SegVGGT: Joint 3D Reconstruction and Instance Segmentation from Multi-View Images

Jinyuan Qu[1,3*], Hongyang Li[2,3*], and Lei Zhang[3†]

[1] Tsinghua University, China
[2] South China University of Technology, China
[3] International Digital Economy Academy (IDEA), China

Abstract. 3D instance segmentation methods typically rely on high-quality point clouds or posed RGB-D scans, requiring complex multi-stage processing pipelines, and are highly sensitive to reconstruction noise. While recent feed-forward transformers have revolutionized multi-view 3D reconstruction, they remain decoupled from high-level semantic understanding. In this work, we present SegVGGT, a unified end-to-end framework that simultaneously performs feed-forward 3D reconstruction and instance segmentation directly from multi-view RGB images. By introducing object queries that interact with multi-level geometric features, our method deeply integrates instance identification into the visual geometry grounded transformer. To address the severe attention dispersion problem caused by the massive number of global image tokens, we propose the Frame-level Attention Distribution Alignment (FADA) strategy. FADA explicitly guides object queries to attend to instance-relevant frames during training, providing structured supervision without extra inference overhead. Extensive experiments demonstrate that SegVGGT achieves the state-of-the-art performance on ScanNetv2 and ScanNet200, outperforming both recent joint models and RGB-D-based approaches, while showing favorable generalization to ScanNet++. The code is available at https://github.com/IDEA-Research/SegVGGT.

Keywords: 3D instance segmentation · Feed-forward 3D reconstruction · Visual geometry grounded transformer

1 Introduction

Understanding 3D scenes at the instance level is fundamental for embodied perception [11], robotics [65], AR/VR [19, 38], and autonomous navigation [57]. Among various scene understanding tasks, 3D instance segmentation plays a central role, as it provides object-level geometric representations that support reasoning and interaction. Current state-of-the-art 3D instance segmentation methods [15, 23, 25, 29, 31, 39, 46] predominantly rely on carefully reconstructed and post-processed point clouds or high-quality posed RGB-D scans as inputs,

* Equal contribution. Work done during an internship at IDEA Research.
† Corresponding author.

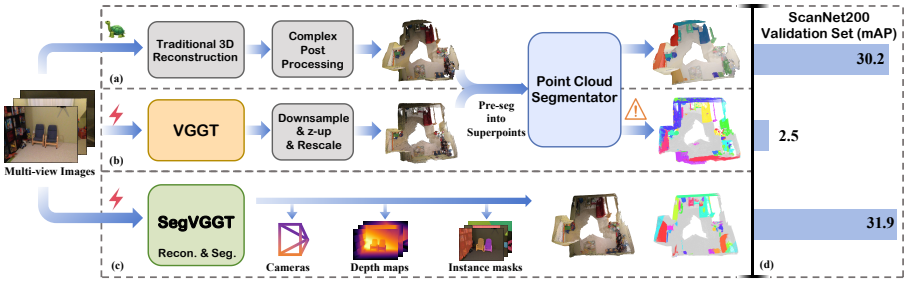

Fig. 1: Comparison of different 3D instance segmentation pipelines. (a) Segmentation methods that rely on high-quality point clouds require complex and time-consuming processing, limiting their applicability. (b) Feed-forward 3D reconstruction improves efficiency, but leads to severely degraded segmentation performance. (c) Our proposed SegVGGT performs 3D reconstruction and instance segmentation simultaneously in a single forward pass, achieving superior segmentation quality. (d) 3D instance segmentation performance on the ScanNet200 validation set.

as shown in Fig. 1 (a). However, their overall pipeline typically involves multiple stages, including multi-view data acquisition, camera registration, fusion and reconstruction, and point cloud denoising before segmentation can be performed. These fragmented workflows are usually time-consuming and considerably increase the system complexity, and their applicability is further limited when depth sensors are unavailable or reconstruction quality is imperfect.

Recent advances in 3D reconstruction have shifted the paradigm from traditional multi-stage pipelines [35, 44, 45] to one-stage feed-forward transformer-based models [20, 26, 52, 53, 55], significantly improving reconstruction efficiency. These approaches eliminate the need for iterative optimization and expensive post-processing, making large-scale 3D reconstruction practical. Nevertheless, they primarily focus on geometry estimation and remain disconnected from high-level semantic understanding. Consequently, 3D scene reconstruction and 3D semantic understanding are still treated as two relatively independent problems.

A straightforward strategy is to decompose the pipeline into two stages, where a feed-forward model first performs 3D reconstruction, followed by a point cloud-based segmentation model, as illustrated in Fig. 1 (b). However, such an approach normally suffers from two major limitations. First, segmentation performance is highly sensitive to reconstruction quality. Feed-forward reconstructions without heavy post-processing often contain noise and misalignment, causing point cloud segmentation methods to degrade significantly. Second, decoupling reconstruction and segmentation limits the potential of leveraging geometric reasoning to enhance instance-level understanding, leading to suboptimal performance.

Although several recent works [21, 28, 66] have explored joint modeling of 3D reconstruction and segmentation, existing approaches still exhibit some limitations. Some of them retain a two-stage paradigm, treating the geometric foundation model [3] as a frozen backbone and additionally introducing DINOv2 [37] and MaskFormer [6] for segmentation, which leads to suboptimal joint learning.

Others train an instance feature head solely with contrastive learning, which requires costly clustering during inference, cannot directly predict semantic categories, and often produces inaccurate segmentation boundaries.

In this work, we propose SegVGGT, a unified end-to-end framework which jointly performs feed-forward 3D reconstruction and 3D instance segmentation directly from multi-view RGB images, as shown in Fig. 1 (c). Our key insight is that instance understanding can be deeply integrated into the visual geometry grounded transformer. To this end, we initialize a set of object queries and let them perform cross-attention with global image tokens after global self-attention in each transformer layer, enabling instance-level representations to progressively evolve together with multi-level geometric features throughout the network. This tight coupling allows object queries to leverage both low-level spatial cues and high-level geometric abstractions. We further employ a semantic feature head [41] to produce cross-view consistent instance-level features, from which 3D instance masks are derived through dot products with the global object queries.

However, a typical scene may contain $10^4 \sim 10^5$ image tokens, and directly introducing object queries to interact with global image tokens without explicit guidance poses significant optimization challenges. We empirically observe that the attention between queries and tokens becomes highly dispersed in this setting. By contrast, transformer-based 3D instance segmentation models [25,39,46] commonly employ mechanisms such as mask-attention or positional embeddings to guide object queries to attend to specific regions. To address this issue within our unified 3D reconstruction and segmentation framework, we propose a Frame-level Attention Distribution Alignment (FADA) strategy. Specifically, we introduce an additional supervision and incorporate a corresponding cost into the Hungarian matching process, encouraging object queries to assign higher attention weights to frames that contain the associated instance while interacting with global image tokens. This mechanism provides a structured supervision for cross-frame attention without incurring additional inference overhead.

Extensive experiments validate the effectiveness of SegVGGT. On the ScanNetv2 [7] and ScanNet200 [43] benchmarks, SegVGGT achieves state-of-the-art segmentation performance compared with both joint 3D reconstruction and segmentation models and RGB-D scan-based methods, demonstrating its effectiveness. Moreover, without any additional training, we evaluate its performance on ScanNet++ [61]. Under the same setting, our method even surpasses approaches trained with ScanNet++, suggesting favorable cross-dataset generalization.

In summary, our contributions are threefold: 1) We present SegVGGT, a unified framework that simultaneously performs feed-forward 3D reconstruction and 3D instance segmentation directly from multi-view RGB images. 2) We integrate instance identification into the visual geometry grounded transformer through object queries and introduce the Frame-level Attention Distribution Alignment (FADA) strategy to efficiently solve the attention dispersion problem over massive image tokens without incurring inference overhead. 3) We achieve the state-of-the-art performance on ScanNetv2 and ScanNet200, surpassing re-

cent joint models and even methods relying on posed RGB-D scans. Moreover, SegVGGT shows favorable generalization to ScanNet++.

2 Related Works

3D Instance Segmentation on Point Clouds. 3D instance segmentation aims to assign both semantic labels and distinct instance identities to 3D geometry. Early works either adopt a proposal-based paradigm that first generates 3D bounding box proposals and subsequently predicts instance masks [8, 13, 22, 59, 62], or follow a grouping-based strategy that clusters points into instances based on learned similarity without explicit proposal generation [5, 17, 18, 30, 50, 54, 64]. Recently, transformer-based methods [1, 23, 25, 29, 31, 32, 36, 39, 46, 47, 51, 63] have become dominant by directly predicting instance masks with object queries. These methods primarily operate on high-quality point clouds, which often require complex post-processing. ODIN [15] instead takes posed RGB-D scans as input, yet still assumes access to reconstructed 3D geometry. Consequently, when only RGB images are available, these approaches are not directly applicable.

Feed-forward 3D Reconstruction. Traditional 3D reconstruction frameworks typically follow a multi-stage pipeline involving feature matching, Structure-from-Motion (SfM) [12, 44], and Multi-View Stereo (MVS) [45]. While being robust, these systems involve complex heuristic tuning and intensive computation, often requiring separate optimization phases to resolve geometric inconsistencies. Recent advances, such as DUSt3R [53] and its variants [16, 26], have shifted to a feed-forward paradigm that directly regresses dense point maps from image pairs. VGGT [52] and its successors [20, 55] further propose architectures capable of processing hundreds of frames in a single forward pass to predict multiple geometric attributes, substantially improving efficiency and scalability. Despite these advances, existing works focus primarily on geometric estimation and do not explicitly consider semantic or instance-level scene understanding.

Joint 3D Reconstruction and Segmentation. Recently, there has been growing interest in bridging reconstruction and understanding in a unified model. LSM [9] and Uni3R [48] extend geometric foundation models [52, 53] with 3D Gaussian Splatting heads to produce semantic features, but they lack instance-level awareness. SIU3R [58] further incorporates instance perception, but supports only two-view inputs. PanSt3R [66] follows a two-stage paradigm, using a frozen MUSt3R [3] model for scene reconstruction and treating it as a backbone combined with DINOv2 [37] to extract features, which are then fed into a separate Mask2Former [6] for panoptic segmentation, leading to suboptimal joint learning. IGGT [28] and UNITE [21] perform segmentation by training an instance-level feature head with contrastive learning. However, they require computationally expensive clustering [33] during inference, which may yield inaccurate segmentation boundaries, and rely on additional VLMs [10, 27, 40] to predict semantic categories. MVGGT [56] introduces the task of multi-view 3D referring expression segmentation along with a baseline model, aiming to reconstruct scenes from sparse views and segment the referred object.

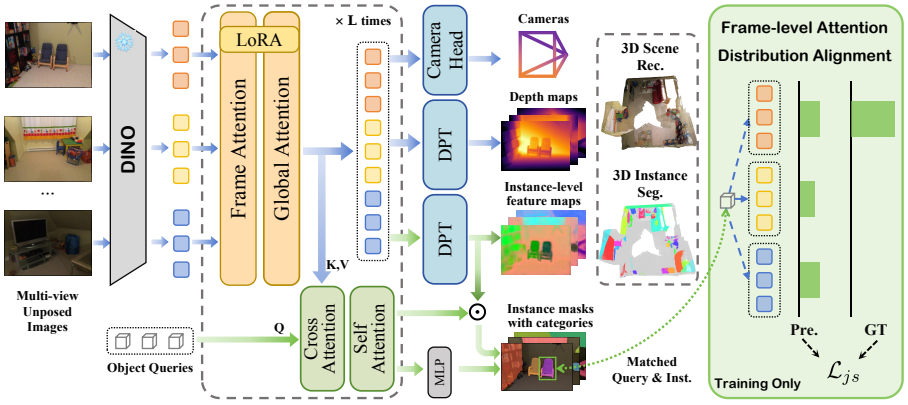

Fig. 2: Overall architecture of SegVGGT. SegVGGT takes multi-view unposed RGB images as input and outputs the camera parameters, depth maps, and instance masks in a single forward pass. In the visual geometry grounded transformer, image tokens and object queries are progressively refined to capture both multi-view geometry and instance semantics. The green arrows indicate the components related to the 3D instance segmentation. During training, the Frame-level Attention Distribution Alignment (FADA) module regularizes the cross-attention weights via Jensen-Shannon Divergence loss $\mathcal{L}_{js}$, effectively mitigating the attention dispersion problem.

3 Method

3.1 Overall Architecture

As illustrated in Fig. 2, SegVGGT takes N unposed RGB images $(I_i)_{i=1}^N$ as input, where $I_i \in \mathbb{R}^{3\times H\times W}$ and $H \times W$ denotes the resolution of images. Alongside the images, we also initialize a set of O learnable object queries $(\mathbf{q}_j)_{j=1}^O$, where $\mathbf{q}_j \in \mathbb{R}^d$ and d denotes the feature dimension. Our unified transformer maps the input images to the corresponding camera parameters, depth maps, and instance-level feature maps in a single forward pass, while simultaneously updating the object queries. This process can be formulated as:

$$f_\theta\Big((I_i)_{i=1}^N, (\mathbf{q}_j)_{j=1}^O\Big) = \Big((\mathbf{g}_i, D_i, F_i)_{i=1}^N, (\hat{\mathbf{q}}_j)_{j=1}^O\Big), \tag{1}$$

Among the model outputs, $\mathbf{g}_i \in \mathbb{R}^9$ denotes the camera intrinsics and extrinsics, and $D_i \in \mathbb{R}^{H\times W}$ is the depth map with the same resolution as the input image. $F_i \in \mathbb{R}^{d\times h\times w}$ represents the instance-level feature map, whose spatial resolution is half of the input image, i.e., $(h, w) = (\frac{H}{2}, \frac{W}{2})$, to reduce memory consumption. $(\hat{\mathbf{q}}_j)_{j=1}^O$ denotes the optimized object queries after aggregating multi-level geometric features in the transformer. From these outputs, SegVGGT can seamlessly derive instance segmentation masks with categories and obtain the final 3D reconstruction and 3D instance segmentation through simple unprojection.

3.2 Geometry Grounded Transformer with Instance Reasoning

To deeply integrate instance reasoning into the visual geometry grounded transformer, we extend this architecture by introducing a set of learnable object queries $(\mathbf{q}_j)_{j=1}^{O}$. These queries are progressively refined across all the L layers in the transformer, enabling them to leverage both low-level spatial cues and high-level geometric abstractions to enhance instance awareness. Within the transformer, image tokens and object queries are optimized jointly. The image tokens primarily capture multi-view geometric and semantic information and are used to regress dense outputs, while the object queries focus on instance-level understanding and reasoning.

Update of Image Tokens. Each input image I_i is first processed by a frozen DINO [37] backbone to extract dense features, producing K image tokens denoted as a matrix $T_i \in \mathbb{R}^{K\times d}$. To construct a global representation of the 3D scene, the image tokens from all N views are concatenated along the sequence dimension to form an aggregated multi-view token matrix $T \in \mathbb{R}^{(NK)\times d}$, where $T = [T_1; T_2; \ldots; T_N]$[1]. These tokens are then fed into the unified transformer, where features are fused and updated through alternating frame attention and global attention. We denote the multi-view image tokens after optimization by the l-th $(1 \le l \le L)$ global attention layer as $T^{(l)}$, where $T^{(0)}$ corresponds to the original concatenated image tokens extracted by DINO.

Update of Object Queries. Different from prior works that focus solely on geometric estimation, SegVGGT aims to construct a transformer architecture that jointly captures geometric and instance-level information. To this end, we initialize O learnable object queries $(\mathbf{q}_j)_{j=1}^{O}$ and update them throughout the transformer layers. Specifically, after the image tokens $T^{(l)}$ undergo global attention for cross-view feature fusion, we insert a cross-attention module that enables the object queries to extract information from them. With the j-th object query acting as the query and the unified image tokens acting as the keys and values, this cross-attention process at the l-th layer can be formulated as:

$$A_j^{(l)} = \mathrm{Softmax}\left(\mathbf{q}_j^{(l-1)}(T^{(l)})^{\top}/\sqrt{d}\right), \quad \mathbf{q}_j'^{(l)} = \mathbf{q}_j^{(l-1)} + A_j^{(l)}T^{(l)}, \tag{2}$$

where $A_j^{(l)} \in \mathbb{R}^{1\times(NK)}$ represents the global cross-attention weight map of the j-th query over all multi-view image tokens, and $\mathbf{q}_j'^{(l)} \in \mathbb{R}^{1\times d}$ denotes the intermediate query updated by the cross-attention. Subsequently, a self-attention layer among the queries is applied to further enhance their feature representations, yielding the final output $\mathbf{q}_j^{(l)}$ for the current layer.

3.3 Synchronous 3D Decoding and Training Objective

Following the L layers of the geometry grounded transformer, the enriched image tokens $T^{(L)}$ and object queries $\mathbf{q}_j^{(L)}$ are fed into the decoding heads to yield the final 3D scene reconstruction and instance segmentation.

[1] In practice, each image is also augmented with a camera token and four register tokens, which are omitted here and in Fig. 2 for clarity.

Geometry Predictions. To reconstruct the multi-view geometry, we follow VGGT [52] by employing a Camera Head to regress the camera parameters $(\mathbf{g}_i)_{i=1}^N$, and a DPT [41] head to predict the dense depth maps $(D_i)_{i=1}^N$.
Instance Predictions. For instance-level understanding, a separate semantic DPT head is introduced to decode the image tokens into cross-view consistent instance-level feature maps $F_i \in \mathbb{R}^{d\times h\times w}$. Meanwhile, each optimized query $\hat{\mathbf{q}}_j = \mathbf{q}_j^{(L)}$ predicts semantic logits $\mathbf{c}_j \in \mathbb{R}^{C+1}$ via an MLP classification head, where C is the number of instance categories. To generate the multi-view 2D masks, we compute the dot product between each object query $\hat{\mathbf{q}}_j$ and the dense feature maps $(F_i)_{i=1}^N$. The probability mask $M_{j,i} \in [0,1]^{h\times w}$ for the j-th query on the i-th view is formulated as:

$$M_{j,i} = \sigma\left(\hat{\mathbf{q}}_j^\top F_i\right), \tag{3}$$

where σ denotes the sigmoid function. During inference, a pre-defined threshold τ_m is applied to binarize $M_{j,i}$. With the synchronously predicted camera parameters $\mathbf{g}_i$ and depth maps D_i, these binarized 2D masks can be directly unprojected into a unified 3D coordinate space, elegantly yielding the final 3D instance segmentation without relying on any complex point cloud post-processing.
Training Objective. The overall training objective $\mathcal{L}_{total}$ is composed of the geometric loss, the instance loss, and our proposed attention alignment loss:

$$\mathcal{L}_{total} = \mathcal{L}_{geo} + \mathcal{L}_{inst} + \lambda_{js}\mathcal{L}_{js}. \tag{4}$$

For 3D scene reconstruction, $\mathcal{L}_{geo} = \lambda_{camera}\mathcal{L}_{camera} + \lambda_{depth}\mathcal{L}_{depth}$ is used to supervise the predicted camera parameters $(\mathbf{g})_{i=1}^N$ and depth maps $(D_i)_{i=1}^N$. Both $\mathcal{L}_{camera}$ and $\mathcal{L}_{depth}$ directly follow the design in VGGT [52], and their detailed formulations are omitted here for brevity. Notably, we adopt a pre-trained VGGT model with frozen parameters as a teacher model to provide geometric supervision, which aims to mitigate overfitting of geometry predictions to noise and biases in the training data.

For 3D instance segmentation, we follow DETR [4] and employ the Hungarian algorithm [24] to match object queries with ground-truth instances, enabling end-to-end training with classification and mask supervision. Interestingly, we observe that by simply flattening the predicted multi-view masks $M_j \in [0,1]^{N\times H\times W}$ of each query into a 1D sequence $m_j \in [0,1]^{NHW}$, the representation becomes mathematically equivalent to a 3D point cloud mask. By applying the identical operation to the multi-view consistent 2D ground-truth masks, we can seamlessly inherit the matching cost and loss formulations from point cloud segmentation methods [23, 47]. Specifically, the pair-wise matching cost $C_{j,k}$ between the j-th prediction and the k-th ground truth is defined as:

$$C_{j,k} = -\lambda_{cls}\, c_{j,c_k} + \lambda_{mask}\left(\mathrm{BCE}(m_j, m_k^{gt}) + \mathrm{Dice}(m_j, m_k^{gt})\right) + \lambda_{js}\, C_{j,k}^{js}, \tag{5}$$

where c_{j,c_k} denotes the predicted probability of the j-th query belonging to the target class c_k, and $m_k^{gt} \in \{0,1\}^{NHW}$ is the flattened ground-truth mask. Here, $\mathrm{BCE}(\cdot,\cdot)$ and $\mathrm{Dice}(\cdot,\cdot)$ represent the binary cross-entropy and dice loss with

Laplace smoothing [34]. Crucially, $C^{js}_{j,k}$ represents our proposed frame-level attention alignment cost based on the Jensen-Shannon divergence, which explicitly encourages queries to match with instances that geometrically align with their attention priors (detailed in Sec. 3.4). Once the optimal assignment is established, the instance loss $\mathcal{L}_{inst}$ is computed over the matched pairs as a weighted sum of the classification loss and the mask segmentation losses:

$$\mathcal{L}_{inst} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{mask}\left(\mathcal{L}_{bce} + \mathcal{L}_{dice}\right). \tag{6}$$

Finally, $\mathcal{L}_{js}$ is our proposed frame-level attention alignment loss, which acts as a structural regularization during training and will be detailed in Sec. 3.4.

3.4 Frame-level Attention Distribution Alignment

A typical 3D scene inherently involves a massive number of aggregated multi-view image tokens (e.g., $NK \approx 10^4 \sim 10^5$) in our architecture. In such a high-dimensional space, the cross-attention weights of the object queries tend to become highly dispersed across the entire sequence. However, due to occlusion and limited fields of view, a specific 3D instance is typically visible in only a sparse subset of the N views. This naturally motivates a frame-level prior that queries should concentrate more attention on frames where their assigned instances are visible. Without explicit constraints, unguided object queries normally struggle to localize the correct views and are prone to absorbing irrelevant background noise, leading to optimization challenges and suboptimal performance.

To bridge this gap, we propose the Frame-level Attention Distribution Alignment (FADA) module to provide an explicit global-to-local guidance. Our core insight is to directly align the queries' attention distribution with the actual visibility of the target instances across the multi-view frames. Instead of solely applying this constraint at the final output, we employ deep supervision across the intermediate transformer layers to progressively refine the queries' geometric awareness. Specifically, at the l-th transformer layer, we extract the cross-attention weight map $A^{(l)}_j \in \mathbb{R}^{1\times(NK)}$ of the j-th query obtained from Equation (2). Since $A^{(l)}_j$ is a normalized probability distribution over all tokens, we can naturally obtain the predicted frame-level attention distribution $\hat{\mathbf{p}}^{(l)}_j \in [0,1]^N$ by marginalizing the weights over the spatial dimension of each view. For the i-th view I_i, its aggregated attention score is calculated as:

$$\hat{p}^{(l)}_{j,i} = \sum_{t\in \text{view}_i} A^{(l)}_{j,t}. \tag{7}$$

Meanwhile, for a ground-truth instance k, we deterministically construct its target visibility distribution $\mathbf{p}^{gt}_k \in [0,1]^N$. Specifically, the target probability $p^{gt}_{k,i}$ for the i-th view I_i is defined proportionally to the number of pixels the instance occupies in that frame relative to its total pixel count across all N views. This area-proportional distribution ensures that frames containing a larger visible portion of the target instance receive correspondingly higher attention priors,

while unobserved frames are naturally assigned a probability of zero. To measure the alignment between the predicted attention focus and the ground-truth visibility, we utilize the Jensen-Shannon (JS) divergence. Unlike the Kullback-Leibler (KL) divergence, the JS divergence is symmetric, inherently bounded, and numerically stable for distributions containing zero probabilities. The proposed FADA module plays a dual role in our end-to-end framework.

Prior-guided Matching. During the bipartite matching phase, we incorporate the JS divergence across all L transformer layers as a structural prior. To ensure numerical stability across different sequence lengths and network depths, the matching cost $C^{js}_{j,k}$ is normalized by the total number of frames N and layers L:

$$C^{js}_{j,k} = \frac{1}{L \cdot N} \sum_{l=1}^{L} \mathrm{JS}(\mathbf{p}^{gt}_{k} \parallel \hat{\mathbf{p}}^{(l)}_{j}). \tag{8}$$

This explicitly encourages queries to be assigned to instances that intrinsically align with their geometric attention focus throughout the entire progressive refinement process, easing the optimization difficulty.

Attention Regularization. After Hungarian matching with the cost in Equation (5), we obtain the optimal assignment $\mathcal{M}$ between object queries and ground-truth instances. We then apply the alignment loss $\mathcal{L}_{js}$ to penalize distribution inconsistency between matched pairs. Consistent with the matching cost, the overall alignment loss is averaged over all layers and frames:

$$\mathcal{L}_{js} = \frac{1}{|\mathcal{M}| \cdot L \cdot N} \sum_{(j,k)\in\mathcal{M}} \sum_{l=1}^{L} \mathrm{JS}(\mathbf{p}^{gt}_{k} \parallel \hat{\mathbf{p}}^{(l)}_{j}). \tag{9}$$

By imposing this structured deep guidance both in the assignment and optimization stages, SegVGGT enforces the object queries to firmly anchor on instance-relevant frames during progressive cross-view feature fusion. Notably, the computational overhead introduced by this module is negligible during training. Rather than requiring additional parameterized layers, it directly reuses the cross-attention weights inherently computed by the transformer. Furthermore, operating at the frame level significantly reduces the dimensionality of calculation, making it highly efficient. Crucially, as this module acts strictly as a regularization mechanism during training, it is completely removed during inference and introduces no additional computational cost or network parameters.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate SegVGGT on three widely adopted indoor datasets: ScanNetv2 [7], ScanNet200 [43], and ScanNet++ [61]. ScanNetv2 comprises 1,613 3D indoor scenes with dense semantic and instance annotations, covering 20 semantic classes and 18 instance categories. Following the official split, it

contains 1,201 training scenes, 312 validation scenes, and 100 test scenes. ScanNet200 shares the same underlying scenes and data splits as ScanNetv2, but significantly expands the label space from 20 to 200 fine-grained semantic categories, among which 198 are instance categories. It exhibits a severe long-tail distribution, presenting a much greater challenge for instance-level understanding. ScanNet++ is a recently introduced high-fidelity dataset comprising 856, 50, and 50 scenes for training, validation, and testing, respectively, with 84 categories for evaluating 3D instance segmentation. All three datasets provide point clouds, RGB-D images, and corresponding camera intrinsics and extrinsics. In these experiments, we train and evaluate on ScanNetv2 and ScanNet200, respectively, while ScanNet++ is used exclusively to assess the generalization capability of our method without any fine-tuning on it.

Implementation Details. We build SegVGGT upon the VGGT [52] architecture with $L = 24$ and initialize it with pre-trained weights. During training, the DINO backbone is kept frozen. To optimize computational efficiency, we employ LoRA [14] to fine-tune the frame attention and global attention modules, while fully training the newly introduced parameters. For each training iteration, we randomly sample 2-24 frames per scene as input. For geometric supervision, we utilize a frozen, pre-trained VGGT as a teacher model to distill robust depth maps and camera parameters. The instance segmentation branch is supervised using multi-view consistent 2D masks provided by the datasets, which are generated by projecting the 3D instance annotations onto the respective 2D frames. We train the models separately on ScanNetv2 and ScanNet200 using 8 NVIDIA A100 GPUs, taking approximately 2 days per dataset. For more implementation details, please refer to the appendix.

Evaluation Metrics. We follow the standard evaluation metric for 3D instance segmentation and report mean Average Precision (mAP), defined as the area under the precision-recall curve across multiple Intersection-over-Union (IoU) thresholds. Concretely, mAP is averaged over IoU thresholds from 50% to 95% with a step of 5%, providing a comprehensive measure of segmentation quality. We also report mAP_{50} and mAP_{25} at fixed IoU thresholds of 50% and 25%, respectively, to evaluate performance under different localization tolerances.

Evaluation Settings. To comprehensively evaluate SegVGGT, we compare it against state-of-the-art 3D instance segmentation methods under two primary input modalities: 1) Point Cloud-based Methods [15,23,25,31,32,39,46]. As the dominant paradigm, these approaches predominantly operate on point clouds sampled from benchmark meshes, constituting a privileged setting with direct access to accurate 3D geometry. This avoids reconstruction artifacts and projection inconsistencies, while assuming pre-reconstructed 3D inputs that may not be readily available in practical deployments. ODIN [15] instead takes posed RGB-D scans provided by the benchmark as input. Although less restrictive, it still relies on additional modalities beyond RGB images. 2) RGB Image-based Methods [28,66]. Approaches in this category, including ours, tackle a more practical yet challenging setting by predicting 3D instances purely from 2D images. Since ground-truth instance masks are defined on the official 3D meshes, pre-

Table 1: Comparison of SegVGGT with point cloud-based methods. Input modalities are denoted as follows: P (point cloud from the benchmark), I (RGB images), D (depth maps), and C (camera parameters). † denotes a two-stage pipeline that first reconstructs a point cloud from the input images using VGGT [52], and then applies OneFormer3D [23] for 3D instance segmentation on the reconstructed point cloud. ‡ denotes a stronger two-stage baseline using VGGT-predicted geometry followed by ODIN [15] under the favorable GT-aligned setting.

Method	Input Modality	ScanNetv2			ScanNet200		
		mAP	mAP_{50}	mAP_{25}	mAP	mAP_{50}	mAP_{25}
Mask3D [46]	P	55.2	73.7	85.3	27.4	37.0	42.3
QueryFormer [32]	P	56.5	74.2	83.3	28.1	37.1	43.4
MAFT [25]	P	59.9	76.5	-	29.2	38.2	43.3
SPFormer [47]	P	56.3	73.9	82.9	-	-	-
SGIFormer-L [60]	P	61.0	81.2	**88.9**	29.2	39.4	44.2
OneFormer3D [23]	P	59.3	78.1	86.4	30.2	40.9	44.6
Relation3D [31]	P	62.5	80.2	87.0	31.6	41.2	45.6
OpenMask3D [49]	P & I & D & C	-	-	-	15.4	19.9	23.1
Open-YOLO 3D [2]	P & I & D & C	-	-	-	24.7	31.7	36.2
SegDINO3D [39]	P & I & D & C	**64.0**	**81.5**	**88.9**	**40.2**	**52.4**	**58.6**
ODIN [15]	I & D & C	50.0	71.0	83.6	31.5	45.3	53.1
ODIN‡ [15]	I	-	-	-	19.3	33.1	46.7
OneFormer3D† [23]	I	5.4	10.2	17.4	2.5	4.1	6.4
Ours	I	**50.4**	**71.7**	**87.0**	**31.9**	**45.7**	**53.7**

dicted multi-view masks must be mapped to these benchmark point clouds for evaluation. We utilize the ground-truth depth maps and camera poses during this mapping stage for fair comparison. Nevertheless, this projection step inevitably introduces noise and misalignments, placing RGB-based approaches at an inherent disadvantage compared to point cloud-based baselines.

4.2 Comparison with Point Cloud-Based Methods

We first perform a class-aware comparison with standard point cloud-based 3D instance segmentation methods on the ScanNetv2 and ScanNet200 validation sets (Table 1). We categorize baselines by input modalities. While conventional methods directly operate on benchmark-provided point clouds and some incorporate additional modalities, ODIN [15] avoids pre-reconstructed meshes but still relies on accurate sensor point clouds to provide geometric information for inference. In contrast, SegVGGT is the only method relying exclusively on unposed 2D RGB images (uniformly sampled every 20 frames as input). This imposes the most stringent evaluation challenges, including inevitable projection misalignments and entirely unobserved ground-truth regions.

Methods without Benchmark Point Clouds. Among methods lacking direct access to the benchmark point clouds (bottom of Table 1), SegVGGT

Table 2: Comparison of SegVGGT with RGB image-based methods on the full validation sets. All methods are evaluated under the class-agnostic setting.

Method	ScanNetv2			ScanNet200			ScanNet++		
	mAP	mAP_{50}	mAP_{25}	mAP	mAP_{50}	mAP_{25}	mAP	mAP_{50}	mAP_{25}
PanSt3R [66]	17.8	35.0	53.9	15.9	30.1	45.5	7.3	16.9	29.1
IGGT [28]	21.2	33.3	46.4	21.1	33.2	46.0	7.7	15.4	28.2
Ours	**50.3**	**76.1**	**89.6**	**45.4**	**70.3**	**84.4**	**10.0**	**25.1**	**45.9**

achieves state-of-the-art results across all metrics on both datasets, despite operating under the strictest input constraints. To further validate the necessity of a unified architecture, we construct two-stage pipelines that combine VGGT [52] reconstruction with widely used 3D segmentation models. With OneFormer3D† [23], the mAP on ScanNet200 drops drastically from 30.2 to 2.5. This drop is partially caused by evaluation projection errors, but more importantly, it reflects the sensitivity of existing point cloud segmentation models to feed-forward reconstruction noise and geometry distribution shifts. We further evaluate a stronger ODIN‡ [15] baseline under a favorable GT-aligned setting, where ODIN uses aligned VGGT-predicted geometry and reaches only 19.3 mAP on ScanNet200, far below ODIN with benchmark RGB-D/cameras (31.5) and our unified model (31.9). These results support the need for joint reconstruction and segmentation rather than cascading off-the-shelf 3D segmentation models after feed-forward reconstruction. More details are provided in the appendix.

Methods on Benchmark Point Clouds. When compared to methods directly segmenting the benchmark point clouds, SegVGGT exhibits a gap in the strict mAP metric on ScanNetv2, but achieves competitive mAP_{25}, comparable to strong point cloud baselines such as Relation3D [31]. Consistent with observations in ODIN [15], we attribute the strict-metric gap to projection errors that lead to misalignments with the ground-truth point clouds during evaluation, while the mAP_{25} result suggests reliable instance localization despite this mismatch. On the significantly more challenging ScanNet200 benchmark, where the expanded fine-grained label space places greater emphasis on semantic appearance cues, SegVGGT outperforms pure point cloud-based methods while using only unposed RGB images. While SegDINO3D [39] achieves higher metrics, it relies on richer input modalities and additionally incorporates a powerful pre-trained 2D perception model [42]. In contrast, our RGB-only framework achieves robust performance with a simpler and more practical RGB-only deployment setup. A more detailed discussion of this comparison under different input assumptions is provided in the appendix.

4.3 Comparison with RGB Image-Based Methods

We further compare our method with recent RGB image-based approaches, PanSt3R [66] and IGGT [28], which concurrently perform 3D reconstruction and instance segmentation. Because IGGT relies on a clustering algorithm [33]

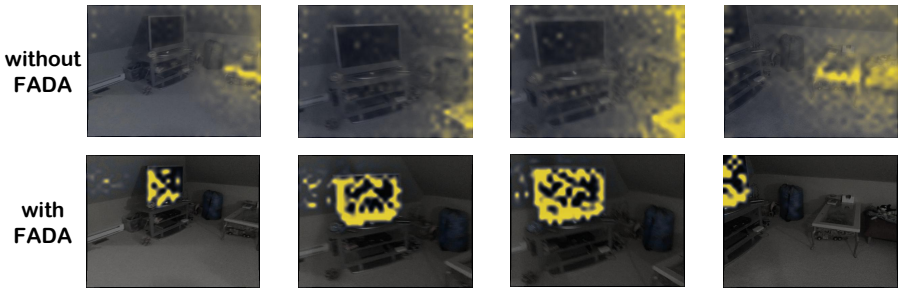

Fig. 3: Visualization of the FADA module's effect. We visualize the similarity map between the query corresponding to the TV instance and the image tokens in the last cross-attention layer. Without the FADA module, the query attention is dispersed across the entire scene. After introducing FADA, the query focuses much more accurately on the TV, indicating that FADA improves the model's ability to attend to target instances and thereby enhances instance segmentation performance.

to generate instance masks and cannot directly predict semantic categories, we conduct evaluations under a class-agnostic setting to ensure a fair comparison. To improve statistical reliability, we evaluate all RGB image-based methods on the full validation sets, including 312 scenes for ScanNetv2/ScanNet200 and 50 scenes for ScanNet++. For consistency, all methods use the same uniformly downsampled input frames, with a 40× sampling rate for ScanNetv2 and ScanNet200 and a 160× sampling rate for the denser ScanNet++ scenes.

As shown in Table 2, SegVGGT achieves state-of-the-art performance across all datasets. Specifically, it delivers substantial improvements on the ScanNetv2 and ScanNet200 benchmarks, demonstrating the efficacy of our model in purely vision-based 3D scene understanding. To further validate its robustness and generalization capability, we evaluate our model on the high-fidelity ScanNet++ dataset. Notably, while the baseline methods are trained on massive datasets that explicitly include ScanNet++ training scenes, our model is trained solely on ScanNet200 and applied to ScanNet++ without any fine-tuning. Despite this setting, SegVGGT consistently outperforms these baselines across all metrics. Particularly in localization precision, our method outperforms IGGT by 17.7 on the mAP_{25} metric, providing more reliable evidence of cross-dataset generalization. Overall, these results highlight the structural advantages of our query-based unified framework. By deeply coupling instance reasoning with geometric features, it effectively addresses the insufficient joint modeling present in PanSt3R, while elegantly bypassing the high-cost and heuristically-tuned post-clustering required by IGGT during inference.

4.4 Ablation Studies and Analysis

We conduct ablation studies on the ScanNet200 validation set to evaluate the effectiveness of different components in our method.

Ablation on FADA Module. As described in Sec. 3.4, we introduce the FADA module to mitigate the attention dispersion problem. To evaluate its effectiveness, we investigate its impact by decoupling its dual roles: attention regularization through the loss term and prior-guided matching through the cost term. As shown in the first row of Table 3, without any explicit guidance, the baseline model only achieves 26.7 mAP, indicating that unguided queries struggle to accurately segment instances. By introducing the JS divergence as a regularization loss during training, the performance substantially improves by +4.4 mAP, as shown in row 2. This significant gain validates that structurally aligning the attention distribution with target visibility effectively mitigates the attention dispersion issue. Furthermore, incorporating the JS divergence into the Hungarian matching cost yields the best performance of 31.9 mAP, as shown in row 3. This confirms that leveraging geometric attention priors to guide bipartite matching successfully reduces optimization difficulty, thereby further enhancing the overall accuracy and localization precision.

Table 3: Ablation on the FADA module. We decouple the FADA module into loss supervision and cost function.

Row	Loss	Cost	mAP	mAP_{50}	mAP_{25}
1	✗	✗	26.7	39.2	48.6
2	✓	✗	31.1	**45.7**	52.3
3	✓	✓	**31.9**	**45.7**	**53.7**

Table 4: Ablation on where to insert query interaction

Inserted Layers	mAP	mAP_{50}	mAP_{25}
Early 12	23.8	36.9	45.6
Late 12	30.5	45.1	53.0
Interleaved	30.4	44.5	51.8
All 24	**31.9**	**45.7**	**53.7**

In addition, we further analyze the effect of the FADA module through visualization. We train two models, one with the FADA module and one without it, and identify the query with the highest score for the TV instance in the same scene. As shown in Fig. 3, we visualize the similarity maps between the selected query and the global image tokens in the last cross-attention layer, and present four views. The results show that without the FADA module, the query attention is highly dispersed and often focuses on incorrect regions. In contrast, after introducing the FADA module, the query attention accurately concentrates on the TV instance. This observation further confirms that the FADA module effectively mitigates attention dispersion and guides the queries to focus on the target instance, thereby improving instance segmentation performance.

Ablation on Multi-level Geometric Features. In SegVGGT, object queries are progressively updated through cross-attention with image tokens. To validate the necessity of this deep integration, we ablate the insertion locations of these query interactions within the 24-layer transformer. As illustrated in Table 4, constraining interactions only to the early 12 layers yields a poor mAP of 23.8, as early tokens primarily capture local 2D appearance cues without an established multi-view geometric consensus. Conversely, interacting solely with the late 12 layers improves the mAP to 30.5, highlighting the critical role of global

3D geometric information for instance reasoning. Similarly, interleaving 12 interaction layers throughout the network yields 30.4 mAP. Ultimately, inserting the interaction modules across all 24 layers achieves the optimal performance. This clearly demonstrates that object queries benefit most when they continuously co-evolve with the image tokens, seamlessly absorbing multi-level features ranging from low-level spatial details to high-level geometric structures.

5 Conclusion

In this work, we presented SegVGGT, a unified feed-forward framework that simultaneously performs 3D scene reconstruction and 3D instance segmentation directly from unposed multi-view RGB images in a single forward pass. By deeply integrating object queries into a visual geometry grounded transformer, instance-level representations progressively co-evolve with multi-view geometric features, enabling tight coupling between geometry estimation and instance reasoning. To address the severe attention dispersion arising from massive global image tokens, we propose the Frame-level Attention Distribution Alignment (FADA) module, which serves a dual role as a structured bipartite matching prior and an attention regularization loss, explicitly anchoring queries to instance-relevant frames without any extra inference overhead. Extensive experiments demonstrate that SegVGGT achieves state-of-the-art performance among RGB image-based methods, surpasses prominent point cloud-native baselines on the challenging ScanNet200 benchmark, and generalizes favorably to ScanNet++. As future work, we will further scale the training data and extend this unified paradigm toward metric-scale reconstruction and open-vocabulary 3D perception.

References

1. Al Khatib, S., El Amine Boudjoghra, M., Lahoud, J., Khan, F.S.: 3D Instance Segmentation via Enhanced Spatial and Semantic Supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 541–550 (2023)
2. Boudjoghra, M.E.A., Dai, A., Lahoud, J., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.S.: Open-YOLO 3D: Towards Fast and Accurate Open-Vocabulary 3D Instance Segmentation. arXiv preprint arXiv:2406.02548 (2024)
3. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: MUSt3R: Multi-view Network for Stereo 3D Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1050–1060 (2025)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020)
5. Chen, S., Fang, J., Zhang, Q., Liu, W., Wang, X.: Hierarchical Aggregation for 3D Instance Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15467–15476 (2021)

6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
8. Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., Nießner, M.: 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9031–9040 (2020)
9. Fan, Z., Zhang, J., Cong, W., Wang, P., Li, R., Wen, K., Zhou, S., Kadambi, A., Wang, Z., Xu, D., et al.: Large Spatial Model: End-to-end Unposed Images to Semantic 3D. Advances in neural information processing systems **37**, 40212–40229 (2024)
10. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
11. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
12. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge university press (2003)
13. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4421–4430 (2019)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank Adaptation of Large Language Models. Iclr **1**(2), 3 (2022)
15. Jain, A., Katara, P., Gkanatsios, N., Harley, A.W., Sarch, G., Aggarwal, K., Chaudhary, V., Fragkiadaki, K.: ODIN: A Single Model for 2D and 3D Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3564–3574 (2024)
16. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3R: Empowering Unconstrained 3D Reconstruction with Camera and Scene Priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1071–1081 (2025)
17. Jiang, H., Yan, F., Cai, J., Zheng, J., Xiao, J.: End-to-End 3D Point Cloud Instance Segmentation Without Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12796–12805 (2020)
18. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition. pp. 4867–4876 (2020)
19. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: AnySplat: Feed-forward 3D Gaussian Splatting from Unconstrained Views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
20. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal Feed-forward Metric 3d Reconstruction. arXiv preprint arXiv:2509.13414 (2025)

21. Koch, S., Wald, J., Matsuki, H., Hermosilla, P., Ropinski, T., Tombari, F.: Unified Semantic Transformer for 3D Scene Understanding. arXiv preprint arXiv:2512.14364 (2025)
22. Kolodiazhnyi, M., Vorontsova, A., Konushin, A., Rukhovich, D.: Top-Down Beats Bottom-Up in 3D Instance Segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3566–3574 (2024)
23. Kolodiazhnyi, M., Vorontsova, A., Konushin, A., Rukhovich, D.: OneFormer3D: One Transformer for Unified Point Cloud Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20943–20953 (2024)
24. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
25. Lai, X., Yuan, Y., Chu, R., Chen, Y., Hu, H., Jia, J.: Mask-Attention-Free Transformer for 3D Instance Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3693–3703 (2023)
26. Leroy, V., Cabon, Y., Revaud, J.: Grounding Image Matching in 3D with MASt3R. In: European conference on computer vision. pp. 71–91. Springer (2024)
27. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven Semantic Segmentation. arXiv preprint arXiv:2201.03546 (2022)
28. Li, H., Zou, Z., Liu, F., Zhang, X., Hong, F., Cao, Y., Lan, Y., Zhang, M., Yu, G., Zhang, D., et al.: IGGT: Instance-Grounded Geometry Transformer for Semantic 3D Reconstruction. arXiv preprint arXiv:2510.22706 (2025)
29. Li, H., Qu, J., Zhang, L.: OVSeg3R: Learn Open-vocabulary Instance Segmentation from 2D via 3D Reconstruction. arXiv preprint arXiv:2509.23541 (2025)
30. Liang, Z., Li, Z., Xu, S., Tan, M., Jia, K.: Instance Segmentation in 3D Scenes Using Semantic Superpoint Tree Networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2783–2792 (2021)
31. Lu, J., Deng, J.: Relation3D: Enhancing Relation Modeling for Point Cloud Instance Segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8889–8899 (2025)
32. Lu, J., Deng, J., Wang, C., He, J., Zhang, T.: Query Refinement Transformer for 3D Instance Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18516–18526 (2023)
33. McInnes, L., Healy, J.: Accelerated Hierarchical Density Based Clustering. In: 2017 IEEE international conference on data mining workshops (ICDMW). pp. 33–42. IEEE (2017)
34. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
35. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE transactions on robotics **33**(5), 1255–1262 (2017)
36. Nguyen, P., Ngo, T.D., Kalogerakis, E., Gan, C., Tran, A., Pham, C., Nguyen, K.: Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4018–4028 (2024)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193 (2023)

38. Park, K.B., Kim, M., Choi, S.H., Lee, J.Y.: Deep Learning-based Smart Task Assistance in Wearable Augmented Reality. Robotics and Computer-Integrated Manufacturing **63**, 101887 (2020)
39. Qu, J., Li, H., Chen, X., Liu, S., Shi, Y., Ren, T., Jing, R., Zhang, L.: SegDINO3D: 3D Instance Segmentation Empowered by Both Image-Level and Object-Level 2D Features. arXiv preprint arXiv:2509.16098 (2025)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
42. Ren, T., Chen, Y., Jiang, Q., Zeng, Z., Xiong, Y., Liu, W., Ma, Z., Shen, J., Gao, Y., Jiang, X., et al.: DINO-X: A Unified Vision Model for Open-World Object Detection and Understanding. arXiv preprint arXiv:2411.14347 (2024)
43. Rozenberszki, D., Litany, O., Dai, A.: Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In: European Conference on Computer Vision. pp. 125–141. Springer (2022)
44. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
45. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise View Selection for Unstructured Multi-View Stereo. In: European Conference on Computer Vision (ECCV) (2016)
46. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8216–8223. IEEE (2023)
47. Sun, J., Qing, C., Tan, J., Xu, X.: Superpoint Transformer for 3D Scene Instance Segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2393–2401 (2023)
48. Sun, X., Jiang, H., Liu, L., Nam, S., Kang, G., Wang, X., Sui, W., Su, Z., Liu, W., Wang, X., et al.: Uni3R: Unified 3D Reconstruction and Semantic Understanding via Generalizable Gaussian Splatting from Unposed Multi-View Images. arXiv preprint arXiv:2508.03643 (2025)
49. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: OpenMask3D: Open-Vocabulary 3D Instance Segmentation. arXiv preprint arXiv:2306.13631 (2023)
50. Vu, T., Kim, K., Luu, T.M., Nguyen, T., Yoo, C.D.: SoftGroup for 3D Instance Segmentation on Point Clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2708–2717 (2022)
51. Wang, D., Liu, J., Gong, H., Quan, Y., Wang, D.: CompetitorFormer: Competitor Transformer for 3D Instance Segmentation. arXiv preprint arXiv:2411.14179 (2024)
52. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual Geometry Grounded Transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
53. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D Vision Made Easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)

54. Wang, X., Liu, S., Shen, X., Shen, C., Jia, J.: Associatively Segmenting Instances and Semantics in Point Clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4096–4105 (2019)
55. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-Equivariant Visual Geometry Learning. arXiv preprint arXiv:2507.13347 (2025)
56. Wu, C., Wang, H., Ji, J., Yao, Y., Du, C., Kang, J., Fu, Y., Cao, L.: MVGGT: Multimodal Visual Geometry Grounded Transformer for Multiview 3D Referring Expression Segmentation. arXiv preprint arXiv:2601.06874 (2026)
57. Xie, C., Xiang, Y., Mousavian, A., Fox, D.: Unseen Object Instance Segmentation for Robotic Environments. IEEE Transactions on Robotics **37**(5), 1343–1359 (2021)
58. Xu, Q., Wei, D., Zhao, L., Li, W., Huang, Z., Ji, S., Liu, P.: SIU3R: Simultaneous Scene Understanding and 3D Reconstruction Beyond Feature Alignment. arXiv preprint arXiv:2507.02705 (2025)
59. Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., Trigoni, N.: Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. Advances in neural information processing systems **32** (2019)
60. Yao, L., Wang, Y., Liu, M., Chau, L.P.: SGIFormer: Semantic-guided and Geometric-enhanced Interleaving Transformer for 3D Instance Segmentation. IEEE Transactions on Circuits and Systems for Video Technology **35**(3), 2276–2288 (2024)
61. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
62. Yi, L., Zhao, W., Wang, H., Sung, M., Guibas, L.J.: GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3947–3956 (2019)
63. Yin, Y., Liu, Y., Xiao, Y., Cohen-Or, D., Huang, J., Chen, B.: SAI3D: Segment Any Instance in 3D Scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3292–3302 (2024)
64. Zhang, B., Wonka, P.: Point Cloud Instance Segmentation Using Probabilistic Embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8883–8892 (2021)
65. Zhuang, C., Li, S., Ding, H.: Instance Segmentation Based 6D Pose Estimation of Industrial Objects Using Point Clouds for Robotic Bin-picking. Robotics and Computer-Integrated Manufacturing **82**, 102541 (2023)
66. Zust, L., Cabon, Y., Marrie, J., Antsfeld, L., Chidlovskii, B., Revaud, J., Csurka, G.: PanSt3R: Multi-view Consistent Panoptic Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5856–5866 (2025)