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

SemGS: Feed-Forward Semantic 3D Gaussian Splatting from Sparse Views for Generalizable Scene Understanding

Sheng Ye[1], Zhen-Hui Dong[1], Ruoyu Fan[1], Tian Lv[1], and Yong-Jin Liu[1,*], *Senior Member, IEEE*

***Abstract*— Semantic understanding of 3D scenes is essential for robots to operate effectively and safely in complex environments. Existing methods for semantic scene reconstruction and semantic-aware novel view synthesis often rely on dense multi-view inputs and require scene-specific optimization, limiting their practicality and scalability in real-world applications. To address these challenges, we propose SemGS, a feed-forward framework for reconstructing generalizable semantic fields from sparse image inputs. SemGS uses a dual-branch architecture to extract color and semantic features, where the two branches share shallow CNN layers, allowing semantic reasoning to leverage textural and structural cues in color appearance. We also incorporate a camera-aware attention mechanism into the feature extractor to explicitly model geometric relationships between camera viewpoints. The extracted features are decoded into dual-Gaussians that share geometric consistency while preserving branch-specific attributes, and further rasterized to synthesize semantic maps under novel viewpoints. Additionally, we introduce a regional smoothness loss to enhance semantic coherence. Experiments show that SemGS achieves state-of-the-art performance on benchmark datasets, while providing rapid inference and strong generalization capabilities across diverse synthetic and real-world scenarios.**

I. INTRODUCTION

Semantic understanding of 3D scenes is a fundamental challenge in computer vision and robotics. For intelligent robots to operate safely and efficiently in unknown environments, they must go beyond low-level appearance perception and gain a high-level semantic understanding of their surroundings. Such semantic awareness is crucial for tasks like navigation [1], [2], obstacle avoidance [3], and decision-making [4]. While recent advances in 3D scene representation — such as Neural Radiance Fields (NeRF) [5] and 3D Gaussian Splatting (3DGS) [6] — have achieved remarkable rendering fidelity, they only provide implicit geometry and appearance details, without semantic reasoning. Thus, there is a growing need to integrate semantic information into these 3D representations. Despite its importance, semantic scene understanding and semantic-aware novel view synthesis under sparse inputs remain under-explored.

Some pioneering works [7], [8], [9] have extended NeRF and 3DGS to incorporate semantics, either through auxiliary semantic branches or embedding semantic features [10], [11]. However, these methods typically rely on dense multi-view images, which are costly to acquire. Furthermore, they are generally optimized in a scene-specific manner. For each new scenario, these methods have to retrain a new model, which severely limits their scalability and real-world applicability.

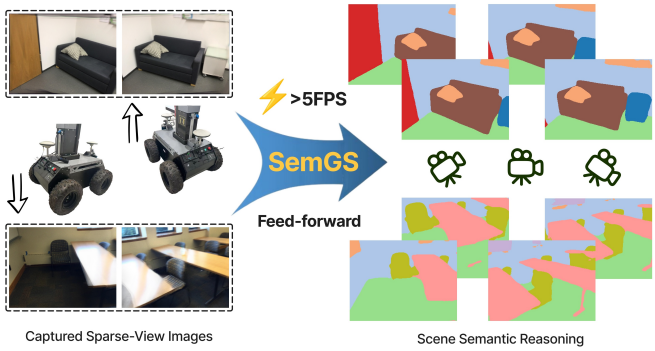

Fig. 1. We propose **SemGS**, a novel framework for generalizable semantic 3DGS. Given sparse-view images of an unseen scene, **SemGS** can rapidly infer semantic maps under novel viewpoints in a single feed-forward pass.

In this work, we aim to learn a generalizable semantic representation that can be trained across multiple scenes and infer semantic maps under novel viewpoints from only sparse input images. As shown in Fig. 1, our method enables fast semantic inference in a single feed-forward pass. Achieving this capability requires the model to possess strong geometric reasoning and generalization abilities, rather than merely overfitting to a specific scene. Our work is motivated by recent feed-forward 3DGS methods such as MVSplat [12], which leverages neural networks to predict Gaussian attributes from sparse views in a feed-forward pass. While these feed-forward methods focus solely on color rendering, we observe that through cost-volume based depth reasoning, such models inherently capture rich geometric priors that are also highly informative for semantic inference. Since visual appearance and semantics are often closely related, extending the feed-forward 3DGS from color to semantic can be both intuitive and beneficial.

To achieve this goal, we propose **SemGS**, a novel feed-forward framework for generalizable semantic 3DGS. Our method employs dual branches for color and semantic feature extraction, each consisting of a CNN backbone and a Swin Transformer [13], with cross-attention mechanisms fusing information across multiple input views. By sharing low-level CNN feature layers between the two branches, **SemGS** enables semantic reasoning to leverage textural and structural cues embedded in appearance representations. Both color and semantic features are decoded into dual-Gaussians (*color Gaussians* and *semantic Gaussians*), whose geometric position and opacity attributes are shared and derived from the

[1]S. Ye, Z.-H. Dong, R. Fan, T. Lv, and Y.-J. Liu are with the Department of Computer Science, Tsinghua University, Beijing, China {yec22, dzh23, fry21, lt22}@mails.tsinghua.edu.cn, liuyongjin@tsinghua.edu.cn
*Corresponding Author

cost-volume based depth. This design allows semantic Gaussians to inherit strong 3D geometric priors from the color reconstruction branch. These Gaussians are then splatted [6] to render novel views and semantic maps.

Inspired by PRoPE [14], we integrate camera intrinsic and extrinsic parameters into the Swin Transformer's attention blocks via relative positional encoding, which enhances the 3D geometric awareness. We further introduce a regional smoothness loss to promote local semantic label consistency across neighboring regions. Experiments show that **SemGS** achieves state-of-the-art performance on the ScanNet [15] and ScanNet++ [16] datasets, and generalizes robustly across synthetic (Replica [17]) and real-world scenarios.

In summary, our contributions are:

- We propose **SemGS**, a novel feed-forward framework for joint radiance and semantic field reconstruction from sparse input images, enabling rapid semantic inference without the need for per-scene optimization.
- We incorporate camera geometry into the Swin Transformer via relative positional encoding to facilitate 3D perception, and introduce a regional smoothness loss to enforce semantic coherence.
- Experiments show that **SemGS** outperforms existing baselines on benchmark datasets, while offering faster inference speed and stronger generalization ability.

II. RELATED WORKS

A. Generalizable Novel View Synthesis

The novel view synthesis task aims to generate photo-realistic images from previously unseen camera viewpoints. NeRF [5] represents a significant breakthrough, modeling 3D scenes as continuous implicit functions encoded by MLPs and synthesizing novel views through volume rendering with high fidelity. Following this, 3D Gaussian Splatting (3DGS) [6] further improves rendering efficiency using a differentiable point-based splatting technique. However, both NeRF and 3DGS generally require per-scene optimization, which limits their scalability and generalizability.

To achieve cross-scene generalization, pixelNeRF [18] conditions the radiance field on input image features, enabling synthesis from sparse input views. Subsequent approaches such as MVSNeRF [19] and GeoNeRF [20] integrate multi-view stereo and geometric priors to enhance rendering quality. As 3DGS offers superior rendering quality and efficiency over NeRF, its generalizable extensions (a.k.a. *feed-forward 3DGS*) have attracted growing interest. For instance, PixelSplat [21] introduces a multi-view epipolar transformer to predict pixel-aligned Gaussian parameters in a feed-forward manner, while MVSplat [12] leverages cost volumes for improved geometry estimation. Despite these advances, existing methods focus primarily on color rendering, leaving the problem of generalizable semantic reasoning largely unaddressed.

B. Semantic Fields and Scene Understanding

Semantic scene understanding is essential for many vision and robotic applications, including navigation [2] and human–robot interaction [22], as it provides interpretable, high-level cues beyond raw appearance. Conventional 3D semantic segmentation methods operate on point clouds [23], [24] or meshes [25], [26], but require expensive 3D annotations for training. Recently, 3D semantic field methods have emerged that allow semantic map rendering from arbitrary viewpoints using only 2D supervision. Specifically, Semantic-NeRF [7] extends the radiance field with a semantic branch to jointly render color and semantic labels. Panoptic NeRF [27] introduces a 3D-to-2D label transfer strategy to optimize both semantic and radiance fields. LERF [8] incorporates CLIP [10] features into NeRF for open-vocabulary semantic segmentation. In the context of 3DGS, several methods have been proposed: LangSplat [9] leverages SAM [28] and CLIP embeddings to learn hierarchical semantic fields, Feature 3DGS [29] enables semantic rendering via 3D feature field distillation, and Semantic Gaussians [30] introduce a 3D semantic network to associate semantic attributes with each Gaussian. Nevertheless, these methods are confined to scene-specific optimization and cannot generalize across scenes.

Research on generalizable semantic fields is still relatively limited. S-Ray [31] learns a generalizable semantic field by building a 3D context space with cross-reprojection attention. GSNeRF [32] aggregates visual features and depth estimates to render novel-view semantic segmentations. In contrast to these works, we propose a novel framework based on feed-forward 3DGS that efficiently constructs generalizable semantic fields from sparse-view inputs, achieving superior performance with faster inference speed.

III. METHOD

We propose **SemGS**, a feed-forward framework for reconstructing 3D scenes with semantic understanding from sparse input views. Given N sparse-view RGB images $\mathcal{I} = \{\mathbf{I}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N}$, and their corresponding camera poses $\mathcal{P} = \{\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}_{i=1}^{N}$, our model predicts a set of Gaussian primitives that jointly represent the geometry, radiance, and semantics of the 3D scene.

A key aspect of our method is the use of a **dual-Gaussian representation**, in which each pixel of the input images is associated with two complementary Gaussians: a *color Gaussian* for radiance modeling and a *semantic Gaussian* for semantic reasoning. To exploit geometric priors and ensure consistency, both Gaussians share the same 3D position $\boldsymbol{\mu}_j$ and opacity α_j, while maintaining their own branch-specific attributes: $(\boldsymbol{\Sigma}_j^c, \mathbf{c}_j)$ for color Gaussians and $(\boldsymbol{\Sigma}_j^s, \mathbf{s}_j)$ for semantic Gaussians, where $\boldsymbol{\Sigma}_j^c, \boldsymbol{\Sigma}_j^s$ denote the covariance matrices, $\mathbf{c}_j$ is the color encoded via spherical harmonics, and $\mathbf{s}_j$ is the semantic class label distribution.

As shown in Fig. 2, our architecture comprises two parallel feature extraction branches (one for color and one for semantics), followed by a Gaussian decoder that outputs shared geometric attributes and branch-specific attributes. Finally, the predicted Gaussians are rendered into novel views and semantic maps via a differentiable rasterizer [6].

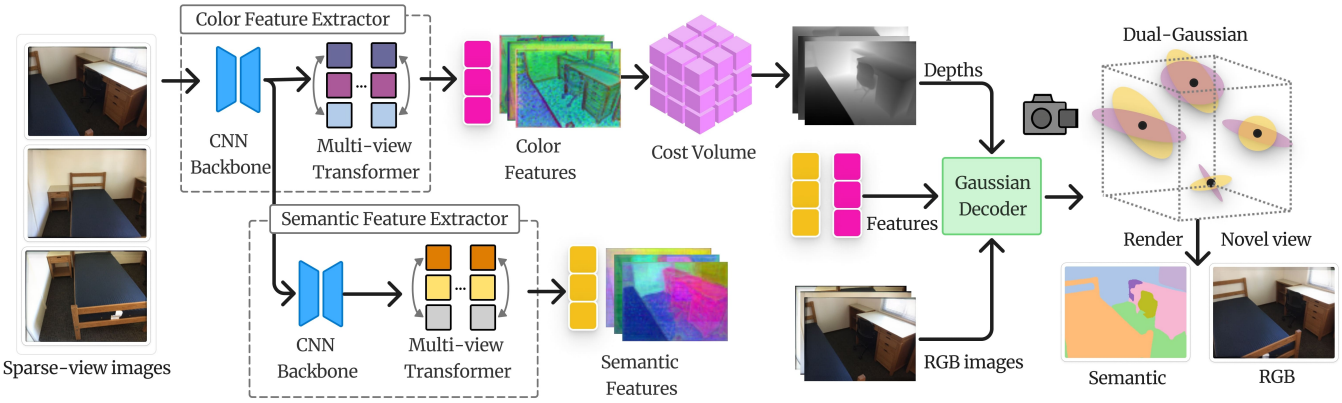

Fig. 2. Overall pipeline of our proposed **SemGS**. Given sparse-view RGB images as inputs, **SemGS** leverages a dual-branch architecture (Sec. III-A) to extract both color and semantic features. These features are used to regress multi-view depth maps (Sec. III-B) and are subsequently decoded into per-pixel dual-Gaussians (Sec. III-C). The resulting dual-Gaussians share geometric attributes while maintaining branch-specific attributes, enabling efficient rasterization for synthesizing both novel RGB views and semantic maps.

A. Multi-View Feature Extraction

1) Dual-Branch Feature Extractor: We extract multi-view features through a dual-branch architecture tailored for capturing color and semantic information. Both branches share the low-level CNN layers to capture fundamental texture and structure patterns, while maintaining branch-specific Swin Transformers [13] for high-level feature learning. Concretely, given input images $\{\mathbf{I}_i\}_{i=1}^N$, the low-level features are first extracted by a shared CNN backbone. In the color branch, these features are directly processed by a color Transformer to obtain per-view color features $\{\mathbf{F}_i^c\}_{i=1}^N$. In parallel, the semantic branch employs an additional CNN to refine the features into semantic cues, which are subsequently fed into a semantic Transformer to yield per-view semantic features $\{\mathbf{F}_i^s\}_{i=1}^N$. This design enables both branches to exploit shared low-level textural and structural information while learning branch-specific high-level features.

2) Camera Pose Injection: Our dual-branch feature extractor utilizes Swin Transformer to learn high-level features through its shifted window mechanism, which effectively captures both local and global context. To further enhance inter-view consistency, Swin Transformers are extended with cross-attention layers that propagate information across multiple input views, enabling powerful 3D geometric reasoning. Accurate geometric and semantic perception relies on camera information, particularly under sparse-view inputs. However, the original Swin Transformer struggles to capture geometric relationships across camera viewpoints. Therefore, we propose to adopt a camera-aware attention mechanism (shown in Fig. 3) inspired by PRoPE [14]. Specifically, we inject the camera poses (*i.e.,* projective transformations) into attention process by applying encoding to the queries, keys, and values of tokens (*i.e.,* image patch embeddings) in Swin Transformer. Formally, the relative camera projective transformation between view i and j is:

$$\tilde{\mathbf{P}}_{i\to j} = \tilde{\mathbf{P}}_i \tilde{\mathbf{P}}_j^{-1},\ \tilde{\mathbf{P}}_i = \begin{bmatrix} \mathbf{K}_i\mathbf{R}_i & \mathbf{K}_i\mathbf{t}_i \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{1}$$

where $\tilde{\mathbf{P}}$ denotes the world-to-camera projection matrix, $\mathbf{K}_i$, $\mathbf{R}_i$ and $\mathbf{t}_i$ represent the camera intrinsic, rotation, and translation of view i, respectively. For a token t, to jointly encode the inter-view relations and intra-view token positional order, we construct a block-diagonal matrix $\mathbf{G}_t \in \mathbb{R}^{d\times d}$:

$$\mathbf{G}_t = \begin{bmatrix} \mathbf{G}_t^{\text{proj}} & 0 \\ 0 & \mathbf{G}_t^{\text{rope}} \end{bmatrix}. \tag{2}$$

In this block-diagonal matrix $\mathbf{G}_t$, $\mathbf{G}_t^{\text{proj}} = \mathbf{I}_{d/8} \otimes \tilde{\mathbf{P}}_{i(t)} \in \mathbb{R}^{\frac{d}{2}\times\frac{d}{2}}$ encodes camera-level information, and

$$\mathbf{G}_t^{\text{rope}} = \begin{bmatrix} \text{RoPE}(x_t) & \mathbf{0} \\ \mathbf{0} & \text{RoPE}(y_t) \end{bmatrix} \in \mathbb{R}^{\frac{d}{2}\times\frac{d}{2}} \tag{3}$$

encodes intra-view positional information via rotary embeddings [33]. Here, $\text{RoPE}(\cdot)$ represents $\frac{d}{4}\times\frac{d}{4}$ rotary position embedding for the (x_t, y_t) coordinate of token t. Following the GTA-style attention [34], $\mathbf{G}_t$ is then applied to the query, key, and value of token t as:

$$\mathbf{Q}'_t = (\mathbf{G}_t)^\top \mathbf{Q}_t,\ \mathbf{K}'_t = (\mathbf{G}_t)^{-1}\mathbf{K}_t,\ \mathbf{V}'_t = (\mathbf{G}_t)^{-1}\mathbf{V}_t. \tag{4}$$

Finally, the attention output $\mathbf{O}_t$ is computed as:

$$\mathbf{O}_t = \sum_u \mathbf{A}_{t,u}\,\mathbf{V}'_u,\ \mathbf{A}_{t,u} = \text{Softmax}\Big(\frac{\mathbf{Q}'_t(\mathbf{K}'_u)^\top}{\sqrt{d}}\Big), \tag{5}$$

where u indexes all attended tokens, d denotes the embedding dimension. By adopting the camera-aware attention, our model strengthens cross-view geometric consistency and enhances semantic reasoning ability in sparse-view settings.

B. Multi-View Depth Estimation

1) Cost Volume Construction: Building on camera-aware multi-view color features $\{\mathbf{F}_i^c\}_{i=1}^N$ extracted in the previous stage, we construct the cost volume using a plane-sweep stereo strategy following MVSplat [12]. For each reference view i, we uniformly sample L depth candidates $\{d_m\}_{m=1}^L$ within a predefined near-to-far range. The color feature $\mathbf{F}_j^c$ from a source view j is then warped to the reference view i at each depth candidate d_m, producing a set of warped features

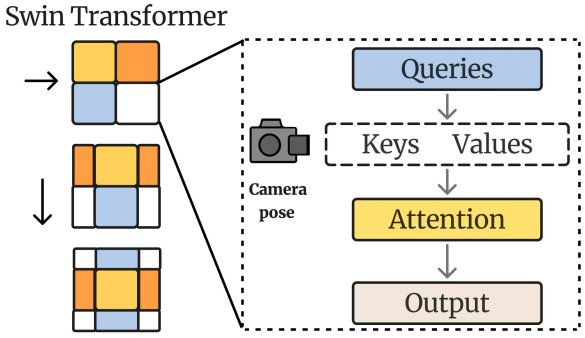

Fig. 3. Swin Transformer with camera-aware attention. Different colors denote attention windows, and the window shifting mechanism enables model to capture cross-window connections. Camera poses are injected into the attention process to enhance the 3D reasoning ability.

$\{\mathbf{F}_{d_m}^{j\to i}\}_{m=1}^{L}$. We then compute the correlation between these warped features and the original reference feature $\mathbf{F}_i^c$, and aggregate the results across all source views to form the 3D cost volume for reference view i, denoted as $\mathbf{C}_i$.

2) Depth Regression: Given the constructed cost volumes $\{\mathbf{C}_i\}_{i=1}^N$ and corresponding Transformer features $\{\mathbf{F}_i^c\}_{i=1}^N$, we predict per-pixel depth map $\{\mathbf{D}_i\}_{i=1}^N$ for each input view. Specifically, the cost volume and Transformer feature of each view are concatenated and fed into a lightweight 2D CNN U-Net to produce per-pixel depth probability distributions. The final depth map is then computed as the expectation over L depth candidates. This per-view depth estimation provides the geometric foundation for predicting Gaussian parameters in the subsequent stage.

C. Semantic Reasoning and Gaussian Parameter Prediction

Leveraging the extracted multi-view features $\{\mathbf{F}_i^c, \mathbf{F}_i^s\}_{i=1}^N$ and estimated multi-view depth maps $\{\mathbf{D}_i\}_{i=1}^N$, we decode a set of Gaussians to jointly represent the geometry, appearance, and semantics of the scene. We adopt a dual-Gaussian representation where each pixel of input views corresponds to two Gaussians: a *color Gaussian* for radiance modeling and a *semantic Gaussian* for semantic reasoning. Both Gaussians share geometric attributes while maintaining branch-specific attributes for their respective purposes.

1) Shared Geometric Attributes: Given the per-view depth maps $\{\mathbf{D}_i \in \mathbb{R}^{H\times W}\}_{i=1}^N$ estimated in the previous stage, the 3D Gaussian positions $\{\boldsymbol{\mu}_j\}_{j=1}^{N\times H\times W}$ are computed by back-projecting all the pixels to 3D points using the corresponding camera parameters. The depth probability distribution shares a similar physical meaning to the opacity (points with higher probabilities are more likely on the surface). Accordingly, the Gaussian opacities $\{\alpha_j \in [0,1]\}_{j=1}^{N\times H\times W}$ are predicted using a CNN from the depth probability distributions.

2) Branch-Specific Attributes: The Gaussian positions and opacities implicitly define the 3D scene geometry. Building on these geometric attributes, we reason about other Gaussian attributes from the multi-view Transformer features to model the appearance and semantic. Given the input images $\{\mathbf{I}_i\}_{i=1}^N$ and color Transformer features $\{\mathbf{F}_i^c\}_{i=1}^N$, a lightweight CNN is used to regress the color coefficients $\{\mathbf{c}_j\}_{j=1}^{N\times H\times W}$ and covariance $\{\mathbf{\Sigma}_j^c\}_{j=1}^{N\times H\times W}$ of *color Gaussians*. Symmetrically, another lightweight CNN is used to regress the semantic class distribution $\{\mathbf{s}_j\}_{j=1}^{N\times H\times W}$ and covariance $\{\mathbf{\Sigma}_j^s\}_{j=1}^{N\times H\times W}$ of *semantic Gaussians*, conditioned on the input images $\{\mathbf{I}_i\}_{i=1}^N$ and semantic Transformer features $\{\mathbf{F}_i^s\}_{i=1}^N$.

With these estimated attributes, the Gaussians can then be rasterized [6] to synthesize novel views and semantic maps. This dual-Gaussian representation jointly encodes the geometry, appearance, and semantics, enabling unified modeling and holistic reasoning of the 3D scene.

D. Training

In our **SemGS**, the color branch of the feature extractor and the depth regression 2D CNN are initialized with pre-trained weights from the feed-forward 3DGS model MVSplat [12], which provides reliable priors on scene geometry and appearance. The semantic branch is trained from scratch. Nevertheless, as *semantic Gaussians* and *color Gaussians* share geometric attributes, semantic inference also leverages the geometric priors of the pre-trained model.

To supervise predicted Gaussians, we employ a semantic cross-entropy loss: $\mathcal{L}_{sem} = -\sum_l \mathbf{S}^l log\hat{\mathbf{S}}^l$, where l indexes the semantic classes, $\mathbf{S}^l$ is the ground-truth class label, and $\hat{\mathbf{S}}^l$ is the predicted class probability distribution, and a color MSE loss: $\mathcal{L}_c = \|\mathbf{I} - \hat{\mathbf{I}}\|_2^2$, where $\mathbf{I}$ and $\hat{\mathbf{I}}$ denote the ground-truth and predicted RGB images, respectively.

Semantic models trained solely with the cross-entropy loss often yield predictions that exhibit poor spatial coherence, leading to noisy or irregular outputs within homogeneous regions. Therefore, we design a regional smoothness loss $\mathcal{L}_{rs}$ that enforces the consistency of predicted semantic class distributions between neighboring pixels. Formally, $\mathcal{L}_{rs}$ is computed as:

$$\begin{aligned}\mathcal{L}_{rs} = &\sum_l \sum_{(i,j)\in\mathcal{R}_l} \mathbf{1}[(i+1,j)\in\mathcal{R}_l]\|\hat{\mathbf{S}}_{i+1,j}^l - \hat{\mathbf{S}}_{i,j}^l\|_1 \\ &+ \sum_l \sum_{(i,j)\in\mathcal{R}_l} \mathbf{1}[(i,j+1)\in\mathcal{R}_l]\|\hat{\mathbf{S}}_{i,j+1}^l - \hat{\mathbf{S}}_{i,j}^l\|_1, \quad (6)\end{aligned}$$

where $\mathcal{R}_l$ denotes the pixel regions belonging to the ground-truth semantic class l, and $\hat{\mathbf{S}}_{i,j}^l$ denotes the predicted probability distribution at pixel (i,j) for class l. This regional smoothness loss promotes semantic local consistency while preserving sharp inter-class boundaries.

In sum, the final training objective for **SemGS** is:

$$\mathcal{L} = \lambda_{sem}\mathcal{L}_{sem} + \lambda_c\mathcal{L}_c + \lambda_{rs}\mathcal{L}_{rs}, \quad (7)$$

where λ_{sem}, λ_c, and λ_{rs} are balancing weights.

IV. EXPERIMENTS

A. Experimental Setup

1) Implementation Details: Our **SemGS** is implemented in PyTorch with a CUDA-based Gaussian rasterizer. In dual-branch feature extractor, both color and semantic branches share a CNN backbone consisting of 6 CNN-residual blocks, while the semantic branch includes 2 extra residual blocks

TABLE I

QUANTITATIVE COMPARISONS ON SCANNET DATASET (AVERAGED OVER 10 SCENES). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	2 Input Views ($N = 2$)				3 Input Views ($N = 3$)				4 Input Views ($N = 4$)			
	mIoU↑	acc.↑	class acc.↑	FPS↑	mIoU↑	acc.↑	class acc.↑	FPS↑	mIoU↑	acc.↑	class acc.↑	FPS↑
S-Ray [31]	0.538	0.772	0.619	0.52	0.563	0.778	0.632	0.41	0.604	0.802	0.673	0.34
GSNeRF [32]	0.529	0.751	0.616	0.25	0.550	0.752	0.648	0.23	0.587	0.796	0.671	0.19
SemGS (Ours)	**0.754**	**0.912**	**0.803**	**8.49**	**0.757**	**0.908**	**0.811**	**7.35**	**0.765**	**0.919**	**0.815**	**6.10**

TABLE II

QUANTITATIVE COMPARISONS ON SCANNET++ DATASET (AVERAGED OVER 10 SCENES). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	2 Input Views ($N = 2$)				3 Input Views ($N = 3$)				4 Input Views ($N = 4$)			
	mIoU↑	acc.↑	class acc.↑	FPS↑	mIoU↑	acc.↑	class acc.↑	FPS↑	mIoU↑	acc.↑	class acc.↑	FPS↑
S-Ray [31]	0.411	0.775	0.502	0.60	0.441	0.808	0.523	0.46	0.467	0.804	0.558	0.38
GSNeRF [32]	0.421	0.760	0.503	0.29	0.450	0.794	0.538	0.26	0.479	0.813	0.562	0.21
SemGS (Ours)	**0.625**	**0.903**	**0.689**	**9.24**	**0.671**	**0.924**	**0.734**	**7.68**	**0.676**	**0.927**	**0.739**	**6.29**

for semantic-specific refinement. The color branch adopts a Swin Transformer with 6 Transformer blocks (each containing self-attention and cross-attention), whereas the semantic branch adopts 3 Transformer blocks, with feature dimension d set to 128. The cost volume is refined using a 2D U-Net and regressed into depth maps through a 2-layer CNN. We sample $L = 128$ depth candidates. Branch-specific attributes of color and semantic Gaussians are predicted by separate 2-layer CNNs. We adopt the Adam optimizer with learning rate of 2×10^{-4}. The loss weights are set as $\lambda_{sem} = 0.1$, $\lambda_c = 1.0$, and $\lambda_{rs} = 0.001$. Training is conducted on 4 A100 GPUs with batch size 2 per GPU for 300k iterations.

2) Baselines & Datasets: Notably, only a few prior works have explored generalizable semantic field reconstruction and semantic novel view synthesis from sparse input images. We compare our **SemGS** with two state-of-the-art methods in this domain: S-Ray [31] and GSNeRF [32]. All methods are tested under sparse input settings with $N = 2, 3, 4$ views.

For quantitative evaluation, we adopt ScanNet [15] and ScanNet++ [16] datasets, both of which provide multi-view images, camera poses, and ground-truth semantic annotations. On ScanNet, all methods are trained with 1,000 scenes and tested on 10 held-out scenes, while on ScanNet++ we use 896 training scenes and 10 testing scenes. To further assess the generalization ability, we directly evaluate models trained on ScanNet on unseen domains without finetuning. Specifically, we use Replica [17] which contains synthetic scenes and real-world sequences collected by a mobile robot, and conduct qualitative comparisons across all models.

3) Evaluation Metrics: Following previous works [31], [32], we report three semantic metrics: mean Intersection-over-Union (*mIoU*), total pixel accuracy (*acc.*), and average per-class pixel accuracy (*class acc.*). These metrics are computed over the 20 classes defined in the ScanNet benchmark, and comprehensively assess both the overall and class-wise segmentation performance. Additionally, we adopt rendering frames per second (*FPS*) to evaluate the inference speed and computational efficiency of different methods.

B. Quantitative Evaluation

We quantitatively evaluate our method against S-Ray [31] and GSNeRF [32] on ScanNet and ScanNet++ datasets under different numbers of input views. The comparison results are summarized in Tables I and II.

On both datasets, **SemGS** consistently outperforms prior works across all metrics and demonstrates stronger semantic reasoning ability. With only 2 input views, **SemGS** already achieves a significant improvement in *mIoU*, showing the effectiveness of our framework under extremely sparse inputs. As the number of views increases, the advantage of **SemGS** remains pronounced. In terms of pixel accuracy (*acc.*) and class accuracy (*class acc.*), our method also exhibits significant gains, indicating that it produces more reliable semantic maps with enhanced inter-class separability. Overall, these improvements show that our **SemGS** enables more accurate and robust semantic field reconstruction.

In addition to the semantic accuracy, **SemGS** also exhibits significantly superior inference efficiency. By leveraging the feed-forward architecture and efficient Gaussian rasterization, our method achieves more than an order of magnitude speedup (see *FPS* metric in Table I and Table II) over existing works. This makes **SemGS** more practical for real-time robotic applications.

C. Qualitative Evaluation

We present more qualitative comparisons in Fig. 4. Overall, our **SemGS** outperforms existing methods and generates semantic maps that are closer to the ground-truth annotations. Specifically, S-Ray often suffers from segmentation errors and blurred boundaries. For example, in the second row of Fig. 4, edges of the table are poorly delineated and parts of areas are misclassified as chairs. GSNeRF yields relatively better predictions but tends to produce noisy or fragmented regions, particularly in challenging cases with fine-grained structures, such as the sinks in the third row.

In contrast, our method can produce sharper object boundaries, fewer misclassified regions, and more spatially coherent segmentations. For instance, the table and chairs in the

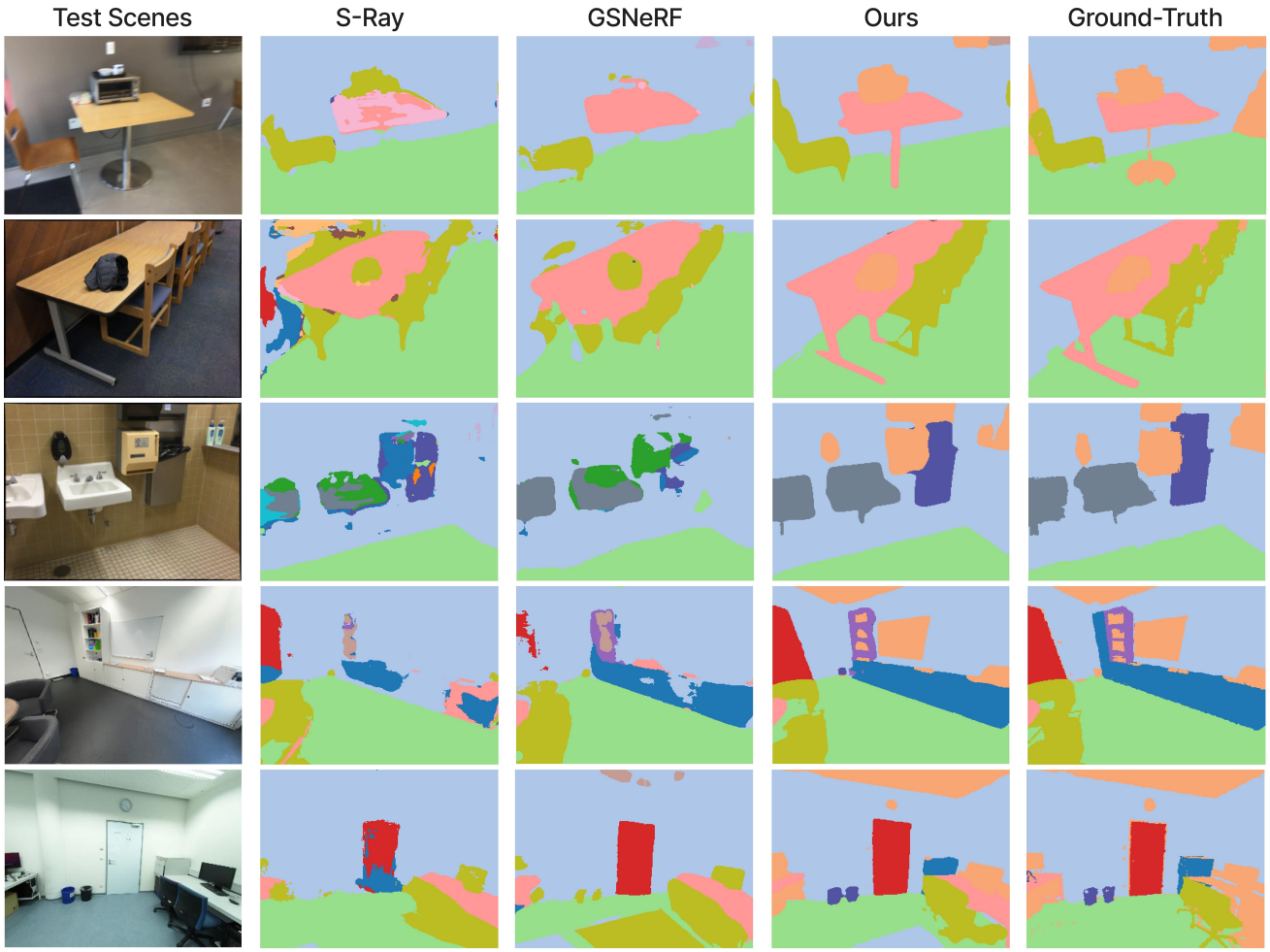

Fig. 4. Qualitative comparisons on ScanNet and ScanNet++ datasets. We demonstrate novel-view semantic rendering results of different methods. Compared to existing works, our method produces sharper object boundaries, fewer misclassified regions, and more spatially consistent segmentations.

second row are properly separated, and sinks in the third row are distinguished without artifacts. Moreover, in cluttered indoor environments (fourth and fifth rows), **SemGS** maintains semantic consistency across large planar regions (*e.g.,* walls and floors), while still capturing small objects (*e.g.,* cabinets and trash cans) with high fidelity.

Qualitative results show that our proposed model ensures global semantic consistency while preserving local details. Consequently, **SemGS** produces more accurate and visually appealing semantic novel views than previous works.

D. Ablation Study

To validate the effectiveness of each proposed component in **SemGS**, we conduct ablation studies on ScanNet dataset and the results are reported in Table III.

Adopting shared CNN layers (*Sh-Layer*) in feature extractor enhances information exchange between dual branches, yielding noticeable gains in *mIoU* and semantic accuracy. Replacing vanilla Transformer Blocks with Swin Transformer [13] Blocks (*Sw-Trans*) further improves model performance, showing that hierarchical window-based attention is more effective for capturing scene structures. The proposed camera pose injection mechanism (*C-Inj*) explicitly encodes multi-view camera geometry, which results in better semantic predictions and consistent improvements. Finally, applying the regional smoothness loss (*Sm-Loss*) regularizes the semantic spatial coherence and reduces local noise. Therefore, it primarily improves pixel accuracy (*acc.*), as this metric is dominated by large-area categories, where the proposed loss can effectively suppress scattered noise and enhance semantic consistency (see Fig. 6).

Overall, these results validate the individual contributions of each proposed component, and the full model achieves the best performance across all evaluated metrics.

E. Generalizability

To further assess the generalizability of each method, we directly apply models trained on ScanNet to unseen domains, including Replica [17] synthetic scenes and real-world robot-captured scenes. The results are shown in Fig. 5.

In synthetic scenes, our method accurately distinguishes fine-grained structures and preserves sharp boundaries. S-Ray and GSNeRF often produce incomplete or fragmented regions and frequently miss small objects. In real-world

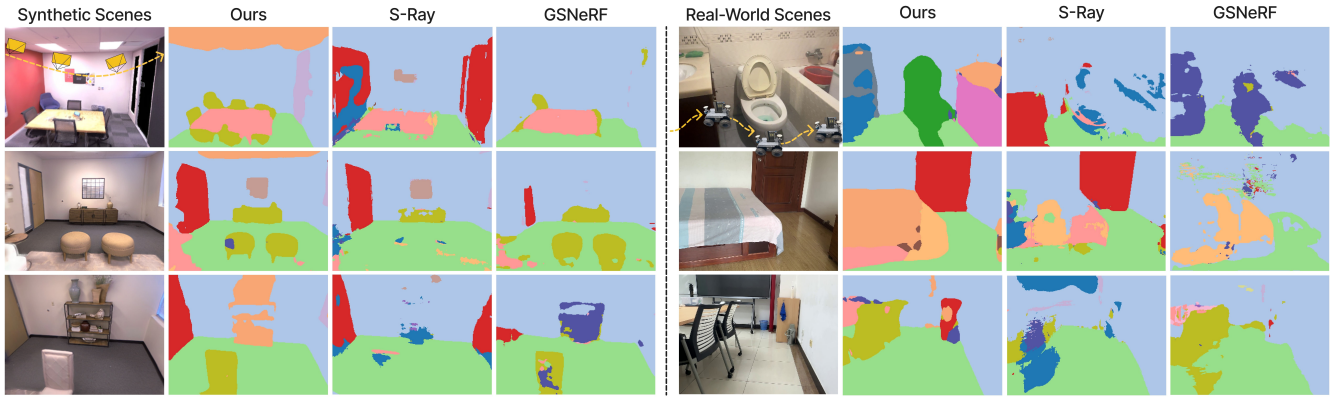

Fig. 5. Generalization ability on unseen domains. Different models trained on ScanNet are directly evaluated on Replica synthetic scenes (left) and real-world robot-captured scenes (right). Our **SemGS** generates more accurate and complete semantic maps than prior works.

TABLE III

ABLATION STUDY ON SCANNET DATASET. THE FULL MODEL, WITH ALL COMPONENTS COMBINED, ACHIEVES THE BEST PERFORMANCE.

Sh-Layer	*Sw-Trans*	*C-Inj*	*Sm-Loss*	mIoU↑	acc.↑	class acc.↑
✗	✗	✗	✗	0.681	0.856	0.751
✓	✗	✗	✗	0.718	0.873	0.776
✓	✓	✗	✗	0.744	0.898	0.795
✓	✓	✓	✗	0.761	0.910	0.812
✓	✓	✓	✓	**0.765**	**0.919**	**0.815**

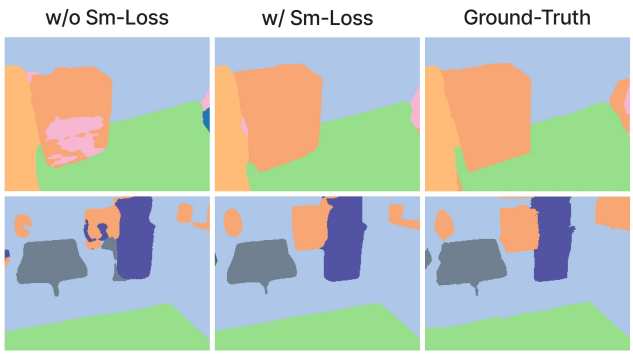

Fig. 6. Visualization of the effect of regional smoothness loss (*Sm-Loss*). The proposed loss suppresses noise and enhances semantic coherence.

robot-captured scenes, S-Ray and GSNeRF suffer from severe noise and misclassify large portions of the scenes, whereas **SemGS** produces semantic maps that align well with the underlying scene geometry and object layouts.

The superior generalizability of **SemGS** stems from three key factors. First, the dual-branch feature extractor enables semantic reasoning to leverage low-level structural and textural cues from color features, which remain reliable across domains. Second, the proposed camera-aware attention mechanism explicitly encodes the inter-view camera relations, thereby enhancing the robustness of 3D perception and scene understanding. Third, by sharing geometric attributes between semantic and color Gaussians, we implicitly enforce the geometric consistency. Together, these designs enable our **SemGS** to achieve superior performance when deployed in previously unseen environments.

V. CONCLUSIONS

In this paper, we present **SemGS**, a feed-forward framework that constructs generalizable semantic fields from sparse input images. **SemGS** employs a dual-branch feature extractor with shared low-level layers, allowing semantic features to leverage the structural and textural cues of color appearance. In addition, we propose to inject camera poses into the Transformer attention mechanism, which enhances the semantic reasoning capability. To further enforce the coherence of semantic predictions, we also design a regional smoothness loss. Experiments on benchmark datasets show that **SemGS** not only achieves superior semantic rendering accuracy, but also offers significant improvements in inference speed and generalization ability to unseen environments. These advantages underscore the potential of our framework for real-world robotic applications.

Despite its strong performance, there still remains room to improve **SemGS**. First, our framework relies on known cameras (or estimated by off-the-shelf tools [35], [36]). Inaccuracies in camera poses can propagate into semantic predictions. A future direction is to jointly optimize the camera parameters within the framework, improving robustness to imperfect camera calibration in an end-to-end manner. Second, although **SemGS** exhibits improved cross-domain generalization, its performance may still degrade when facing drastic domain gaps such as outdoor scenes with highly dynamic objects. Scaling up training with more diverse datasets, combined with strong 2D foundation model features [10], [28], could further enhance robustness in such challenging scenarios. We leave addressing these challenges as promising directions for future work.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China (62332019, 62461160309).

REFERENCES

[1] I. Kostavelis, K. Charalampous, A. Gasteratos, and J. K. Tsotsos, "Robot navigation via spatial and temporal coherent semantic maps," *Engineering Applications of Artificial Intelligence*, vol. 48, pp. 173–187, 2016.

[2] A. Cosgun and H. I. Christensen, "Context-aware robot navigation using interactively built semantic maps," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 254–276, 2018.

[3] M. Hua, Y. Nan, and S. Lian, "Small obstacle avoidance based on rgb-d semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[4] T. S. Veiga, P. Miraldo, R. Ventura, and P. U. Lima, "Efficient object search for mobile robots in dynamic environments: Semantic map as an input for the decision maker," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2745–2750.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, vol. 12346, 2020, pp. 405–421.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

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

[8] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19 729–19 739.

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

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[11] B. Li, K. Q. Weinberger, S. J. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[12] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–386.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[14] R. Li, B. Yi, J. Liu, H. Gao, Y. Ma, and A. Kanazawa, "Cameras as relative positional encoding," *arXiv preprint arXiv:2507.10496*, 2025.

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

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

[17] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[18] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4578–4587.

[19] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 124–14 133.

[20] M. M. Johari, Y. Lepoittevin, and F. Fleuret, "Geonerf: Generalizing nerf with geometry priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 365–18 375.

[21] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19 457–19 467.

[22] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. Reid, S. Gould, *et al.*, "Semantics for robotic mapping, perception and interaction: A survey," *Foundations and Trends® in Robotics*, vol. 8, no. 1–2, pp. 1–224, 2020.

[23] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4558–4567.

[24] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, "Graph attention convolution for point cloud semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 296–10 305.

[25] A. Kundu, X. Yin, A. Fathi, D. Ross, B. Brewington, T. Funkhouser, and C. Pantofaru, "Virtual multi-view fusion for 3d semantic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 518–535.

[26] Z. Hu, X. Bai, J. Shang, R. Zhang, J. Dong, X. Wang, G. Sun, H. Fu, and C.-L. Tai, "Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 488–15 498.

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

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

[29] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 676–21 685.

[30] J. Guo, X. Ma, Y. Fan, H. Liu, and Q. Li, "Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting," *arXiv preprint arXiv:2403.15624*, 2024.

[31] F. Liu, C. Zhang, Y. Zheng, and Y. Duan, "Semantic ray: Learning a generalizable semantic field with cross-reprojection attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 386–17 396.

[32] Z.-T. Chou, S.-Y. Huang, I. Liu, Y.-C. F. Wang, *et al.*, "Gsnerf: Generalizable semantic neural radiance fields with enhanced 3d scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 806–20 815.

[33] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[34] T. Miyato, B. Jaeger, M. Welling, and A. Geiger, "GTA: A geometry-aware attention mechanism for multi-view transformers," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[35] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 501–518.

[36] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.