding ground-truth match, we treat its IoU as zero. The final mIoU is obtained by averaging the IoUs over all predicted instances across all views for a given scene.

Panoptic Instance Segmentation

Task Setting Given original training data of a scene, the goal is to perform panoptic segmentation over the reconstructed scene. We first conduct experiments in 3DGS reconstruction scenes using both contrastive learning-based and classification-based approaches. Subsequently, we adapt Gaussian Grouping algorithms and evaluate OMEGAS in the 2DGS reconstruction setting. For our method, we ensure that the same object observed from different views is consistently assigned the same query ID during decoding, and that the IDs in the 3D segmentation results are aligned with those in the 2D masks of all views.

Results and Analysis Quantitative results of baselines and IQGS are shown in Table 1. Visualization results of 2D and 3D segmentation, along with comparisons to other methods, are shown in Figure 3. Among the methods applicable to 3DGS, contrastive clustering-based approaches such as SA3D-GS (29.62%), Omniseg3D-GS (31.92%), and SAGA (42.52%) demonstrate relatively low performance in panoptic segmentation.. In contrast, the classification-based method Gaussian Grouping achieves significantly better results (57.17%). However, when directly transferring the classification strategy of Gaussian Grouping to the 2DGS setting, a sharp drop in segmentation accuracy is observed (36.34%). In contrast, our proposed IQGS (66.6%) demonstrate superior performance on 2DGS compared to the aforementioned classification-based method. IQGS also achieves state-of-the-art performance across most datasets.

Generalization Evaluation

Task Setting To evaluate the generalization ability of those models, we conduct a transfer-based experiment across different scenes. Specifically, we select scene pairs and transfer a pretrained model trained on one source scene directly to its paired counterpart target scene without any fine-tuning. We then assess how well the learned segmentation model generalizes to this new scene. Since contrastive clustering methods do not provide transferable parameters for inference on novel scenes, we instead compare with the classification-based approach Gaussian Grouping.

Results and Analysis The quantitative comparison of mIoU is shown in Table 2, while the qualitative results are visualized in Figure 1(e) and Figure 4. Most object IDs are incorrectly segmented in the Gaussian Grouping results, and the predicted masks often exhibit fragmented and noisy regions within individual objects. This indicates that Gaussian Grouping suffers from almost no generalization (1.56%) across different datasets. Compared to Gaussian Grouping, IQGS demonstrates initial generalization capability (11.26%), segmenting more objects correctly and achieving higher accuracy. In particular, we construct a variant, *IQGS-H*, by replacing the ID matching module in IQGS with a Hungarian matching mechanism. IQGS-H achieves stronger cross-dataset generalization (34.18%) by avoiding

Method	teatime	figurines	bonsai	room	kitchen	counter
Gau-Group	1.56	1.70	1.10	0.82	1.55	2.65
IQGS	10.11	11.72	9.86	12.78	10.56	12.56
IQGS-H	31.09	41.78	39.92	33.67	28.45	30.17

Table 2: Generalization performance (%) of *Gaussian Grouping*, *IQGS* and *IQGS-H* (replaced ID-based matching in IQGS with Hungarian matching) when trained on different source datasets and directly evaluated on the target scene *ramen* without fine-tuning.

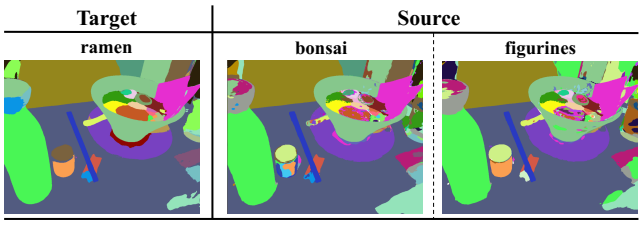


Figure 4: IQGS-H shows pretty strong generalization capability.

fixed ID supervision. Hungarian matching enables queries to learn transferable object representations rather than memorizing scene-specific IDs, thereby enhancing their adaptability to unseen scenes.

Ablation Study

Effect of Query Number and Feature Dimension. We ablate two key hyperparameters—object query number N_q and feature dimension D. Increasing N_q improves performance by better covering complex scenes. A smaller D, such as 16, already captures instance identity well, balancing accuracy and efficiency.

N_q	D	LERF (%)	360_v2 (%)
50	16	46.21	57.13
150	16	50.34	63.46
300	16	**59.36**	**72.02**
300	8	52.68	64.62
300	32	53.57	66.59

Table 3: Ablation of query count N_q and feature dimension D with mIoU results on LERF and 360_v2 datasets.

Effect of different loss components and weights. We ablate the contribution of each loss term by removing them individually and adjusting their weights. As shown in Table 4, removing either focal loss or dice loss significantly reduces mIoU, indicating their critical roles in mask quality. Moreover, we observe that tuning the weights of λ_{mask} and λ_{dice} jointly can slightly improve performance, but overly decreasing them harms accuracy.

Effect of model composition. We perform ablation studies to assess the contribution of key components in our framework. Replacing the original DEVA+ID matching with Hungarian matching leads to performance degradation, confirming the effectiveness of query-ID to mask-ID alignment over conventional Hungarian assignment.

Loss Setup	LERF	360_v2
(a) Loss Component Removal		
All losses (default weights)	**59.36**	**72.02**
w/o $\mathcal{L}_{\text{focal}}$	33.91	48.69
w/o $\mathcal{L}_{\text{dice}}$	49.43	59.34
w/o $\mathcal{L}_{\text{reg3D}}$	51.75	60.08
w/o $\mathcal{L}_{\text{consist}}$	54.33	66.17
(b) Weight Configuration Variants		
$\lambda_{\text{mask}} = 10.0$ ($\downarrow$ from 20.0)	51.15	63.31
$\lambda_{\text{dice}} = 2.0$ ($\uparrow$ from 1.0)	56.52	69.95
$\lambda_{\text{reg}} = 0.5$ ($\downarrow$ from 1.0)	56.85	70.34
$\lambda_{\text{geom}} = 0.5$ ($\downarrow$ from 1.0)	58.44	71.49

Table 4: Ablation study on loss settings with $N_q = 300$, $D = 16$. Results are reported as mIoU(%) on LERF and 360_v2 datasets.

We also modify the query branch to predict per-pixel IDs, following classification-based paradigms, and supervise using ID maps instead of binary masks. This yields only marginal gains over direct application of Gaussian Grouping on 2DGS, likely due to the presence of self- and cross-attention in our query design. These results indicate that classification-style methods underperform on 2DGS due to overly strict supervision, while our design alleviates this issue and achieves improved segmentation. Quantitative results are reported in Table 5.

Module Setup	LERF (%)	360_v2 (%)
(a) Matching Method		
ID Matching	**59.36**	**72.02**
Hungarian Matching	49.36	58.21
(b) Supervision Method		
Binary-Mask Supervision	**59.36**	**72.02**
ID maps Supervision	40.02	49.27

Table 5: Ablation study on model composition. Results are reported as mIoU on LERF and 360_v2 datasets.

Conclusion

In this work, we identify the limitations of existing 3DGS-based segmentation methods, including a performance degradation when applied to 2DGS and poor generalization. To address this, we propose IQGS. Extensive experiments validate that IQGS achieves state-of-the-art results across multiple complex real-world datasets, both in segmentation and generalization tasks. We hope this work inspires further research into query-based representation learning for view-consistent scene understanding.

Acknowledgments

This work is supported by Strategic Priority Research Program of Chinese Academy of Sciences (XDA28040000), National Natural Science Foundation of China (62372433), the Chinese Academy of Sciences and the science and technology projects of the Ministry of Agriculture and Rural Affairs of China.

References

Azizi, S.; et al. 2023. Nerfies: Deformable neural radiance fields. *arXiv preprint arXiv:2304.06162*.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5855–5864.

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

Cen, J.; Fang, J.; Yang, C.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2025. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1971–1979.

Cen, J.; Zhou, Z.; Fang, J.; Shen, W.; Xie, L.; Jiang, D.; Zhang, X.; Tian, Q.; et al. 2023. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36: 25971–25990.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, 333–350. Springer.

Chen, K.; Yuan, Z.; Mao, T.; and Wang, Z. 2025a. Dual-level precision edges guided multi-view stereo with accurate planarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2105–2113.

Chen, K.; Yuan, Z.; Xiao, H.; Mao, T.; and Wang, Z. 2025b. Learning multi-view stereo with geometry-aware prior. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chen, X.; Tang, J.; Wan, D.; Wang, J.; and Zeng, G. 2023a. Interactive segment anything nerf with feature imitation. *arXiv preprint arXiv:2305.16233*.

Chen, X.; et al. 2023b. ISRF: Image-Conditioned Semantic Radiance Fields for Open-World 3D Understanding. *arXiv preprint arXiv:2310.01884*.

Cheng, H. K.; Oh, S. W.; Price, B.; Schwing, A.; and Lee, J.-Y. 2023. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1316–1326.

Cho, S.; et al. 2023. Gaussmrf: Gaussian-based semantic mapping with Markov random fields. *arXiv preprint arXiv:2306.06613*.

Choi, S.; Song, H.; Kim, J.; Kim, T.; and Do, H. 2024. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, 289–305. Springer.

Fan, H.; et al. 2023. DFFs: Distilled Feature Fields for 3D Recognition. *arXiv preprint arXiv:2304.09149*.

Fan, Z.; Wang, P.; Jiang, Y.; Gong, X.; Xu, D.; and Wang, Z. 2022. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*.

Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Fu, X.; Zhang, S.; Chen, T.; Lu, Y.; Zhu, L.; Zhou, X.; Geiger, A.; and Liao, Y. 2022. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, 1–11. IEEE.

Goel, R.; Sirikonda, D.; Saini, S.; and Narayanan, P. 2023. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4201–4211.

Hao, Z.; et al. 2023. Nerfbrowse: Interactive free-viewpoint exploration of large-scale 3d scenes. *arXiv preprint arXiv:2304.06291*.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, 1–11.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, J.; et al. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Li, J.; et al. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Li, X.; et al. 2023b. TinCan: Self-Supervised Multi-Modal Scene Understanding with Neural Radiance Fields. *arXiv preprint arXiv:2305.17951*.

Li, Z.; Müller, T.; Evans, A.; Taylor, R. H.; Unberath, M.; Liu, M.-Y.; and Lin, C.-H. 2023c. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8456–8465.

Lin, C.-H.; et al. 2023a. Nerfplayer: A streaming architecture for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21113–21123.

Lin, K.; et al. 2023b. LeRF: Language Embedded Radiance Fields. *arXiv preprint arXiv:2303.08733*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, H.; et al. 2023. Semanticrf: Semantic neural radiance fields for multi-view segmentation. *arXiv preprint arXiv:2303.07713*.

Lu, Y.; et al. 2023. Sparseactive: Unifying sparse supervision and active learning for 3d semantic segmentation. *arXiv preprint arXiv:2310.07296*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

MIRZAEI, A.; AUMENTADO-ARMSTRONG, T. T.; DERPANIS, K. G.; BRUBAKER, M. A.; Gilitschenski, I.; and Levinshtein, A. 2024. Multi-view segmentation and perceptual inpainting with neural radiance fields. US Patent App. 18/228,472.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5480–5490.

Oquab, M.; Darcet, T.; Moutakanni, T.; Neverova, N.; Ficheux, E.; Haziza, D.; Ponce, J.; Massa, F.; Caron, M.; and Gkioxari, G. 2023. DINOv2: Learning Robust Visual Features without Supervision. In *arXiv preprint arXiv:2304.07193*.

Qin, M.; Li, W.; Zhou, J.; Wang, H.; and Pfister, H. 2024. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20051–20060.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, B.; Chen, L.; and Yang, B. 2022. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*.

Wang, L.; Zhou, F.; Yu, B.; Cao, P.; and Yin, J. 2025. OMEGAS: Object Mesh Extraction from Large Scenes Guided by Gaussian Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, W.; et al. 2022. Git: A generative image-to-text transformer for vision and language. In *Advances in Neural Information Processing Systems*, 11038–11050.

Wu, C.; et al. 2022. Sparse neural representations for high-fidelity interactive rendering. In *SIGGRAPH Asia 2022 Conference Papers*, 1–11.

Xu, Q.; et al. 2023. N3F: Neural Feature Fusion Fields for Semantic Scene Completion. *arXiv preprint arXiv:2306.09953*.

Ye, M.; Danelljan, M.; Yu, F.; and Ke, L. 2024. Gaussian grouping: Segment and edit anything in 3d scenes. In *European conference on computer vision*, 162–179. Springer.

Ying, H.; Yin, Y.; Zhang, J.; Wang, F.; Yu, T.; Huang, R.; and Fang, L. 2024. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20612–20622.

Zhou, Y.; et al. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. *arXiv preprint arXiv:2306.00997*.