The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Unifying Street Scene Point Cloud Semantic Segmentation with Deformable Mesh-based Neural Representation

Yuzhou Zhou[1]

[1] Department of Computer Science, University of Oxford, Oxford, OX1 2JD, UK - yuzhou.zhou@cs.ox.ac.uk

Keywords: Point cloud, semantic segmentation, scene understanding, neural representation, street scene reconstruction.

Abstract

Accurate semantic segmentation of urban point clouds is important for applications such as urban planning and autonomous driving. Recently, neural scene representations have been extended to merge semantic information across modalities and spatial dimensions. While 3D Gaussian Splatting (3DGS) enables efficient and high-quality reconstruction, its semantic understanding performance in street scenes is influenced by trajectory-constrained viewpoints, where Gaussian densification introduces occlusions and semantic ambiguity. This paper explores the use of NeRF-based neural representation for street scene point cloud semantic segmentation. Specifically, deformable neural mesh primitives (DNMPs) are used to compactly represent spatial geometry and simplify ray sampling. Then, neural fields including density, RGB, and semantics are constructed based on mesh vertex feature interpolation and MLPs. The sampled neural field values are accumulated via ray rendering and supervised using original images and corresponding semantic label maps generated by pre-trained models. Point cloud semantics are then predicted by interpolating neighboring samples within the learned field. The method is validated on the KITTI-360 and Waymo datasets. Results show that the proposed approach achieves improved semantic segmentation performance while maintaining competitive rendering quality, and supports both novel view synthesis and semantic rendering.

1. Introduction

Accurate point cloud semantic segmentation enables reliable scene understanding for autonomous driving and urban mapping applications (Han et al., 2024). Conventional methods largely rely on carefully designed feature extractors and require large-scale 3D annotations. Despite notable progress, these methods are inherently limited by the high cost of 3D labeling and domain gaps across datasets, which restrict their generalization ability. Recently, neural scene representations such as Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) have revolutionized 3D reconstruction. Neural scene representation offers a promising approach to bridge 2D supervision and 3D properties, which enables lifting 2D semantic clues to 3D space. Building upon NeRF and 3DGS, subsequent works have extended to semantic neural modeling, where semantic fields are jointly learned with geometry and appearance, enabling semantic information to be transferred between 2D images and 3D space within a unified framework (Zhi et al., 2021, Siddiqui et al., 2023, Zhai et al., 2025).

Although 3DGS has recently become the dominant neural representation approach due to its explicit and efficient formulation, it is primarily designed for 360° object-centric scenes with dense viewpoint coverage. In large-scale street scenes, however, data is typically captured from ego-vehicle platforms, where viewpoints are constrained along driving trajectories and exhibit strong directional bias. This leads to highly uneven spatial coverage. Under such conditions, the optimization of 3DGS becomes suboptimal, particularly during the Gaussian densification process. As new Gaussians are iteratively introduced, overlaps and occlusions between Gaussians tend to accumulate, resulting in ambiguous geometry and entangled semantics. This issue is especially detrimental for semantic modeling, where inter-Gaussian occlusion may introduce semantic inconsistencies and bias, degrading point cloud segmentation performance in complex urban environments (Yan et al., 2024).

Motivated by these limitations, we revisit NeRF-based representations in the era of 3DGS. Instead of relying on fully implicit volumetric sampling or purely explicit Gaussian primitives, we seek a hybrid formulation that combines the structural constraints of surface-based representations with the efficiency of explicit modeling. In particular, we build upon Deformable Neural Mesh Primitives (DNMP) (Lu et al., 2023a), which provide a compact and geometry-aware representation by encoding local surfaces as deformable mesh elements. Compared with volumetric NeRF sampling, DNMP enables surface-constrained sampling through ray–mesh intersections, significantly improving efficiency. Based on this representation, a neural semantic field is constructed by supervising the rendered semantic map with pseudo semantic labels predicted using pre-trained image semantic segmentation networks. Point cloud semantic segmentation is performed via point-wise interpolation based on the optimized mesh vertex feature fields.

Specifically, input point clouds are first voxelized, and each voxel is represented by a deformable mesh primitive parameterized by a latent code. The geometry is optimized via depth supervision, after which learnable feature embeddings are assigned to the mesh vertices to encode spatial information. During rendering, ray–mesh intersections are computed, and vertex features are interpolated and accumulated along rays to predict RGB values, semantic distributions, and opacity values. The rendered outputs are supervised by RGB images and pseudo semantic labels generated by pre-trained image semantic segmentation networks. Finally, point-wise semantic labels are inferred through spatial interpolation of the learned neural field.

This formulation unifies neural scene representation and point cloud semantic segmentation in a single framework without requiring 3D annotations. By leveraging the explicit yet structured representation of DNMP, our approach achieves efficient

sampling, robust geometry modeling, and improved semantic consistency under limited-viewpoint conditions. Experiments on the KITTI-360 (Liao et al., 2022) and Waymo (Sun et al., 2020) datasets demonstrate that the proposed method achieves competitive performance in both point cloud semantic segmentation and 3D reconstruction, highlighting its potential for scalable urban scene understanding.

2. Related Work

2.1 Point Cloud Semantic Segmentation

Extensive research has been conducted on point cloud semantic segmentation. Early representative works include networks based on voxels (Graham et al., 2018) or point-wise features (Qi et al., 2017b, Thomas et al., 2019). RandLA-Net significantly improves processing efficiency for large-scale point clouds using feature propagation based on random sampling (Hu et al., 2020). These methods improve local and contextual feature encoding and achieve competitive performance. However, they are trained on fully annotated point cloud datasets. To reduce the reliance on 3D annotations, recent research has focused on exploiting weakly supervised point cloud segmentation through contrastive learning (Jiang et al., 2021, Li et al., 2022) or feature interpolation (Hu et al., 2022). Cross-modality supervision and multi-modal integration have also been explored by some pioneering works, which mainly use image features (Yan et al., 2022, Sautier et al., 2022) or labels (Genova et al., 2021) as guidance to avoid reliance on annotations.

2.2 NeRF and 3DGS in Street Scenes

NeRF represents a scene using an MLP that maps spatial coordinates and viewing directions to densities and radiances (Mildenhall et al., 2021). In its follow-up works, the spatial partitioning and sampling for ray rendering of NeRF have been greatly improved for efficiency and accuracy (Barron et al., 2021, Barron et al., 2022, Liu et al., 2020, Chen et al., 2022). However, these methods densely sample along rays during rendering, which causes low computational efficiency. Point clouds indicate regions of high spatial density and can serve as sampling guidance (Lu et al., 2023a, Deng et al., 2022, Lu et al., 2023b). To scale NeRF to urban scenes, some works have proposed to partition the scene into sub-modules (Tancik et al., 2022, Turki et al., 2023, Turki et al., 2022). When building the urban neural radiance fields with the assistance of point clouds, a key challenge is to balance the simplicity of spatial sampling and the capacity to describe and store spatial features.

More recently, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has emerged as an efficient alternative to NeRF through an explicit representation of anisotropic Gaussians. Some works have extended 3DGS to urban street scenes. Representative approaches, such as StreetGaussians (Yan et al., 2024), aim to disentangle dynamic components from the static scene, enabling improved modeling of scene dynamics and flexible manipulation of moving objects (Huang et al., 2024, Cheng et al., 2024, Wu et al., 2023, Zhou et al., 2024a, Zhou et al., 2024c). These methods typically introduce additional modules for motion modeling or object-level representation, allowing dynamic elements such as vehicles to be separated and reconstructed independently from the static environment.

2.3 Semantic Modeling of Neural Representation

NeRF models use MLPs to regress position-based and view-dependent spatial properties under 2D supervisions. Inspired by neural radiance field modeling, other spatially varying fields including semantics can be constructed in a similar manner (Zhi et al., 2021, Fu et al., 2022, Liu et al., 2023). Semantic-NeRF firstly proposed to render semantic fields and applied it to various semantic tasks (Zhi et al., 2021). Then a series of works extend this to panoptic field construction by explicitly encoding object labels (Fu et al., 2022, Bing et al., 2023, Kundu et al., 2022, Siddiqui et al., 2023). These works investigate the label transfer across 2D and 3D spaces and show promising results. However, Panoptic-NeRF (Fu et al., 2022) still requires 3D coarse primitives as supervision.

Recent works further extend semantic and instance-level representations to the explicit 3DGS, enabling scene understanding and editing capabilities (Qiu et al., 2024, Ye et al., 2024, Zhou et al., 2024b, Wu et al., 2024, Gu et al., 2024, Qin et al., 2024). To reduce reliance on costly 3D annotations, many approaches leverage 2D supervision derived from pre-trained vision foundation models (e.g., DINO, SAM, CLIP) to guide semantic learning in 3D (Oquab et al., 2023, Ravi et al., 2025, Radford et al., 2021). This paradigm has also been explored in urban street scenes, where semantic information is mainly learned for static environments under image-based supervision (Yan et al., 2024).

3. Method

Given a sequence of posed RGB images in urban scenes, we adopt a pre-trained semantic segmentation model to predict pixel-wise semantic labels (Tao et al., 2020). We further use multi-view stereo or LiDAR points to generate depth maps. The point clouds are also used as input to the model. Based on these inputs, our goal is to infer a semantic distribution $k \in \mathcal{K}$ for any 3D point $\boldsymbol{x} \in \mathbb{R}^3$ within the scene.

Unlike previous methods that construct semantic fields directly from coordinates (Zhi et al., 2021) or voxel space (Siddiqui et al., 2023), We employ deformable neural mesh primitives (DNMPs) proposed in (Lu et al., 2023a) to interpolate the neural fields. An overview of the framework is shown in Figure 1. This method simplifies the spatial sampling by supervising only the accumulated samples at ray-mesh intersections.

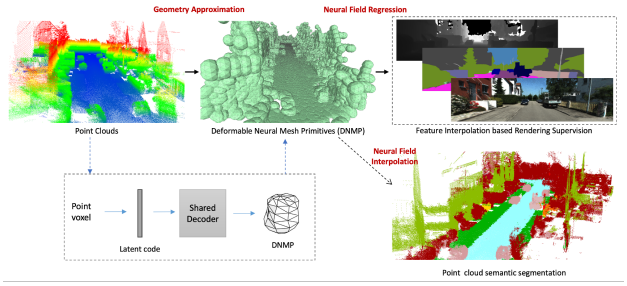


Figure 1. DNMP-based semantic scene representation overview.

3.1 DNMP-based Scene Representation

A DNMP is a watertight mesh composed of tens of triangular faces and vertices, and is proposed as a proxy for point voxels in neural scene representation (Lu et al., 2023a). The shape of a DNMP is defined by the coordinates of vertices $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^3 \mid i = 1, \ldots, m\}$, where m denotes the number

of vertices in a mesh primitive. To build a compact learnable representation of the vertex coordinates, an auto-encoder consisting of a shape encoder and a latent code decoder is trained, such that a single latent code represents the geometry of the mesh primitive. The auto-encoder is built upon PointNet (Qi et al., 2017a) and trained by comparing the 3D coordinate Chamfer Distance between the input and output point samples on mesh faces (Lu et al., 2023a). The shape decoder $G(\boldsymbol{z})$ is defined in Eq. 1, where $\boldsymbol{z} \in \mathbb{R}^{\mu}$ is a μ-dimensional normalized ($\|\boldsymbol{z}\|_2 = 1$) latent code.

$$G(\boldsymbol{z}): \ \boldsymbol{z} \in \mathbb{R}^{\mu} \mapsto \mathcal{X} = \left\{ \boldsymbol{x}_i \in \mathbb{R}^3 \mid i = 1, \ldots, m \right\}. \quad (1)$$

The input point clouds from multi-view stereo (MVS) or laser scanning are firstly voxelized, and a latent code that decodes into a nearly spherical mesh primitive is initialized and assigned to each voxel. We use $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{M}$ to denote the latent codes of all primitives, where M is the number of voxels. Given $\mathcal{Z}$ and the camera parameters, the depth map is rendered as $D(\mathcal{Z})$, where the depth value for each ray is defined by the distance from the camera to the first ray-mesh intersection. Latent codes $\mathcal{Z}$ are optimized to regress the original depth map D rendered from the input point clouds under the supervision of the geometry approximation loss:

$$L_{geometry} = \|\hat{D}(\mathcal{Z}) - D\|_1. \quad (2)$$

In this way, the shapes of the mesh primitives are trained to describe the geometry represented by the original point clouds.

A learnable feature embedding vector $\boldsymbol{f}_i \in \mathbb{R}^{\lambda}$ $(i = 1, \ldots, m)$ is assigned to each vertex of the mesh primitive, which encodes spatially varying properties such as radiance and semantics. Therefore, each mesh primitive is represented as $\boldsymbol{V} = \{\boldsymbol{v}_i = (\boldsymbol{x}_i, \boldsymbol{f}_i) \mid i = 1, \ldots, m\}$, where $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{m}$ is decoded from the learnable latent code $\boldsymbol{z}_i$. In subsequent neural field modeling, the vertex coordinates $\mathcal{X}$ of each primitive are fixed, and the learnable features $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^{m}$ are optimized as the parameters of the neural fields.

3.2 Semantic Neural Field Modeling and Point Cloud Segmentation

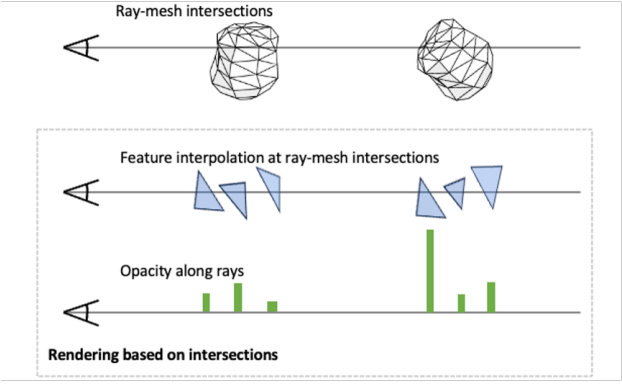


Figure 2. Neural field rendering based on DNMPs. For each viewing direction, ray–mesh intersections with neural mesh primitives are computed. Vertex feature embeddings are interpolated at these intersections and accumulated along rays according to predicted opacities to render different fields.

To build the 3D neural fields under the supervision of input images and semantic maps, the mapping is defined from each 3D point $\boldsymbol{x} \in \mathbb{R}^3$ and viewing direction $d \in \mathbb{S}^2$ to an opacity $\sigma \in \mathbb{R}^+$, an RGB color $\boldsymbol{c} \in \mathbb{R}^3$, and a semantic probability distribution $\boldsymbol{k}$ over the class list $\mathcal{K}$. The opacity σ is the probability that the ray terminates here (Attal et al., 2022). Then for each ray, the predicted pixel values $\hat{\boldsymbol{c}}_p \in \mathbb{R}^3$ and $\hat{\boldsymbol{k}}_p$ over the class set $\mathcal{K}$ are compared with input image pixels. In this process, two aspects are important to consider: 1) how samples are selected and accumulated along rays; 2) how spatial samples are mapped to neural field outputs. These aspects are introduced in the following sections. Figure 2 illustrates the rendering process.

Ray Sampling based on Intersections. For ray sampling and accumulation, this method leverages the property of neural mesh primitives. Since the scene represented by DNMP has been built, indicating the locations of object surfaces, we only sample at the intersections between rays and mesh primitives denoted as $\boldsymbol{t}_j = (\boldsymbol{x}_j, \boldsymbol{f}_j)$, where $\boldsymbol{x}_j$ and $\boldsymbol{f}_j$ denote the spatial coordinates and feature embedding of the intersection. Each sample $\boldsymbol{t}_j$ is mapped to an opacity value $\sigma_j \in \mathbb{R}^+$. Neural field values at these samples (i.e., ray–mesh intersections) are then accumulated along each ray.

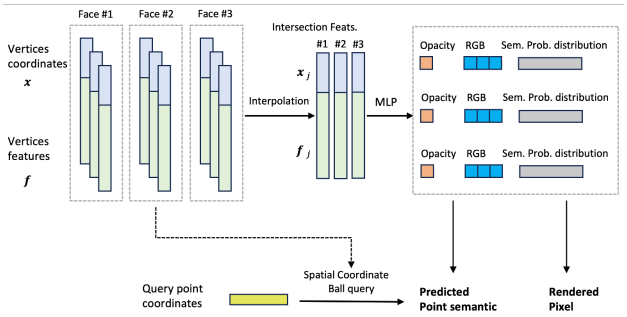


Figure 3. Neural field interpolation and rendering. First, vertex feature embeddings are interpolated at each ray–mesh intersection. Then, along each ray, the intersection coordinates $\boldsymbol{x}_j$ and features $\boldsymbol{f}_j$ are fed into an MLP to predict opacity, RGB values, and semantic distributions. These values are accumulated along the ray based on the predicted opacities. Point cloud semantic labels are inferred via spatial interpolation of the predicted neural field outputs using a ball query.

Neural Field Modeling. Given a spatial sample $\boldsymbol{t}_j$ on a triangular mesh face $\boldsymbol{J}_{\text{inter}} = (\boldsymbol{v}_\alpha, \boldsymbol{v}_\beta, \boldsymbol{v}_\gamma)$, where $\boldsymbol{v} = (\boldsymbol{x}, \boldsymbol{f})$ denotes a face vertex, the feature embedding of the sample $\boldsymbol{f}_j$ is interpolated based on the barycentric coordinates of the intersection on the face $\boldsymbol{J}_{\text{inter}}$. This interpolation is described as:

$$P(\boldsymbol{x}_j): \boldsymbol{x}_j \in \mathbb{R}^3; \boldsymbol{J}_{\text{inter}} = (\boldsymbol{v}_\alpha, \boldsymbol{v}_\beta, \boldsymbol{v}_\gamma) \mapsto \boldsymbol{f}_j. \quad (3)$$

Then we adopt the Semantic-NeRF network structure (Zhi et al., 2021) to model the neural representation from spatial coordinates and interpolated features to neural field values using an MLP, and the input is formed by concatenating the positional encoding (PE) (Mildenhall et al., 2021) of position $\boldsymbol{x}_j$ and the feature embedding $\boldsymbol{f}_j$. Unlike the original Semantic-NeRF, a softmax layer is added before semantic output produce a probability distribution instead of logits (Siddiqui et al., 2023) for accuracy. This network is described as

$$(\sigma_j, \boldsymbol{c}_j, \boldsymbol{k}_j) = F_\theta(\boldsymbol{x}_j, \boldsymbol{f}_j, d_j) \quad (4)$$

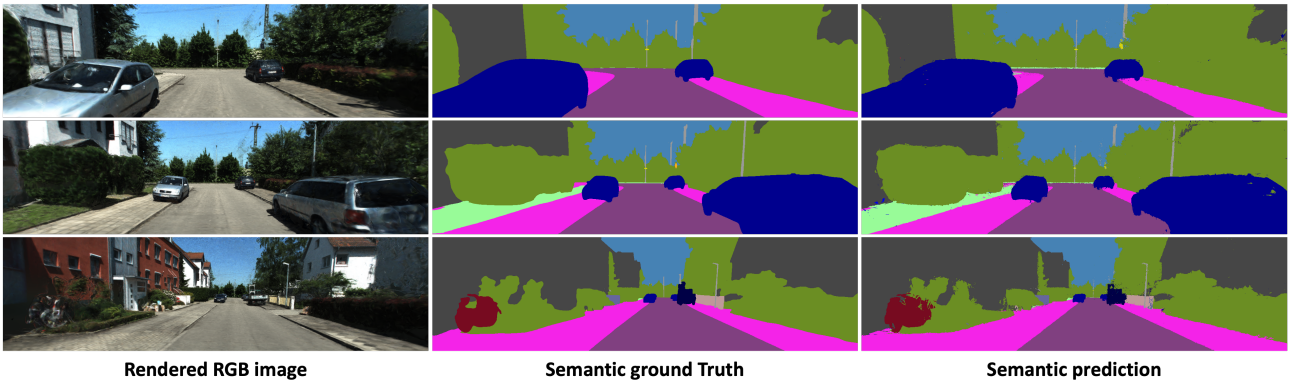


Figure 4. Qualitative results of novel view synthesis and semantic rendering on KITTI-360. From left to right: rendered RGB images, semantic ground truth, and predicted semantic maps. We produce consistent semantic predictions across viewpoints while maintaining high-quality appearance rendering.

Pixel Value Rendering. After predicting neural field values at samples along a ray r, the RGB values and the semantic distribution are computed as follows (Figure 3):

$$\hat{\boldsymbol{c}}_r = \sum_{j=1}^{L} T_j \sigma_j \boldsymbol{c}_j, \quad \hat{\boldsymbol{k}}_r = \sum_{j=1}^{L} T_j \sigma_j \boldsymbol{k}_j, \tag{5}$$

where $T_j = \prod_{i=1}^{j-1} (1 - \sigma_i)$ denotes the accumulated transmittance along the ray up to the j-th sample and L is the number of samples along the ray r.

Loss Function. For a batch of rays $\mathcal{R}$, the appearance loss L_{rgb} is computed as:

$$L_{rgb}(\mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left\| \boldsymbol{c}_r - \hat{\boldsymbol{c}}_r \right\|^2, \tag{6}$$

where $\boldsymbol{c}_r$ is the ground truth value. The semantic loss L_{sem} is modeled as a cross-entropy loss:

$$L_{\text{sem}}(\mathcal{R}) = -\frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{K}} \boldsymbol{k}_r(l) \log \hat{\boldsymbol{k}}_r(l), \tag{7}$$

where l indexes semantic classes, and $\boldsymbol{k}_r$ denotes the ground truth one-hot semantic label. The training loss L is:

$$L = L_{rgb} + w L_{\text{sem}}, \tag{8}$$

where w is a weight empirically set to 0.1 to balance the loss magnitude.

Point Semantic Label Prediction. After training F_θ, given any point position $\boldsymbol{x}$, the opacity σ and semantic distribution $\boldsymbol{k}$ can be predicted. Also, the vertex embeddings, which are optimized during training, encode spatial features. Therefore, to predict the semantic label of any 3D coordinate $\boldsymbol{x} \in \mathbb{R}^3$, we adopt the ball-query-based interpolation (Figure 3). Firstly, neighboring DNMP vertices $\boldsymbol{V}_b = \{\boldsymbol{v}_i = (\boldsymbol{x}_i, \boldsymbol{f}_i) \mid i = 1, \ldots, B\}$ whose features $\boldsymbol{f}_i$ have been adjusted during training are queried using a spatial ball query centered at $\boldsymbol{x}$, where B is the number of neighboring points. If $B = 0$, the point is assigned an undefined label. Then, we use both opacity and inverse distance weights to interpolate the semantic label of the point. For the given $\boldsymbol{V}_b$, we use F_θ to predict $\{\sigma_i, \boldsymbol{k}_i\}_{i=1}^{B}$. The semantic label is computed as:

$$\text{sem}(\boldsymbol{x}) = \underset{l \in \mathcal{K}}{\text{argmax}} \sum_i \sigma_i \frac{w_i}{\sum w_i} \boldsymbol{k}_i(l), \tag{9}$$

where $w_i = \frac{1}{\|\boldsymbol{x}_i - \boldsymbol{x}\|}$ denotes the inverse distance weight.

4. Experiments and Discussion

4.1 Implementation Details

The proposed method is validated and evaluated on the KITTI-360 (Liao et al., 2022) and Waymo (Sun et al., 2020) datasets. We use two road segments from each dataset. We perform a series of experiments including point cloud segmentation, novel view image synthesis, and semantic map synthesis. COLMAP (Schönberger and Frahm, 2016, Schönberger et al., 2016) is used to reconstruct point clouds from input images for geometric approximation. The pre-trained image semantic segmentation network (Tao et al., 2020) is used for generating semantic supervision. Every second image in the KITTI-360 image sequence is used for testing. For the Waymo dataset, every fifth image is used for testing. During training, each batch contains 4096 rays. The geometry approximation of mesh primitives is trained for 5k iterations, followed by 50k iterations for neural field modeling. The hierarchical voxelization strategy proposed in (Lu et al., 2023a) is adopted. The model is run on an NVIDIA RTX 3090 Ti GPU.

4.2 Experiment Results

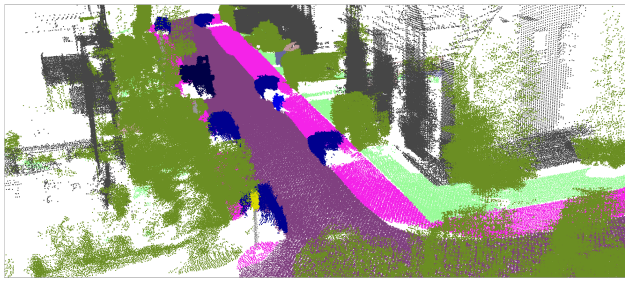
Qualitative Analysis. We qualitatively evaluate our method from three aspects: point cloud semantic segmentation, novel view synthesis, and semantic rendering. As shown in Figure 4, our approach produces high-quality semantic predictions for novel views that are well aligned with the ground truth across different viewpoints. The rendered RGB images and corresponding semantic maps demonstrate that the proposed neural representation effectively model geometry, appearance, and semantics in a consistent manner. Figure 5 presents the point cloud semantic segmentation results on laser scanning points. Our method successfully transfers semantic information from image space to 3D point clouds, producing consistent and spatially smooth predictions. Compared with the ground truth, most major semantic categories are correctly recognized, indicating that the learned neural field effectively captures scene-

Table 1. Quantitative comparison of street scene semantic segmentation with StreetGaussians (mIoU, %).
TL: traffic lights. TS: traffic signs. Veg.: vegetation. Terr.: terrain.

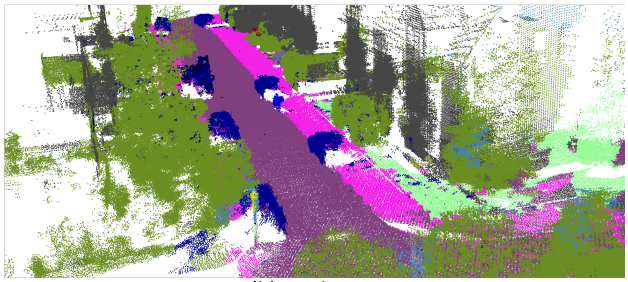
Method	mIoU	Road	Sidewalk	Building	Wall	Fence	Pole	TL	TS	Veg.	Terr.	Car	Truck	Others
Ours	**51.7**	87.9	63.8	73.3	36.3	35.3	26.1	30.2	39.0	85.6	21.3	78.0	67.6	27.6
StreetGaussians	43.5	75.3	56.0	45.6	27.9	40.5	19.5	12.3	25.2	77.1	26.3	72.4	57.8	30.2

Table 2. Quantitative comparison in terms of rendering (PSNR, SSIM) and novel view semantic segmentation (mIoU).

	KITTI-360			**Waymo**		
Method	PSNR ↑	SSIM ↑	mIoU ↑	PSNR ↑	SSIM ↑	mIoU ↑
Ours	23.01	**0.852**	**0.705**	29.31	0.916	**0.621**
StreetGaussians	**25.44**	0.840	0.688	**33.27**	**0.920**	0.609

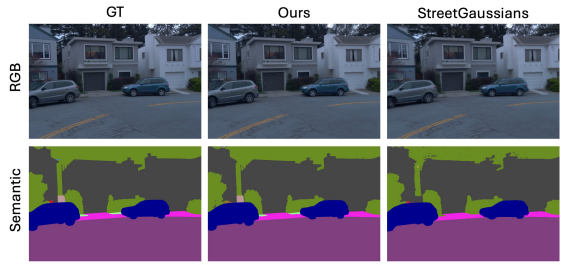


(a) Ground truth

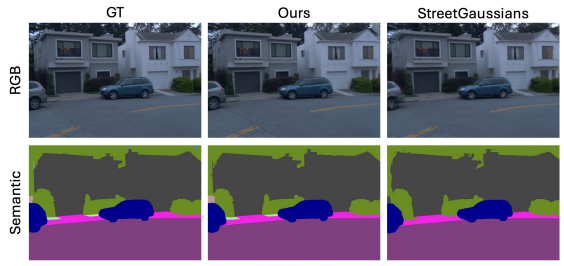


(b) Prediction

Figure 5. Qualitative results of point cloud semantic segmentation on the KITTI-360 dataset.



(a) View 1



(b) View 2

Figure 6. Qualitative comparison on the Waymo dataset. For each view, the first row shows RGB renderings, and the second row shows semantic predictions. From left to right: ground truth (GT), ours, and StreetGaussians.

level semantic structures without requiring any 3D annotations. In addition, we compare our results with StreetGaussians (Yan et al., 2024) in Figure 6. Both methods achieve realistic novel view synthesis, and our results exhibit more coherent semantic predictions. This improvement can be attributed to the surface-constrained sampling of DNMP, which alleviates ambiguity and inter-sample occlusion commonly that are observed in Gaussian-based representations under limited viewpoints. Overall, these results demonstrate that the we achieve competitive point cloud semantic segmentation performance, and support high-quality novel view synthesis and semantic rendering within a unified framework.

Quantitative analysis. As shown in Table 1, our method outperforms StreetGaussians in point cloud semantic segmentation, achieving higher mIoU (51.7% vs. 43.5%). We obtain point-wise predictions for StreetGaussians by applying region voting on Gaussian centers. However, the densification and inter-Gaussian occlusion in 3DGS introduce semantic ambiguity, which degrades segmentation performance. In Table 2, StreetGaussians achieves higher PSNR due to denser spatial sampling during reconstruction, while our results achieve comparable SSIM. More importantly, our approach consistently yields better mIoU on novel view semantic segmentation across datasets, demonstrating improved semantic consistency across views.

5. Conclusion

In this paper, we revisit NeRF-based representations for point cloud semantic segmentation and propose a DNMP-based semantic neural modeling method. Under limited viewpoints in street scenes, 3DGS suffers from densification-induced occlusion and semantic ambiguity. To address this, we exploit a semantic-aware neural representation based on deformable neural mesh primitives that combines surface-constrained sampling with explicit geometric structure. Our method achieves point cloud semantic segmentation without requiring 3D annotations, while also supporting high-quality novel view synthesis and semantic rendering. Experiments on KITTI-360 and Waymo demonstrate that our approach achieves superior semantic segmentation performance compared to StreetGaussians, while maintaining competitive performance in novel view synthesis and semantic rendering.

Acknowledgements

The author gratefully acknowledges the valuable suggestions from Professor Niki Trigoni and Professor Andrew Markham. This work is supported by the Clarendon Fund Scholarship and the China Oxford Scholarship Fund. The author gratefully acknowledges their support.

References

Attal, B., Huang, J.-B., Zollhöfer, M., Kopf, J., Kim, C., 2022. Learning neural light fields with ray-space embedding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19819–19829.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P. P., 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Bing, W., Chen, L., Yang, B., 2023. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *The Eleventh International Conference on Learning Representations*.

Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H., 2022. Tensorf: Tensorial radiance fields. *European Conference on Computer Vision*, Springer, 333–350.

Cheng, K., Long, X., Yang, K., Yao, Y., Yin, W., Ma, Y., Wang, W., Chen, X., 2024. Gaussianpro: 3d gaussian splatting with progressive propagation. *International Conference on Machine Learning*.

Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D., 2022. Depth-supervised nerf: Fewer views and faster training for free. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

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

Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., Funkhouser, T., 2021. Learning 3d semantic segmentation with only 2d image supervision. *2021 International Conference on 3D Vision (3DV)*, IEEE, 361–372.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3d semantic segmentation with submanifold sparse convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9224–9232.

Gu, Q., Lv, Z., Frost, D., Green, S., Straub, J., Sweeney, C., 2024. Egolifter: Open-world 3d segmentation for egocentric perception. *The European Conference on Computer Vision*, Springer, 382–400.

Han, X., Liu, C., Zhou, Y., Tan, K., Dong, Z., Yang, B., 2024. WHU-Urban3D: An urban scene LiDAR point cloud dataset for semantic instance segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209, 500–513.

Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A., 2022. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. *European Conference on Computer Vision*, Springer, 600–619.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11108–11117.

Huang, N., Wei, X., Zheng, W., An, P., Lu, M., Zhan, W., Tomizuka, M., Keutzer, K., Zhang, S., 2024. S^3Gaussian: Self-Supervised Street Gaussians for Autonomous Driving. *arXiv preprint arXiv:2405.20323*.

Jiang, L., Shi, S., Tian, Z., Lai, X., Liu, S., Fu, C.-W., Jia, J., 2021. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6423–6432.

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

Kundu, A., Genova, K., Yin, X., Fathi, A., Pantofaru, C., Guibas, L. J., Tagliasacchi, A., Dellaert, F., Funkhouser, T., 2022. Panoptic neural fields: A semantic object-aware neural scene representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12871–12881.

Li, M., Xie, Y., Shen, Y., Ke, B., Qiao, R., Ren, B., Lin, S., Ma, L., 2022. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14930–14939.

Liao, Y., Xie, J., Geiger, A., 2022. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 3292–3310.

Liu, F., Zhang, C., Zheng, Y., Duan, Y., 2023. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17386–17396.

Liu, L., Gu, J., Zaw Lin, K., Chua, T.-S., Theobalt, C., 2020. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33, 15651–15663.

Lu, F., Xu, Y., Chen, G., Li, H., Lin, K.-Y., Jiang, C., 2023a. Urban Radiance Field Representation with Deformable Neural Mesh Primitives. *arXiv preprint arXiv:2307.10776*.

Lu, F., Xu, Y., Chen, G., Li, H., Lin, K.-Y., Jiang, C., 2023b. Urban radiance field representation with deformable neural mesh primitives. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 465–476.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

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

Qiu, R.-Z., Yang, G., Zeng, W., Wang, X., 2024. Language-driven physics-based scene synthesis and editing via feature splatting. *European Conference on Computer Vision*, Springer, 368–383.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L. et al., 2025. Sam 2: Segment anything in images and videos. *International Conference on Learning Representations*.

Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R., 2022. Image-to-lidar self-supervised distillation for autonomous driving data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9891–9901.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *The European Conference on Computer Vision*, Springer, 501–518.

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

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B. et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P. P., Barron, J. T., Kretzschmar, H., 2022. Block-nerf: Scalable large scene neural view synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Tao, A., Sapra, K., Catanzaro, B., 2020. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6411–6420.

Turki, H., Ramanan, D., Satyanarayanan, M., 2022. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12922–12931.

Turki, H., Zhang, J. Y., Ferroni, F., Ramanan, D., 2023. Suds: Scalable urban dynamic scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12375–12385.

Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., Zhang, J., 2024. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *The Conference on Neural Information Processing Systems*, 19114–19138.

Wu, Z., Liu, T., Luo, L., Zhong, Z., Chen, J., Xiao, H., Hou, C., Lou, H., Chen, Y., Yang, R. et al., 2023. Mars: An instance-aware, modular and realistic simulator for autonomous driving. *CAAI International Conference on Artificial Intelligence*, Springer, 3–15.

Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z., 2022. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. *European Conference on Computer Vision*, Springer, 677–695.

Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S., 2024. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. *European Conference on Computer Vision*, Springer, 156–173.

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

Zhai, H., Li, H., Li, Z., Pan, X., He, Y., Zhang, G., 2025. Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14114–14124.

Zhi, S., Laidlow, T., Leutenegger, S., Davison, A. J., 2021. In-place scene labelling and understanding with implicit scene representation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15838–15847.

Zhou, H., Shao, J., Xu, L., Bai, D., Qiu, W., Liu, B., Wang, Y., Geiger, A., Liao, Y., 2024a. Hugs: Holistic urban 3d scene understanding via gaussian splatting. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21336–21345.

Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A., 2024b. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21676–21685.

Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.-H., 2024c. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21634–21643.