ISPRS Journal of Photogrammetry and Remote Sensing 223 (2025) 274–295

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

3D semantic segmentation: Cluster-based sampling and proximity hashing for novel class discovery

Jing Du [a], Linlin Xu [b], Lingfei Ma [c], Kyle Gao [a], John Zelek [a,*], Jonathan Li [a,d,**]

[a] Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[b] Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China
[d] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Point cloud
Semantic segmentation
Novel class discovery
Deep learning
Neighborhood spatial partitioning

ABSTRACT

Novel Class Discovery (NCD) in 3D semantic segmentation is crucial for applications requiring the ability to learn and segment previously unknown classes in point cloud data, such as autonomous driving and urban planning. Traditional 3D semantic segmentation methods often build upon a fixed set of known classes, which restricts their ability to discover classes not covered in the original training data. To overcome these limitations, we propose a novel framework specifically designed for NCD in 3D semantic segmentation. The framework integrates the Voxel-Geometry Data Integration module, the Cluster-based Representative Sampling module, the Neighborhood Spatial Partitioning module, and the Spatial Feature Attention Mechanism. These modules collectively enhance the model's capability to integrate spatial and geometric information, identify key representative points, map neighborhoods effectively, and synthesize localized and global features. Experimental results on benchmark datasets, including S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS, demonstrate the proposed method's superior performance in discovering novel classes and improving overall segmentation quality. For instance, in the SemanticPOSS-4^0 split, the method achieves a mean Intersection over Union (mIoU) of 43.68% for novel classes, compared to 35.70% achieved by NOPS. These results highlight the framework's effectiveness in handling complex scenes and its potential to advance NCD in 3D semantic segmentation.

1. Introduction

Real-world 3D environments are often highly dynamic, with new object categories continually emerging in applications such as autonomous driving, robotics, and urban planning. However, most traditional 3D semantic segmentation pipelines (Qi et al., 2017; Zhou and Tuzel, 2018; Zhu et al., 2021; Zhao et al., 2021) rely on predefined classes, limiting their ability to adapt to novel classes not included in the training set. This limitation has led to the rise of Novel Class Discovery (NCD) in 3D semantic segmentation, whereby models build on existing labeled 3D data while autonomously segmenting previously unseen classes without depending on additional manual annotations (Riz et al., 2023).

Formally, NCD is defined as the task of classifying samples from unlabeled data into novel classes by leveraging the knowledge gained from labeled classes (Han et al., 2019). Base classes refer to categories present in the labeled dataset, whereas novel classes are those absent from the labeled dataset but found in unlabeled data (Riz et al., 2023). Therefore, NCD in 3D semantic segmentation involves assigning each point in a 3D point cloud a semantic label from either a base class or a novel class, enabling dynamic 3D scenes to be segmented into distinct, meaningful categories.

NCD is critical for bridging the gap between supervised learning and real-world scenarios, where labeling data for all possible classes is often impractical (Chi et al., 2022). In dynamic and unstructured environments, such as those encountered in autonomous systems and robotics, new object classes frequently emerge. The cost and time required for manual annotation make it infeasible to label every new class (Jia et al., 2021). NCD enables models to autonomously discover and learn novel classes from unlabeled data, enhancing their adaptability and ability to generalize without extensive human intervention (Zhong et al., 2021a). By integrating NCD into 3D semantic segmentation, models can improve generalization to novel situations, enhance performance across

* Corresponding author.
** Corresponding author at: Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
E-mail addresses: jzelek@uwaterloo.ca (J. Zelek), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.isprsjprs.2025.03.001
Received 16 July 2024; Received in revised form 16 January 2025; Accepted 1 March 2025
Available online 20 March 2025

diverse scenarios, and reduce dependence on comprehensive labeled datasets (Riz et al., 2023).

NCD has been explored in fields such as 2D image classification and semantic segmentation (Zhong et al., 2021b; Zhao et al., 2022). However, its application to 3D semantic segmentation introduces unique challenges. The high dimensionality and unstructured nature of point cloud data complicate the integration of spatial and semantic information, which is crucial for accurate segmentation and novel class discovery. Moreover, processing large-scale point clouds is computationally intensive, and existing methods often fail to manage this complexity without compromising spatial details. NOPS (Riz et al., 2023) extends NCD to 3D point cloud semantic segmentation. Unlike 2D methods that assume constraints such as a single novel class per image and leverage saliency detection, NOPS adapts to the complexities of 3D data. It clusters unlabeled points based on semantic similarity, performs batch-level clustering with continuous prototype updates, and introduces a queuing strategy to address potentially missing classes.

However, although NOPS (Riz et al., 2023) is a pioneering effort in introducing NCD to 3D point cloud semantic segmentation, several key challenges remain. First, NOPS primarily leverages voxel-based representations but does not thoroughly integrate the original spatial coordinates, which may limit the preservation of fine-grained geometric details, especially in scenes with multiple novel classes. Second, in highly irregular or overlapping regions, the method's approach to defining and updating neighborhoods can lead to ambiguities between base and novel categories. Third, NOPS offers only limited mechanisms for fusing local geometric information with broader global context, potentially yielding suboptimal feature representations and confusion at class boundaries.

These issues with NOPS exemplify fundamental challenges that persist across existing methods for NCD in 3D semantic segmentation. The key research questions that emerge revolve around efficiently processing high-dimensional point cloud data to preserve vital spatial and semantic details and balancing the computational complexity with the need to retain critical information for novel class discovery. Additionally, the unstructured nature and irregular distribution of points complicate defining and leveraging local neighborhood structures, which are essential for improving feature representation. Another crucial question is how to effectively integrate spatial and semantic information to enhance the model's capability for discovering and learning novel classes within complex 3D environments. Addressing these research questions is crucial to advancing NCD in 3D semantic segmentation and expanding the applicability of models to dynamic, real-world settings.

To address these challenges, a comprehensive deep learning framework is proposed that integrates voxel-based representations with original spatial coordinates using an index mapping mechanism. This integration combines structured voxel formats with detailed spatial information, maintaining spatial integrity and enabling efficient handling of point clouds. Although existing methods like PV-RCNN (Shi et al., 2020) and PVCNN (Liu et al., 2019) have effectively combined voxel and point-based feature extraction for 3D data analysis, utilizing the complementary strengths of both representations, our Voxel-Geometry Data Integration module (VGDI) offers a distinct approach to point-voxel integration. Differing from PV-RCNN, which emphasizes keypoint-based proposal refinement, or PV-CNN, which employs trilinear interpolation for feature extraction balance, VGDI performs index mapping between voxels and their original spatial coordinates. By extracting voxel features, VGDI enables efficient encoding of global context and reduces computational complexity. Through its index mapping mechanism, VGDI ensures that each voxel's features remain directly linked to their originating spatial coordinates, thereby preserving the geometric characteristics of each point and enhancing the representation of local neighborhoods. This adaptive integration captures both global structures and local geometric details, mitigating interpolation losses and minimizing spatial precision loss due to quantization. VGDI effectively supports downstream processes that require detailed spatial information and adaptive neighborhood formation. Voxel features facilitate clustering and attention mechanisms, while the original spatial coordinates ensure accurate spatial operations like precise distance calculations and neighbor search, critical for effective neighborhood formation and proximity-based searches.

Additionally, the proposed method employs a Cluster-based Representative Sampling module (CRSM), a Neighborhood Spatial Partitioning module (NSPM), and the Spatial Feature Attention Mechanism (SFAM). An innovative sampling strategy is introduced in the CRSM, which significantly reduces computational complexity by identifying and targeting key representative points within clusters in the high-dimensional feature space. This strategy focuses on a subset of points without sacrificing critical information, enhancing the model's effectiveness in segmenting and learning novel classes. The NSPM utilizing Proximity Hash Mapping (PHM) is implemented, mapping dense feature vectors to a lower-dimensional hash space to manage high-dimensional feature data and segment neighborhoods within point clouds. This approach organizes and analyzes local geometries and relationships, improving neighborhood searches essential for novel class discovery. The SFAM is designed, refining the understanding of spatial relationships by integrating localized feature vectors from neighboring points with global feature context. This mechanism dynamically adjusts focus based on spatial proximity and context, effectively extracting localized and global spatial features, leading to improved segmentation performance.

In summary, our contributions are as follows:

- A novel framework is proposed that integrates voxel and spatial coordinates for novel class discovery in 3D semantic segmentation.
- Specialized strategies are further developed to efficiently handle large-scale, high-dimensional point clouds and facilitate novel class segmentation. By selectively focusing on key representative points, computational complexity is reduced while retaining crucial information. Mapping local geometries to a lower-dimensional hash space is then introduced, which proves particularly beneficial for organizing sparse or irregular point distributions and improving neighborhood searches. Furthermore, local and global features are adaptively integrated to refine spatial representations. Collectively, these improvements strengthen the model's capacity to discover newly emerging semantic categories in complex 3D environments.
- Comprehensive experimental validation is conducted on four benchmark datasets—S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS. The results demonstrate that the proposed framework outperforms baseline methods in novel class discovery for 3D semantic segmentation, confirming its effectiveness and potential for real-world applications.

2. Related work

In the field of novel class discovery, significant advancements have been made over the years, starting with early solutions that introduced pairwise similarity predictions and learnable clustering objectives to enhance cross-domain and cross-task transfer learning (Hsu et al., 2018). These foundational works paved the way for more sophisticated methodologies, such as modifications to deep embedded clustering (Han et al., 2019), self-supervised learning frameworks (Han et al., 2020), and multi-modal data approaches (Jia et al., 2021). Recent advancements have introduced various innovative strategies, including two-branch learning frameworks for visual category discovery (Zhao and Han, 2021), neighborhood contrastive learning (Zhong et al., 2021a), and unified objectives that streamline the discovery process and improve model robustness (Fini et al., 2021; Roy et al.,

2022). Additionally, meta-learning-based approaches (Chi et al., 2022) and class-incremental frameworks such as FRoST have been developed to handle the dynamic nature of NCD (Roy et al., 2022). Further contributions to the field include the development of frameworks for specific challenges, such as leveraging mutual knowledge distillation (Gu et al., 2023), and implementing new loss functions like Spacing Loss to improve latent space separability (Joseph et al., 2022b). The integration of these diverse methodologies has led to significant improvements in the ability to discover and learn novel classes effectively, even in complex and dynamic environments (Zang et al., 2023; Yang et al., 2023; Li et al., 2023). This section provides a comprehensive review of these advancements, highlighting the key contributions and their impact on the field of novel class discovery.

Hsu et al. (2018) can be considered the first solution to the novel class discovery problem (Troisemaine et al., 2023). Hsu et al. (2018) presents a novel approach to transfer learning across domains and tasks through clustering, focusing on leveraging pairwise similarity information. The primary challenge addressed is how to use this similarity to enhance clustering and unsupervised transfer learning. By introducing a learnable clustering objective into neural networks, the method effectively improves cross-domain and cross-task transfer, marking a significant advancement in transfer learning and providing a robust framework for further research in clus-tering-based transfer methods. Building upon the concept of clustering for knowledge transfer, the study Han et al. (2019) addresses the problem of discovering novel object categories in unlabeled images by leveraging known related classes to reduce clustering ambiguity. The core research question focuses on how to use prior knowledge to improve the quality of discovered classes and accurately estimate the number of novel classes in unlabeled data without explicit labels. The proposed method extends Deep Embedded Clustering (DEC) to a transfer learning framework, introducing a representational bottleneck, temporal ensembling, and consistency constraints to improve clustering and representation learning. Furthermore, a novel approach is developed to estimate the number of classes in unlabeled data by utilizing known classes as probes. This dual strategy, which combines an enhanced DEC with class number estimation, effectively enables the discovery and recognition of new object categories within unlabeled data.

In 2020, Han et al. (2020) introduces the term "Novel Category Discovery", which laid the foundation for the novel class discovery problem (Troisemaine et al., 2023). The main challenge tackled in this research is identifying novel visual classes from unlabeled data without introducing bias from labeled data. This study presents a framework to address this issue using three key strategies: self-supervised learning to train a general-purpose image representation across both labeled and unlabeled data; leveraging rank statistics to effectively transfer knowledge for clustering unlabeled images; and a joint objective function that optimizes both classification and clustering to improve performance on known and novel classes. Subsequently, Zhao and Han (2021) proposes a dual-branch framework for novel visual category discovery in unlabeled data by leveraging knowledge transfer from labeled classes. One branch focuses on local part-level details, while the other captures global characteristics. The problem addressed is discovering new visual categories without any labeled instances, a more complex challenge than conventional semi-supervised learning. Similarly, Jia et al. (2021) addresses the problem of novel category discovery in single- and multi-modal data, where the challenge lies in automatically partitioning unlabeled data from unknown categories into semantic groups using knowledge from labeled data of different but related categories. The proposed end-to-end framework jointly learns representations and clusters unlabeled data, extending contrastive learning with category discrimination and cross-modal discrimination to handle both labeled and unlabeled data. By analyzing the difficulty of learning from sparse labeled data and the need for accurate clustering, it introduces Winner-Take-All (WTA) hashing to generate pairwise pseudo-labels for more robust knowledge transfer and clustering.

In 2021, Zhong et al. (2021a) expands previous research to define "Novel Class Discovery" as a distinct research area, establishing a more focused framework for studies in this field (Troisemaine et al., 2023). Zhong et al. (2021a) tackles the problem of how to effectively discover novel classes in unlabeled data by leveraging the structure learned from labeled datasets. Key challenges include ensuring discriminative feature learning for clustering and accurately identifying positive and hard negative pairs. The research questions center on how to exploit the neighborhood relationships in the embedding space to improve pseudo-positive pair selection and how to generate hard negatives to enhance contrastive learning without misclassifying true positives. By addressing these, the proposed framework significantly improves the accuracy of NCD tasks. Complementing this, Fini et al. (2021) introduces a Unified Objective (UNO) for NCD, addressing the challenges of combining supervised and unsupervised objectives within a single framework. It identifies the problem of disparate objective functions in current NCD approaches and simplifies the process by treating cluster pseudo-labels in the same way as ground truth labels. By leveraging a multi-view self-labeling strategy and employing multi-head and over-clustering techniques, UNO unifies learning on both labeled and unlabeled data using a single cross-entropy loss, effectively eliminating the need for multiple objectives and self-supervised pre-training, thereby improving novel class discovery. Additionally, Zhong et al. (2021b) introduces OpenMix, a method for NCD by mixing labeled and unlabeled data from disjoint classes. The main challenge addressed is leveraging labeled data to discover new classes in unlabeled data through reliable pseudo-label generation and clustering. OpenMix employs two strategies: mixing labeled and unlabeled data to enhance pseudo-label reliability and using reliable anchors to refine clustering, effectively preventing overfitting on incorrect pseudo-labels and improving performance.

As a related approach, NCDwF (Joseph et al., 2022a) introduces a model that incrementally discovers novel classes from unlabeled data while preserving the accuracy of previously learned categories. The approach uses pseudo-latent representations to mitigate forgetting and a mutual-information based regularizer to enhance novel class discovery. NCDwF analyzes the challenge of balancing between retaining knowledge of known classes and effectively clustering new, unseen categories without access to the original labeled data, presenting a key strategy for overcoming catastrophic forgetting in NCD. Furthermore, Zhao et al. (2022) presents the first approach for NCD in 2D semantic segmentation, addressing the problem of segmenting novel classes in unlabeled images using disjoint labeled data. The Entropy-based Uncertainty Modeling and Self-training (EUMS) framework is proposed, which utilizes entropy ranking to clean noisy pseudo-labels and dynamically reassigns data for self-supervised learning. By effectively analyzing the challenges of noisy pseudo-labels, inaccurate saliency maps, and the coexistence of multiple novel classes within an image, the EUMS framework enhances the model's ability to generalize to novel classes. This strategy is pivotal for advancing NCD in complex segmentation scenarios. In another approach, Yang et al. (2022) addresses the problem of Generalized Novel Class Discovery (GNCD), where existing methods struggle to generalize across both base and novel classes. The key challenge is balancing separability between these sets while maintaining discriminability within them, which is often overlooked in traditional NCD approaches. To tackle this, the authors introduce Compositional Experts (ComEx), leveraging batch-wise and class-wise experts to achieve comprehensive representation and improved clustering. By focusing on both global-to-local alignment and local consistency, the method effectively enhances clustering performance, offering a robust approach to GNCD.

Moving forward, Zang et al. (2023) addresses challenges in novel category discovery under open-set domain adaptation (ODA) and universal domain adaptation (UNDA) settings by introducing a Soft-contrastive All-in-one Network (SAN). The main problems identified are view-noise in data augmentation impacting feature transfer and

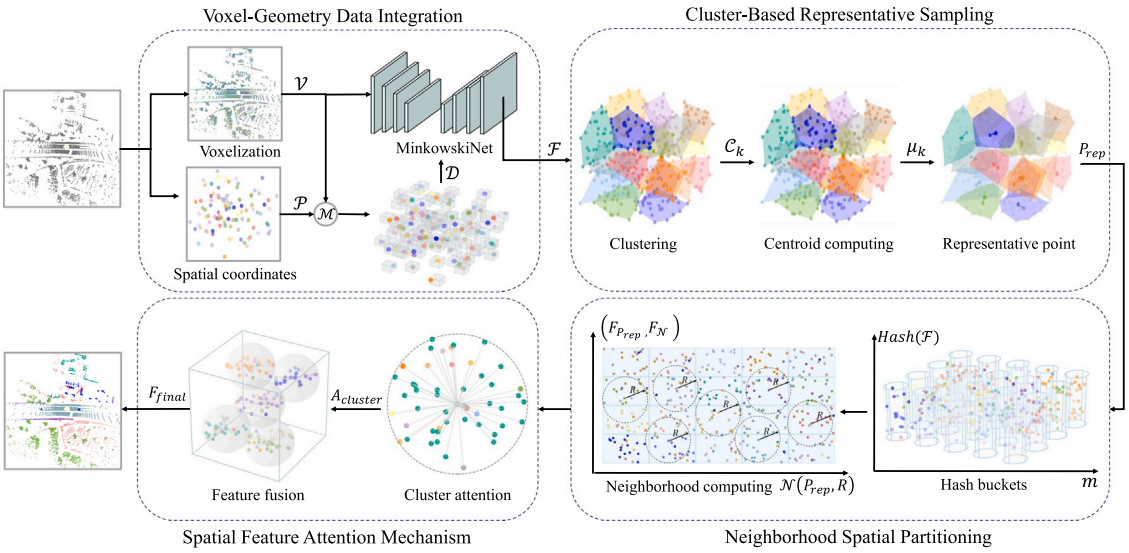

Fig. 1. Framework of the proposed network. The proposed network consists of four functional modules: Voxel-Geometry Data Integration module, Cluster-based Representative Sampling module, Neighborhood Spatial Partitioning module, and Spatial Feature Attention Mechanism. $\mathcal{V}$: Voxelized representation of the point cloud. $\mathcal{P}$: Spatial coordinates of the original point cloud. $\mathcal{M}$: Mapping function that links voxels to their originating points. $\mathcal{D}$: Enriched dataset combining voxelized data with original spatial details. $\mathcal{F}$: High-dimensional feature representation extracted from the enriched dataset. C_k: Clusters identified by MiniBatchKMeans. μ_k: Centroid of each cluster in feature space. P_{rep}: Representative point of each cluster. Hash($\mathcal{F}$): Feature vectors mapped to a lower-dimensional hash space. m: Number of hash buckets. $(F_{P_{\text{rep}}}, F_{\mathcal{N}})$: Localized feature vectors of representative points and their neighbors. $\mathcal{N}(P_{\text{rep}}, R)$: Neighborhood points within radius R of the representative point. $\mathbf{A}_{\text{cluster}}$: Attention vector for each cluster. $\mathbf{F}_{\text{final}}$: Final feature set integrating localized attention with global features.

overconfidence in novel category classifiers affecting recognition performance. SAN employs Soft Contrastive Learning (SCL) to mitigate noisy gradients during domain transfer and introduces an All-in-One (AIO) classifier to provide smoother classification boundaries. By analyzing these issues, SAN enhances strategies for NCD, focusing on robust feature alignment and improved category recognition. Furthermore, Gu et al. (2023) tackles the challenge of learning novel classes without supervision based on labeled known-class data. Existing methods often fail to capture the semantic relationship between known and novel classes, limiting knowledge transfer. To address this, the authors introduce a class-relation representation to retain inter-class relations during training. The key idea is to prevent information loss in class relations through a knowledge distillation framework, which utilizes a learnable weighting function to transfer knowledge adaptively based on the semantic similarity between novel and known classes, thereby improving novel class discovery.

Moreover, Yang et al. (2023) investigates the challenge of NCD in imbalanced data distributions. It analyzes the limitations of existing NCD methods, which assume a uniform class distribution, and addresses the research question of how to effectively discover novel classes without this assumption. The proposed BYOP method iteratively refines the class prior based on model predictions, improving pseudo-label accuracy. A dynamic temperature technique refines predictions, enhancing clustering performance in distribution-agnostic scenarios. BYOP outperforms current methods across benchmarks, providing robust solutions for imbalanced NCD. Similarly, Li et al. (2023) explores NCD by proposing a method to model both inter-class and intra-class constraints using symmetric Kullback–Leibler divergence (sKLD). A key issue is the limited use of the disjoint nature between labeled and unlabeled classes, leading to weak feature separability. Furthermore, there is a need for more effective regularization to maintain consistency between samples and their augmentations. The proposed approach introduces an inter-class sKLD constraint to enhance class separability and an intra-class sKLD constraint to ensure stable feature learning, effectively resolving these challenges.

3. Method

3.1. Motivation

NCD in 3D semantic segmentation presents unique challenges due to the high dimensionality, unstructured nature, and large volume of point cloud data. While NCD has seen progress in 2D image classification and segmentation, its application to 3D point clouds remains underdeveloped, primarily due to several intertwined issues. Firstly, the complexity of high-dimensional data is a major hurdle. Point clouds often consist of millions of points, each carrying spatial information, making their processing computationally demanding. Traditional methods struggle to scale efficiently without compromising spatial and semantic details crucial for segmentation and novel class detection. Accurately capturing these aspects is particularly challenging when novel classes have subtle distinctions.

Furthermore, efficient determination of local neighborhoods and effective feature representation in high-dimensional spaces are problematic. Standard nearest-neighbor searches are computationally costly, and ineffective feature representations can hinder the discovery of novel classes. Processing all points in large point clouds is computationally prohibitive, and indiscriminate downsampling risks losing critical information. Thus, intelligent sampling strategies are essential to retain key features while reducing computational overhead. Moreover, real-world 3D environments are dynamic, with constantly evolving objects and conditions. Models must adapt to new data distributions and remain robust to variations in object appearance and geometry, especially when labeled data for novel classes is unavailable. This necessitates methods capable of generalizing to unseen classes and adapting to changing environments.

To address the challenges of NCD in 3D semantic segmentation, the following research questions (RQ) are formulated:

RQ1: How can high-dimensional 3D point cloud data be processed while preserving critical spatial and semantic details necessary for NCD?

RQ2: What strategies can effectively reduce computational complexity while retaining key information necessary for segmenting novel classes in large-scale point clouds?

RQ3: How can local neighborhood structures within point clouds be identified and utilized to improve feature representation, thereby enhancing NCD performance?

RQ4: What methods can be developed to effectively integrate spatial and semantic information, enhancing the model's capability to identify and segment novel classes in complex environments?

A comprehensive deep learning framework is proposed to address these questions, overcoming the challenges associated with NCD in 3D semantic segmentation, as shown in Fig. 1. The proposed framework addresses each research question as follows:

- Preserving Spatial and Semantic Details (RQ1): The framework leverages voxel-based representations alongside original spatial coordinates to handle high-dimen-sional point cloud data while maintaining crucial spatial details for accurate NCD and segmentation.
- Balancing Complexity and Information Preservation (RQ2): An innovative sampling strategy is introduced to reduce computational load by focusing on informative regions without losing essential spatial and semantic features, enhancing the model's ability to segment novel classes effectively.
- Enhancing Neighborhood Representation (RQ3): The PHM is employed to identify and utilize neighborhood structures within point clouds, improving feature representation and supporting more effective NCD.
- Integrating Spatial and Semantic Information (RQ4): A spatial feature attention mechanism is implemented to integrate both local and global features, enabling the model to discover novel classes by synthesizing relevant spatial and semantic contexts.

3.2. Voxel-geometry data integration

In the approach to 3D point cloud semantic segmentation, we present a novel methodology termed Voxel-Geometry Data Integration module (VGDI) that combines voxelized representations with original spatial coordinates. This method maintains the intricate spatial details inherent in the original point cloud while providing a structured voxel format that facilitates accurate segmentation. The dual-representation mechanism is essential for achieving precise segmentation, particularly in scenarios involving the discovery of novel classes.

Given a point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N}$, where each point $p_i = (x_i, y_i, z_i) \in \mathbb{R}^3$, our goal is to process $\mathcal{P}$ in a manner that maintains its spatial integrity while rendering it amenable to computational models. To this end, we define a dual-mapping mechanism that involves the transformation of $\mathcal{P}$ into a voxelized representation $\mathcal{V}$ and the preservation of original point coordinates through a mapping $\mathcal{M}$.

The process begins with a voxelization function, f_{vox}, which aggregates the point cloud $\mathcal{P}$ into a structured voxel grid $\mathcal{V}$, where each voxel $v_j \in \mathcal{V}$ encapsulates a subset of points $\mathcal{P}_j \subseteq \mathcal{P}$ within a predefined volumetric unit δ^3, formalized as:

$$\mathcal{V} = f_{\text{vox}}\ (\mathcal{P}). \tag{1}$$

To bridge the voxelized representation and the point cloud's original spatial framework, we establish a bijective mapping $\mathcal{M}$ that links each voxel v_j to its originating subset of points $\mathcal{P}_j$, preserving and exploiting detailed spatial information crucial for segmentation accuracy:

$$\mathcal{M} : \mathcal{V} \leftrightarrow \{\mathcal{P}_j\}_{j=1}^{M}. \tag{2}$$

The integration of voxelized data with original spatial coordinates is orchestrated through an integration function, f_{int}, which synthesizes the information from both $\mathcal{V}$ and $\mathcal{M}$, preparing the enriched dataset, $\mathcal{D}$, for segmentation:

$$f_{\text{int}}\ (\mathcal{V}, \mathcal{M}) \rightarrow \mathcal{D}, \tag{3}$$

where $\mathcal{D}$ encompasses the enriched voxelized representations integrated with original spatial details, providing a structured and detailed input for the segmentation model.

Upon obtaining the enriched dataset $\mathcal{D}$, the next critical step involves the extraction of features conducive to segmentation. This is achieved through the MinkowskiNet model (Choy et al., 2019), a widely recognized and well-established method for voxel-based feature extraction (Riz et al., 2023). It employs sparse tensor operations to efficiently handle the high dimensionality and sparsity of 3D point clouds, allowing for accurate and comprehensive capture of spatial and semantic details. Additionally, MinkowskiNet employs specialized Minkowski convolution layers to efficiently handle sparse tensors, significantly reducing memory overhead in the voxel space while preserving voxel continuity and the ability to model global context. Compared to conventional 3D convolutional networks, Minkowski convolutions focus computational resources on non-empty voxels in sparse scenarios, thereby avoiding the redundant overhead caused by the abundance of empty voxels. This design leads to higher efficiency and superior accuracy in large-scale 3D scene processing. The official implementation also integrates optimized GPU kernels, enabling faster training and inference. These advantages have earned MinkowskiNet broad recognition and adoption for various indoor and outdoor 3D semantic segmentation tasks. In our framework, the use of MinkowskiNet further enhances high-dimensional feature representation and facilitates the precise discovery and segmentation of novel classes in large-scale, sparsely distributed point cloud scenes. The model processes the dataset $\mathcal{D}$ to output a high-dimensional feature representation for each voxel, enhancing the segmentation task:

$$\mathcal{F} = \text{MinkowskiNet}(\mathcal{D}) = \{f_i\}_{i=1}^{N}, \tag{4}$$

where $f_i \in \mathbb{R}^d$ denotes the feature vector associated with the ith element in $\mathcal{D}$, and N is the number of points.

3.3. Cluster-based representative sampling

The Cluster-based Representative Sampling module (CRSM) is integral for organizing complex spatial information into a comprehensible format for analysis. Utilizing features extracted via the MinkowskiNet model, this module categorizes the voluminous point cloud data into clusters and identifies representative points that epitomize the essential characteristics of each cluster.

The clustering phase organizes these points into k groups based on their feature similarities. This organization is achie-ved through the MiniBatchKMeans algorithm, expressed mathematically as:

$$C_1, C_2, \ldots, C_k = \text{MiniBatchKMeans}(\mathcal{F}, k), \tag{5}$$

where C_k denotes the set of indices of points belonging to the kth cluster. For each cluster C_k, we compute its centroid μ_k in the feature space, which represents the average of the features within the cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} f_i, \tag{6}$$

where $|C_k|$ is the number of points in C_k. The representative point for each cluster, P_{rep}, is then selected as the one whose feature vector is nearest to the centroid, minimizing the Euclidean distance to the centroid:

$$P_{\text{rep}}\ = \underset{i \in C_k}{\text{argmin}} \left\| f_i - \mu_k \right\|_2. \tag{7}$$

This process results in a succinct representation of the point cloud through representative points, each symbolizing the core features of its respective cluster. This selection is pivotal for capturing the nuanced distinctions across various segments of the point cloud, thereby facilitating a focused examination of critical areas.

In summary, the CRSM streamlines the segmentation process by efficiently delineating point cloud data into distinct clusters and pinpointing representative points within each cluster. This technique enhances the segmentation model's precision by focusing on localized spatial patterns and crucial structural elements of the point cloud. By balancing

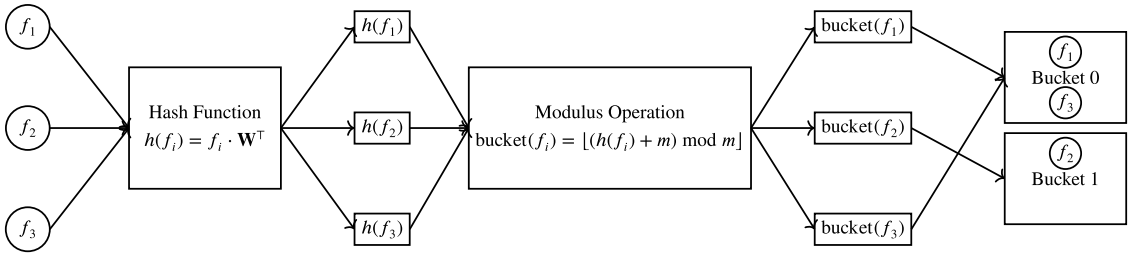

Fig. 2. Illustration of the Proximity Hash Mapping (PHM) process. High-dimensional feature vectors f_i are projected using the hash function $h(f_i) = f_i \cdot \mathbf{W}^\top$, resulting in real-valued hash outputs $h(f_i)$. The modulus operation maps these values into discrete bucket indices $\text{bucket}(f_i)$. The computation steps show how each feature vector is assigned to a bucket based on its bucket index. In this example, f_1 and f_3 are assigned to Bucket 0, while f_2 is assigned to Bucket 1, demonstrating how similar features are grouped together for efficient neighbor retrieval.

computational efficiency with the retention of essential spatial details, this approach simplifies the complexity of point cloud data while preserving its inherent semantic value. These representative points capture each cluster's distinguishing features and serve as reference points for understanding localized spatial patterns. By systematically clustering and focusing on these representatives, the model efficiently captures the intricate details and nuances present across different point cloud regions.

3.4. Neighborhood spatial partitioning

The Neighborhood Spatial Partitioning module (NSPM) is pivotal for dissecting and analyzing the spatial relationships within the data, emphasizing the segmentation of local neighborhoods around designated representative points. The NSPM commences with the application of PHM to the feature space extracted from the point cloud. PHM serves as an effective strategy for managing high-dimensional data, rendering it exceptionally beneficial for the segmentation of neighborhoods within point clouds.

In this implementation, the hash function h is constructed using random projection matrices. Specifically, we initialize k random hash functions represented by a matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$, where d is the dimensionality of the feature vectors. Each feature vector f_i is projected onto these hash functions to obtain hash values:

$$h(f_i) = f_i \cdot \mathbf{W}^\top, \tag{8}$$

where $h(f_i) \in \mathbb{R}^k$ is a vector of real numbers resulting from the dot product between the feature vector f_i and the transpose of the hash function matrix $\mathbf{W}$. To map these real-valued hash values to discrete hash buckets, we apply the modulus operation, which is well-defined for real numbers and returns the remainder after division. The hash bucket for each feature vector is computed as:

$$\text{bucket}(f_i) = \left\lfloor \left(h(f_i) + m\right) \bmod m \right\rfloor, \tag{9}$$

where m is the number of buckets, and the addition of m ensures that the hash values are non-negative. The modulus operation mod m is applied element-wise to $h(f_i)+m$, resulting in a vector within the range $[0, m)$. The floor operation $\lfloor \cdot \rfloor$ converts these real numbers into integer bucket indices. The modulus operation for real numbers is defined as:

$$a \bmod m = a - m \cdot \left\lfloor \frac{a}{m} \right\rfloor, \tag{10}$$

which ensures the result lies in the interval $[0, m)$. This definition allows the modulus operation to be applied to real numbers, making it suitable for our hashing process. The process is governed by:

$$\text{Hash}(\mathcal{F}) = \bigcup_{i=1}^{N} \text{bucket}(f_i), \tag{11}$$

where $\text{Hash}(\mathcal{F})$ represents the aggregation of all hash buckets containing the feature vectors f_i. The index variable i iterates over each point in the sequence, ranging from 1 to N, where N represents the total number of points in the point cloud. The union operation ($\bigcup$) aggregates all such hash buckets, ensuring a comprehensive coverage of the feature space.

This hashing mechanism efficiently reduces the high-dimensional feature space into a lower-dimensional representation, allowing for rapid retrieval of neighboring points. By using random projections, similar feature vectors are likely to produce similar hash values, increasing the probability that they reside in the same or nearby buckets.

As illustrated in Fig. 2, the PHM process begins with high-dimensional feature vectors f_1, f_2, and f_3. These vectors are input into the hash function h to obtain real-valued hash outputs $h(f_i)$. The modulus operation is then applied to map these hash values into discrete bucket indices $\text{bucket}(f_i)$. Following this, each feature vector f_i is assigned to the corresponding hash bucket based on the computed bucket index. The detailed computation steps and formulas demonstrate how the feature vectors are effectively grouped through the hashing function and modulus operation, facilitating efficient neighbor retrieval within the feature space.

Following the PHM application, for each cluster identified in the preceding clustering phase, a representative point P_{rep} is selected based on its proximity to the cluster's centroid. This selection is crucial for accurately reflecting the spatial characteristics of the cluster. The radius R surrounding P_{rep} is dynamically computed to define the local neighborhood, which is adapted according to the spatial distribution within the cluster. The calculation of this radius is critical for encompassing neighboring points and is defined by:

$$R = \gamma \cdot \rho_{\text{rep}}, \tag{12}$$

where γ serves as a scaling factor, adjusting the neighborhood size to accommodate the spatial variability within the cluster. Here, ρ_{rep} is the range value obtained from the feature vectors, specifically for the representative point, determined through network training.

The subsequent step involves identifying neighboring points within the radius of each representative point. Utilizing the hash buckets from the PHM, we efficiently retrieve candidate neighboring points. For each representative point, we examine the buckets corresponding to its hash codes and collect indices of points within those buckets. This significantly reduces the search space compared to evaluating all points in the dataset.

The selection of neighboring points $\mathcal{N}$ is governed by their spatial proximity to the representative point P_{rep}. A key aspect here is the utilization of the original spatial coordinates from the point cloud to calculate distances. Neighbors are determined based on a threshold distance from the representative point, ensuring spatial relevance and precision. This identification is encapsulated as:

$$\mathcal{N}\left(P_{\text{rep}}, R\right) = \left\{ p_j \in \mathcal{P} \mid \left\| p_j - P_{\text{rep}} \right\|_2 \leq R \right\}, \tag{13}$$

where $\mathcal{N}\left(P_{\text{rep}}, R\right)$ denotes the set of points within a radius R from a representative point P_{rep}, indicating the neighborhood of interest. Here, P_{rep} is the selected representative point for a cluster, chosen to best represent its spatial features. The radius R defines the extent of the

neighborhood, enclosing points p from the overall point cloud $\mathcal{P}$ that lie within this spatial boundary. The inclusion criterion $\left\| p - P_{\text{rep}} \right\|_2 \leq R$, utilizing the Euclidean distance, ensures that only points within R of P_{rep} are considered for neighborhood analysis. This mechanism focuses on analyzing local geometries and spatial relationships by establishing a precisely defined area around representative points, thereby facilitating the detailed exploration of local patterns crucial for high-precision segmentation tasks.

By leveraging the hash buckets to limit the set of candidate points and then applying a distance threshold, we efficiently identify spatially relevant neighbors for each representative point. This two-step process – hash-based filtering followed by precise distance computation – balances computational complexity with accuracy in neighborhood formation. This proximity-based approach is pivotal in maintaining contextual accuracy and ensuring that only spatially relevant points are considered for neighborhood formation. By focusing on representative points and their respective neighborhoods, this method enables a detailed analysis of local geometries and relationships within the point cloud. It captures local variations and patterns crucial for high-precision segmentation tasks and reduces computational complexity by concentrating on key spatial areas rather than processing the entire point cloud uniformly. In summary, the NSPM utilizes PHM to efficiently partition the feature space and identify spatially relevant neighborhoods around representative points. This approach effectively addresses the challenges of managing high-dimensional data and computational complexity, enhancing the model's ability to discern complex spatial structures and discover novel classes within 3D point cloud data.

3.5. Spatial feature attention mechanism

The Spatial Feature Attention Mechanism module (SFAM) refines the understanding of spatial relationships within the data. It leverages the outcomes of NSPM, emphasizing significant spatial patterns through an attention-based process. This module integrates localized feature vectors derived from neighboring points identified in the previous module. The representative points, central to each cluster, and their neighboring points' features form the basis of this process. These features are then subjected to a specialized attention mechanism, designed to discern and emphasize significant spatial patterns within these localized segments. This module underscores the importance of localized spatial relationships and mitigates computational complexity by focusing attention on a subset of points rather than the entire point cloud.

The SFAM leverages a multi-head attention mechanism that operates on the principle of selectively aggregating information from the neighboring points of each cluster's representative point. The process begins with linear transformations of the feature vectors for the representative point P_{rep} and its neighbors $\mathcal{N}$, generating query Q, key K, and value V vectors. The attention scores are computed as the dot product between Q and K, normalized through a softmax function to derive attention weights α. These weights are then used to calculate a weighted sum of the value vectors, producing an attention vector A that encapsulates the synthesized neighborhood information.

Mathematically, the transformation and attention computation can be expressed as follows:

$$Q = W^Q \cdot F_{P_{rep}}, K = W^K \cdot F_{\mathcal{N}}, V = W^V \cdot F_{\mathcal{N}}, \tag{14}$$

where W^Q, W^K, and W^V represent the weight matrices for the query, key, and value vectors, respectively, and $F_{P_{\text{rep}}}, F_{\mathcal{N}}$ are the feature vectors of the representative point and its neighbors.

The attention scores are calculated by the dot product of Q and K, followed by a softmax normalization to produce attention weights α, which are then used to compute a weighted sum of the value vectors, yielding an attention vector A that encapsulates combined neighborhood information:

$$\alpha = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \tag{15}$$

$$\mathbf{A} = \alpha V, \tag{16}$$

where d_k is the dimensionality of the key vectors, ensuring the attention scores' stability across varying scales. Q and K are the transformed feature vectors for the representative point and its neighbors, facilitating the computation of attention scores. α denotes the attention weights, determining the significance of each neighbor's features relative to P_{rep}. A is the attention vector, representing the aggregated information from P_{rep} and its neighborhood, enhanced by the attention mechanism.

Next, the attention vector $\mathbf{A}_{\text{cluster}}$ for each cluster is computed by averaging the attention results across the feature dimension for the cluster's representative point and its neighbors:

$$\mathbf{A}_{\text{cluster}} = \frac{1}{d}\sum_{i=1}^{d} \mathbf{A}_i, \tag{17}$$

where d is the feature dimension, and $\mathbf{A}_i$ denotes the ith feature in the attention result for the cluster.

The computed attention vector $\mathbf{A}_{\text{cluster}}$ is then distributed to all points within the cluster, ensuring that the localized attention effect permeates the entire cluster:

$$\mathbf{F}_{\text{updated}} = \mathcal{F}_{C_k} + \mathbf{A}_{\text{cluster}}, \tag{18}$$

where $\mathbf{F}_{\text{updated}}$ represents the updated feature vector for each point in the cluster, incorporating both the original features $\mathcal{F}_{C_k}$ and the attention vector $\mathbf{A}_{\text{cluster}}$. The final step involves merging the localized attention features with global spatial features extracted from the dataset, culminating in an enriched feature set $\mathbf{F}_{\text{final}}$:

$$\mathbf{F}_{\text{final}} = \mathbf{F}_{\text{global}} \oplus \mathbf{F}_{\text{updated}}, \tag{19}$$

where $\oplus$ signifies the fusion operation, and $\mathbf{F}_{\text{global}}$ represents global features $\mathcal{F}$ extracted at an earlier stage of the pipeline.

In summary, the SFAM operates by applying linear transformations to the features of the representative point and its neighbors. The transformed features are then partitioned into multiple heads, allowing the model to process information in parallel, thereby capturing different aspects of the spatial relationships in the data. The attention mechanism computes scores by performing a dot product between the query (features of the representative point) and keys (features of neighboring points). These scores are then normalized using the softmax function, resulting in attention weights. These weights represent the significance of each neighboring point's features in relation to the representative point. The attention weights are used to compute a weighted sum of the value vectors, producing an attention vector for each representative point. This vector encapsulates aggregated spatial information of the representative point and its neighborhood. A critical aspect of this module is the propagation of attention across all points within each cluster. The attention vector computed for each cluster's representative point is distributed to all points within that cluster. This distribution ensures that the localized attention influences the entire cluster, enriching the feature representation with both local and global spatial contexts. The final step in this module involves the fusion of localized attention vectors with the global features extracted earlier in the pipeline. This fusion results in a comprehensive feature set that incorporates both localized attention and broader spatial context, enhancing the model's ability to perform detailed segmentation.

4. Experiments

4.1. Datasets

The S3DIS dataset (Armeni et al., 2016) includes detailed 3D point clouds of five extensive indoor areas from three different buildings, offering a diverse exploration into architectural styles and functionalities.

These areas encompass office spaces, educational facilities, exhibition halls, and various common areas such as lobbies, staircases, and restrooms. Altogether, the S3DIS dataset comprises over 215 million points and extends across more than 6000 square meters. The spaces covered range in size from approximately 450 to 1900 square meters, with some environments spanning multiple floors, offering a comprehensive view of indoor architectural diversity. It also introduces a fine-grained classification of 13 semantic elements, including structural components like ceilings, floors, and walls, as well as common furnishings such as tables, chairs, and sofas. This level of detail enables its application to 3D indoor space understanding, indoor navigation, architectural design, and virtual environment creation.

The Toronto-3D dataset (Tan et al., 2020) is a large-scale urban outdoor point cloud dataset designed for semantic segmentation of urban roadways. It was collected in Toronto, Canada, using a Mobile Laser Scanning (MLS) system. The dataset encompasses approximately 1 km of urban roadways, with a total of about 78.3 million points. It includes 8 labeled object classes for detailed analysis — road, road marking, natural (excluding grass and bare soil), building, utility line, pole, car, and fence, plus an additional class for unclassified points. By providing a detailed and accurately labeled dataset, Toronto-3D offers a valuable resource for developing and testing deep learning models in real-world urban scene understanding.

The SemanticSTF dataset (Xiao et al., 2023) is a large-scale, point-wise annotated dataset designed for 3D semantic segmentation under various adverse weather conditions. It leverages the STF (Bijelic et al., 2020) benchmark, collected in Germany, Sweden, Denmark, and Finland, capturing LiDAR scans in conditions such as snow, dense fog, light fog, and rain. The dataset includes 2076 scans from a Velodyne HDL64 S3D LiDAR sensor, carefully selected for geographical diversity and split into 1326 training scans, 250 validation scans, and 500 testing scans.

The SemanticPOSS dataset (Pan et al., 2020) is a comprehensive LiDAR point cloud dataset designed for 3D semantic segmentation tasks. It is collected using a vehicle equipped with a Pandora sensor module, integrating a 40-channel LiDAR with 0.33-degree vertical resolution, a forward-facing color camera, four wide-angle mono cameras for 360-degree coverage, and GPS/IMU for precise localization and annotation support. The SemanticPOSS dataset contains 2988 LiDAR scans, offering a comprehensive view of various and complex environments. These scans were meticulously gathered around Peking University, providing a real-world backdrop that includes teaching buildings, school gates, main roads, parking lots, and dynamic interactions with pedestrians and vehicles. Unlike many existing datasets, SemanticPOSS emphasizes the inclusion of a large number of dynamic instances, such as people, cars, and riders. This focus is crucial for autonomous driving systems, which must accurately detect and interact with moving objects in their vicinity.

In summary, we include one large-scale indoor dataset (S3DIS) and three outdoor datasets (Toronto-3D, SemanticSTF, and SemanticPOSS) to ensure thorough evaluation under diverse real-world conditions and scene complexities. S3DIS provides a rich indoor environment with cluttered office spaces and distinct architectural elements, effectively testing our framework's ability to handle fine-grained, structured interiors. Meanwhile, the three outdoor datasets differ markedly in geographic scale, weather conditions, and object distributions: Toronto-3D spans broad urban road networks and building facades; SemanticSTF encompasses various regional settings and challenging weather scenarios (fog, rain, snow); and SemanticPOSS offers dynamic interactions with large numbers of vehicles and pedestrians. By covering these different outdoor contexts, we can verify whether our method generalizes to both urban and non-urban settings, distinct sensor modalities, and a diverse array of object classes. Examining performance across these indoor–outdoor scenarios further highlights the framework's scalability and robustness to varying spatial densities, environmental conditions,

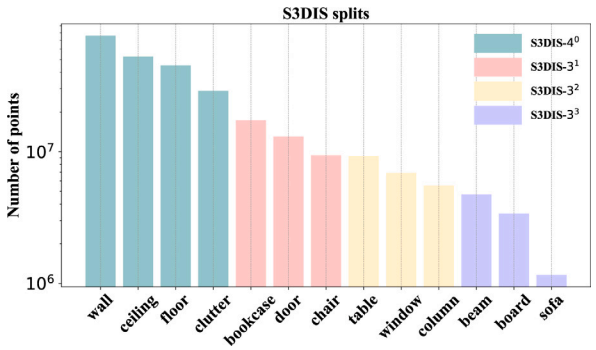

Fig. 3. Histograms illustrating the number of points for each class on the S3DIS dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

and novel classes in structured (indoor) as well as unstructured (outdoor) domains. This balanced set of four benchmarks aims not only to demonstrate overall accuracy but also to showcase the model's adaptability and broader applicability for NCD in real-world 3D semantic segmentation tasks.

4.2. Dataset preparation

For the S3DIS dataset, we adhere to the widely-used convention of using Area-5 as the test set and the remaining areas as the training set. For the Toronto-3D dataset, we follow its standard splitting guideline by using the L002 area as the test set, while L001, L003, and L004 are used as the training set. Considering that each file in the Toronto-3D dataset contains tens of millions of points, we partition the dataset into fixed subsets of 100,000 points each. If a subset initially contains points from only one label, we extend its range by adding exactly 100,000 more points at a time until multiple labels are represented. If the final subset contains fewer than 100,000 points or consists of only a single label, it is merged with the previous subset to maintain label diversity and size consistency. This preprocessing step ensures balanced and manageable data splits for efficient training and evaluation. Both the with offset and without offset versions of the Toronto-3D dataset were processed using the same splitting and preprocessing steps outlined above. For the SemanticSTF dataset, we follow the official guidelines for splitting the training set. Since the test set does not provide labels, we use the validation set as our test set. Additionally, as the SemanticSTF dataset includes different weather conditions (dense fog, light fog, rain, and snow), we split the training data into four groups, where each group contains only the training data corresponding to a single weather condition. In contrast, the test set is kept as a single group that includes all weather conditions without further subdivision. For the SemanticPOSS dataset, we adhere to the official splitting guidelines, dividing the data into six parts. To remain consistent with NOPS, we use part 03 as the test set and the remaining parts as the training set.

To divide base and novel classes, we first calculate the proportion of points for each class in the dataset. All classes are sorted in descending order based on these proportions, ensuring that classes with higher proportions of points are prioritized. To maintain consistency and fairness in comparison, we follow a standardized division criterion in line with prior work (Riz et al., 2023). The sorted classes are then divided into four groups as evenly as possible. Specifically, the number of classes in each group is determined to be roughly equal, and if the classes cannot be perfectly evenly divided, the excess classes are distributed among the initial groups to maintain balance. For each experimental arrangement, one of these groups is designated as the novel class group, while the remaining three groups form the base class group. This ensures that the novel classes are balanced across different splits. Figs. 3, 4, 5, and

Table 1
Novel class discovery results on the S3DIS dataset (%). Pink highlighted values are the novel classes in each split. "Novel" denotes the mIoU of novel classes, "Base" indicates the mIoU of non-novel classes, and "All" shows the mIoU of all classes.

Split	Model	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter	Novel	Base	All
S3DIS-4^0	NOPS	43.35	65.45	25.72	0.00	24.92	17.56	57.39	65.42	72.74	49.56	60.24	7.66	14.49	37.25	39.50	38.81
	Ours	78.75	63.92	36.84	0.00	31.26	28.07	54.11	67.07	80.40	41.76	60.75	10.73	4.08	45.90	41.57	42.90
S3DIS-3^1	NOPS	86.99	93.03	67.32	0.00	32.65	23.52	17.39	70.42	43.24	35.12	33.97	20.15	42.07	31.53	47.13	43.53
	Ours	85.61	93.36	66.61	0.00	31.01	28.43	35.74	69.92	68.26	30.04	50.38	12.89	42.28	51.46	46.02	47.27
S3DIS-3^2	NOPS	85.50	92.99	68.01	0.00	10.06	15.37	57.10	53.23	76.50	39.21	62.86	15.43	42.56	26.22	54.02	47.60
	Ours	86.63	93.42	66.01	0.00	10.14	11.05	57.59	64.73	81.44	33.61	62.70	13.61	44.29	28.64	53.93	48.09
S3DIS-3^3	NOPS	87.97	93.72	67.41	0.10	27.76	22.20	57.88	66.48	75.89	22.30	58.94	6.01	40.77	9.47	59.90	48.26
	Ours	87.68	92.80	65.67	0.00	38.85	31.18	58.00	69.84	78.83	29.79	65.27	6.86	41.42	12.22	62.95	51.25

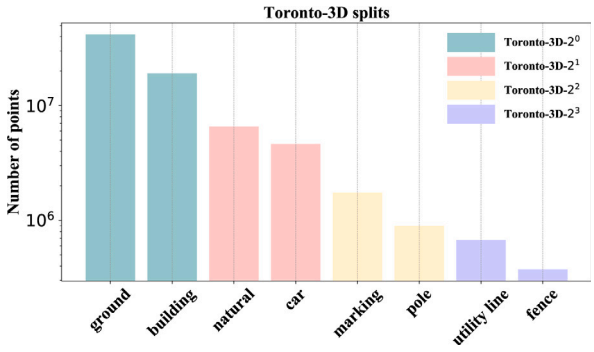

Fig. 4. Histograms illustrating the number of points for each class on the Toronto-3D dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

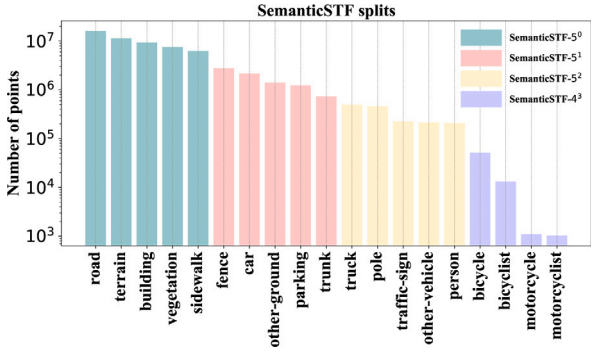

Fig. 5. Histograms illustrating the number of points for each class on the SemanticSTF dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

6 correspond to the four datasets – S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS – respectively, illustrating the distribution of points for each class and highlighting the novel classes in the corresponding split.

4.3. Experimental setup

In the experimental setup, we use a learning rate of 1e-2, a voxel size of 0.05, downsampling to 60,000 points per point cloud file, and a batch size of 4 across all datasets. We train for 500 epochs on the S3DIS and Toronto-3D datasets, and 100 epochs on the SemanticSTF dataset. These settings align with the baseline NOPS to ensure fair comparison.

4.4. Evaluation metrics

In this study, we use Intersection over Union (IoU) and mIoU as evaluation metrics for point cloud semantic segmentation. We evaluate the IoU for each class and compute the mIoU for novel classes, base classes, and all classes, as shown in Tables 1, 2, 3, and 4.

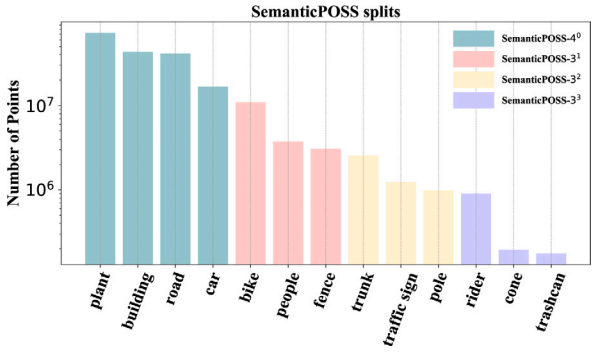

Fig. 6. Histograms illustrating the number of points for each class on the SemanticPOSS dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

The IoU measures the overlap between the predicted segmentation and the ground truth segmentation. Mathematically, it is expressed as:

$$\mathrm{IoU}_i = \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \tag{20}$$

where P_i represents the set of points predicted to belong to class i, and G_i represents the set of points in the ground truth that actually belong to class i.

In addition to evaluating the IoU for each class, we also calculate the mIoU for different subsets of classes to provide a comprehensive assessment. Specifically, we compute the mIoU for novel classes ($\mathrm{mIoU}_{\mathrm{Novel}}$), base classes ($\mathrm{mIoU}_{\mathrm{Base}}$), and all classes ($\mathrm{mIoU}_{\mathrm{All}}$). These are defined as follows:

$$\mathrm{mIoU}_{\mathrm{Novel}} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{IoU}_j, \tag{21}$$

$$\mathrm{mIoU}_{\mathrm{Base}} = \frac{1}{B}\sum_{k=1}^{B} \mathrm{IoU}_k, \tag{22}$$

$$\mathrm{mIoU}_{\mathrm{All}} = \frac{1}{C}\sum_{i=1}^{C} \mathrm{IoU}_i, \tag{23}$$

where N is the number of novel classes, B is the number of base classes, and C is the total number of classes, including both base and novel, and IoU_j, IoU_k, IoU_i are the IoU of the respective classes. These metrics provide a comprehensive evaluation of our method, reflecting its performance across different class subsets.

4.5. Evaluation and visualization results on S3DIS

We evaluate our framework using a detailed experimental protocol for 3D novel class discovery. Following the guidelines set by NOPS, the first method to apply NCD to 3D point cloud semantic segmentation, we adopt similar criteria for creating dataset splits. Currently, NOPS is the only publicly available method for 3D NCD, making it a key baseline for comparison.

The results presented in Table 1 demonstrate our framework's superior performance on the S3DIS dataset compared to NOPS across various splits, particularly in segmenting novel classes. In the S3DIS-4^0

Table 2
Novel class discovery results on the Toronto-3D dataset without offset (%). Pink highlighted values are the novel classes in each split. "Novel" denotes the mIoU of novel classes, "Base" indicates the mIoU of non-novel classes, and "All" shows the mIoU of all classes.

Split	Model	road	road marking	natural	building	utility line	pole	car	fence	Novel	Base	All
Toronto-3D-2^0	NOPS	52.71	8.32	85.08	56.48	32.59	44.10	45.11	23.97	54.60	39.86	43.55
	Ours	74.50	7.90	89.33	60.33	48.51	49.87	55.22	16.23	67.41	44.51	50.24
Toronto-3D-2^1	NOPS	77.74	14.06	46.84	61.97	45.27	53.12	15.95	19.27	31.40	45.24	41.78
	Ours	85.67	7.90	56.03	65.02	61.49	61.38	4.68	21.05	30.36	50.42	45.40
Toronto-3D-2^2	NOPS	74.88	9.66	86.41	61.43	30.84	55.56	51.90	23.15	32.61	54.77	49.23
	Ours	84.93	9.56	88.94	64.05	50.66	55.07	47.75	21.26	32.32	59.60	52.78
Toronto-3D-2^3	NOPS	77.22	14.84	86.65	59.01	34.28	54.22	51.29	14.47	24.38	57.21	49.00
	Ours	86.39	9.81	88.10	62.14	38.65	52.01	51.28	11.34	25.00	58.29	49.97

Table 3
Novel class discovery results on the SemanticSTF dataset (%). Pink highlighted values are the novel classes in each split. "Novel" denotes the mIoU of novel classes, "Base" indicates the mIoU of non-novel classes, and "All" shows the mIoU of all classes.

Split	Model	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	Novel	Base	All
SemanticSTF-5^0	NOPS	66.77	17.18	19.02	28.73	21.25	45.33	57.13	11.59	47.40	19.77	13.57	20.23	23.57	41.01	21.80	19.02	24.54	29.79	32.95	26.17	30.70	29.51
	Ours	59.38	23.70	14.00	37.31	35.64	43.22	61.02	22.10	60.65	23.04	13.33	23.01	26.68	42.42	23.73	23.02	23.66	28.11	35.00	29.61	33.64	32.58
SemanticSTF-5^1	NOPS	30.41	23.06	9.62	31.74	26.72	46.43	53.82	11.14	77.17	12.38	46.20	5.24	65.34	30.40	67.84	4.99	59.29	36.39	32.97	16.68	41.98	35.32
	Ours	29.89	31.64	13.62	38.32	33.55	48.25	62.33	18.66	78.77	11.46	49.25	0.07	69.37	37.04	60.52	5.50	57.10	38.77	37.68	16.79	45.56	37.99
SemanticSTF-5^2	NOPS	68.31	22.51	20.73	15.87	0.48	13.72	41.92	11.86	78.55	21.71	47.78	25.91	63.45	47.53	66.85	28.45	60.06	24.92	14.80	13.96	43.26	35.56
	Ours	68.96	32.38	14.61	12.94	1.04	12.29	64.36	22.89	79.62	24.35	50.52	28.65	71.54	48.16	62.40	30.96	57.64	25.43	14.30	13.20	46.93	38.05
SemanticSTF-4^3	NOPS	71.95	12.47	0.00	30.36	27.94	48.46	18.63	8.17	78.48	21.36	48.93	23.83	67.57	49.53	65.99	26.92	59.34	40.86	39.05	9.82	46.70	38.94
	Ours	70.30	16.15	5.59	34.74	34.05	50.71	22.10	8.04	80.62	24.61	51.28	30.31	72.79	49.08	60.91	29.67	59.58	35.31	38.71	12.97	48.18	40.77

Table 4
Novel class discovery results on the SemanticPOSS dataset (%). Pink highlighted values are the novel classes in each split. "Novel" denotes the mIoU of novel classes, "Base" indicates the mIoU of non-novel classes, and "All" shows the mIoU of all classes.

Split	Model	bike	building	car	cone	fence	road	people	plant	pole	rider	traffic-sign	trashcan	trunk	Novel	Base	All
SemanticPOSS-4^0	NOPS	35.47	30.35	1.24	13.52	24.13	69.14	44.70	42.07	19.19	47.65	24.44	8.17	21.82	35.70	26.57	29.38
	Ours	42.44	54.83	1.67	17.59	32.20	58.33	44.51	59.90	25.98	42.33	25.73	4.56	20.32	43.68	28.41	33.11
SemanticPOSS-3^1	NOPS	29.35	71.35	28.70	12.21	3.94	78.24	56.78	74.21	18.29	38.88	23.31	13.74	23.51	30.02	38.24	36.35
	Ours	8.36	77.48	35.07	28.55	23.54	78.73	44.09	76.68	18.19	45.69	25.08	4.42	20.33	25.33	41.02	37.40
SemanticPOSS-3^2	NOPS	37.16	71.81	29.74	14.64	28.38	77.53	52.09	73.00	11.51	47.11	0.54	10.20	14.79	8.95	44.17	36.04
	Ours	43.03	78.32	33.36	24.32	36.76	76.83	55.81	76.62	9.61	50.13	3.87	6.57	17.43	10.30	48.18	39.44
SemanticPOSS-3^3	NOPS	38.55	70.36	30.91	0.00	29.38	76.50	55.98	71.84	17.03	31.87	26.15	0.95	22.57	10.94	43.93	36.32
	Ours	38.72	75.49	31.80	13.66	29.94	77.84	54.83	74.53	18.23	32.39	30.52	1.69	23.61	15.91	45.55	38.71

split, our method achieves an IoU of 78.75% for the novel ceiling class, significantly surpassing the 43.35% IoU achieved by NOPS. Additionally, our approach yields a higher IoU for the novel wall class, with 36.84% compared to NOPS's 25.72%. For this split, our method attains an mIoU of 45.90% for novel classes, improving upon NOPS's 37.25%, and an mIoU of 42.90% across all classes, indicating a balanced performance across both novel and base classes.

In the S3DIS-3^1 split, substantial improvements are demonstrated in novel classes such as chair, where our framework achieves an IoU of 68.26%, 25.02% higher than the 43.24% obtained by NOPS. Similarly, for the door and bookcase classes, our approach reaches 35.74% and 50.38% IoU, respectively, significantly higher than NOPS's performance. For this split, our method achieves an mIoU of 51.46% for novel classes, compared to NOPS's 31.53%. The mIoU for base classes and all classes are 46.02% and 47.27%, respectively, indicating a consistent and well-rounded segmentation performance across different classes.

In the S3DIS-3^2 split, our framework consistently outperforms NOPS, achieving higher IoU scores for novel classes such as column (10.14%) and table (64.73%), compared to NOPS's scores of 10.06% and 53.23%, respectively. This results in an mIoU of 28.64% for novel classes in this split, slightly exceeding NOPS's 26.22%. In the S3DIS-3^3 split, which includes challenging classes such as beam, sofa, and board as novel classes, our framework achieves an IoU of 29.79% for sofa, outperforming NOPS by 7.49%. This enhanced performance is also reflected in the mIoU for novel classes, which reaches 12.22%, compared to NOPS's 9.47%. Overall, our framework consistently achieves higher mIoU scores across novel, base, and all classes than NOPS, highlighting its robustness and adaptability in accurately segmenting both familiar and novel classes in complex 3D environments.

The significant improvements in novel class segmentation can be attributed to the innovative modules integrated into our framework. The VGDI combines voxelized data with original spatial coordinates, preserving fine spatial details. The CRSM identifies key representative points within each cluster, enabling the model to capture and utilize localized spatial patterns effectively. The NSPM uses locality-sensitive hashing for precise and efficient spatial analysis. Finally, the SFAM employs multi-head attention to synthesize localized and global features, improving the model's ability to discern and segment novel classes.

While our approach demonstrates substantial improvements over the existing method, there are notable limitations and potential failure cases that warrant discussion. In certain classes, our method underperforms. For instance, in the S3DIS-4^0 split, our method achieves only 4.08% IoU for the clutter class. Similarly, in the S3DIS-3^1 split, the proposed method achieves just 12.89% IoU for the board class. Additionally, across all four splits, the beam class consistently achieves 0.00% IoU, similar to NOPS. These lower performances underscore the challenges in accurately segmenting classes that are highly irregular, sparsely represented, or easily confused due to similar shapes and proximity in many scenes. Furthermore, in certain scenes of the S3DIS dataset, the proposed method shows lower IoU than NOPS for some base classes. For example, in the S3DIS-3^1 split, our method attains an mIoU of 46.02% for base classes, compared to 47.13% for NOPS; similarly, in the S3DIS-3^2 split, our method achieves an mIoU of 53.93% for base classes, whereas NOPS reaches 54.02%.

Several underlying factors contribute to these discrepancies. First, while relying on voxelization preserves large-scale spatial information, it can still lose some fine-grained details of small or thin structures, adversely affecting classes with slender shapes or complex boundaries. This limitation may disproportionately affect certain base classes in

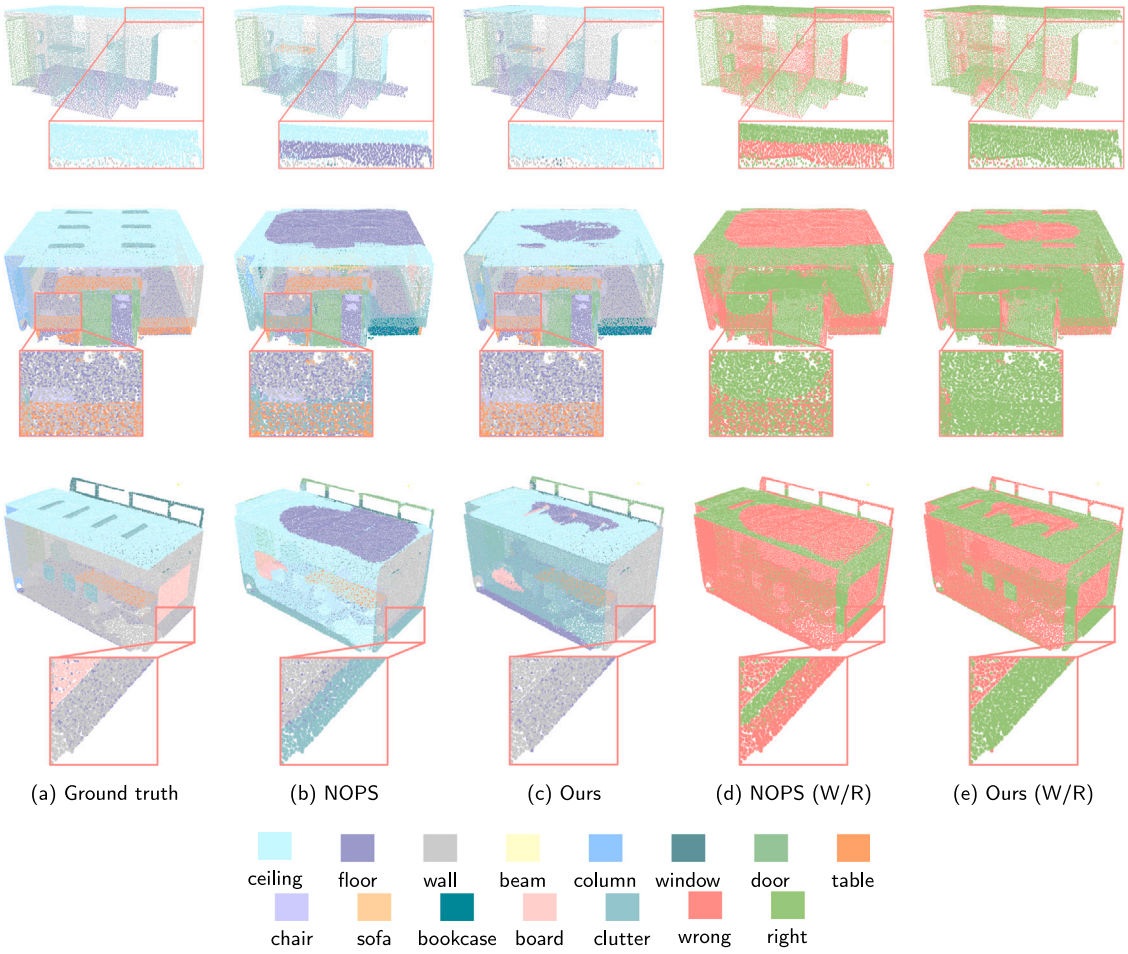

Fig. 7. Segmentation results on the S3DIS-4^0 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes as indicated by the legend. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. Zoom-in views are provided to highlight specific detailed areas. It can be observed that NOPS missegments parts of the ceiling as floor, wall as clutter, and board as wall. Our method shows improvements in these areas, although some board segments are still incorrectly identified as wall.

indoor scenes. Second, the CRSM module emphasizes finding representative points that capture the most distinguishing features for novel class discovery. This design can inadvertently reduce coverage of subtle variations within some base classes—especially those that appear infrequently or share large, uniform surfaces (e.g., walls, boards) with minimal texture cues. Third, some base classes in S3DIS appear adjacent or partially occluded. For example, sofas are often placed near walls, creating ambiguous boundaries that can hinder feature discrimination. Consequently, the method's clustering step may group certain points of the base classes suboptimally in these configurations.

These limitations highlight areas for future research and optimization. Future work could address these issues by exploring more adaptive sampling strategies that jointly consider both base and novel class distributions, which may help strike a better balance between capturing novel class features and maintaining robust performance on base classes. Additionally, adopting more granular voxelization for small or thin structures, refining the clustering mechanism, or incorporating additional class-specific constraints to preserve accurate representations of base and novel classes could further enhance the accuracy of the framework for NCD in 3D point cloud segmentation.

The visualization results in Figs. 7, 8, 9, and 10 illustrate the segmentation outcomes for different splits of the S3DIS dataset, specifically S3DIS-4^0, S3DIS-3^1, S3DIS-3^2, and S3DIS-3^3. Each figure compares the ground truth, the predictions from NOPS, and our method, along with comparisons of correctly and incorrectly segmented regions. The colors in the visualizations represent different semantic classes, and the labels "wrong" and "right" indicate regions of incorrect and correct segmentation, respectively. Overall, the visualizations clearly indicate that our method consistently achieves better segmentation results than NOPS across various splits of the S3DIS dataset, particularly in novel classes highlighted in pink.

4.6. Evaluation and visualization results on Toronto-3D

The experimental results on the Toronto-3D dataset are detailed in Table 2. The table includes various splits (Toronto-3D-2^0, Toronto-3D-2^1, Toronto-3D-2^2, Toronto-3D-2^3), with novel classes indicated in pink. The novel class splits align with the criteria used by NOPS, ensuring consistency in evaluation. The proposed method achieves notable improvements across several classes, reflecting the strengths of its innovative modules. For instance, in the Toronto-3D-2^0 split, the method outperforms NOPS significantly in the road and building classes. Specifically, the road class shows an IoU of 74.50% compared to NOPS's 52.71%, and the building class achieves 60.33% compared to NOPS's 56.48%. The overall improvement in IoU for novel classes in the Toronto-3D-2^0 split is 67.41% compared to NOPS's 54.60%, showcasing a significant advancement. This improvement is attributed to the robust integration of voxel-geometry data, which captures the fine spatial details essential for these large, continuous surfaces.

Similarly, in the Toronto-3D-2^1 split, the method excels in the natural and utility line classes, achieving IoUs of 56.03% and 61.49%, respectively, compared to NOPS's 46.84% and 45.27%. The CRSM

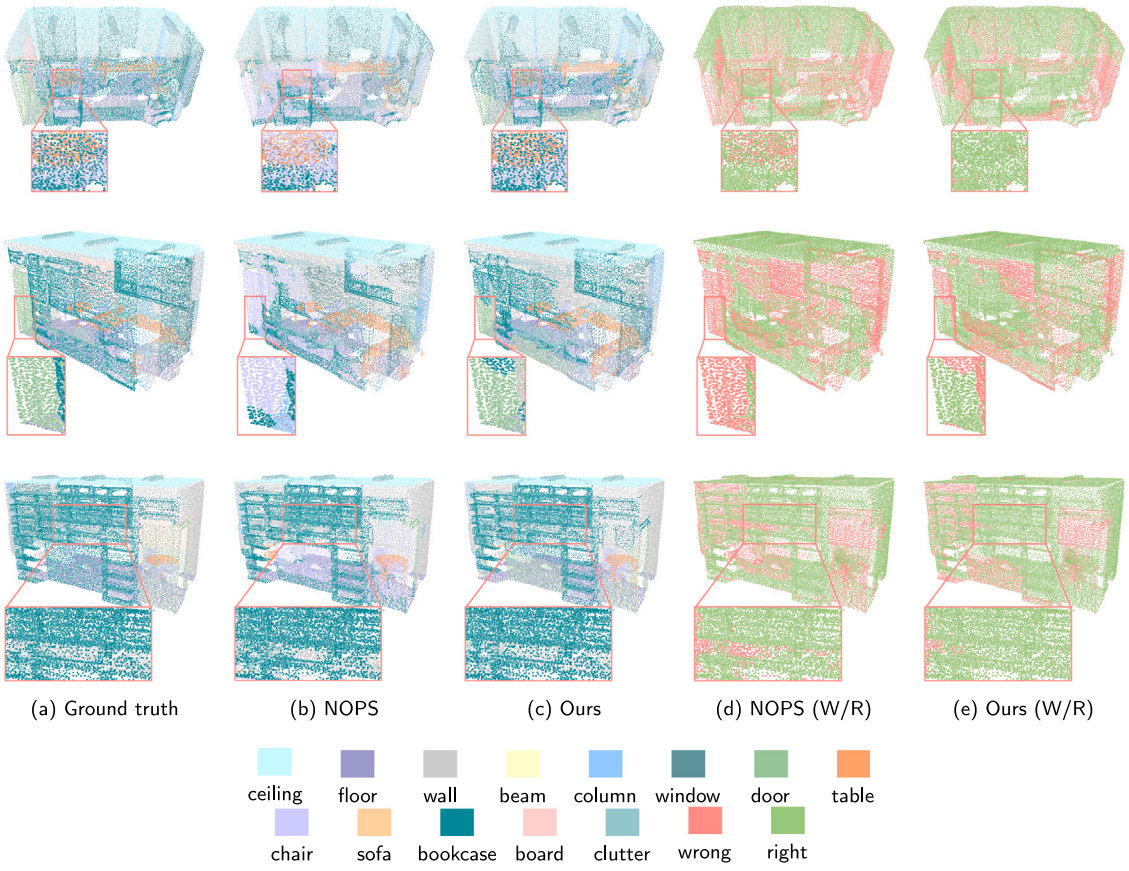

Fig. 8. Segmentation results on the S3DIS-3^1 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views display detailed segmentation results for classes such as bookcase, table, chair, and door. NOPS missegments parts of the bookcase as chair, door as chair, and bookcase as wall. Our method, in contrast, shows a reduction in these specific errors, leading to relatively better performance.

effectively identifies key representative points, enhancing the model's ability to segment objects with intricate shapes and varying geometries. In the Toronto-3D-2^2 split, the method shows superior performance in the road and utility line classes, achieving IoUs of 84.93% and 50.66%, respectively, compared to NOPS's 74.88% and 30.84%. This improvement highlights the effectiveness of the CRSM and SFAM, which enhances the model's ability to capture fine-grained spatial details. In the Toronto-3D-2^3 split, the method excels in the road and natural classes, achieving IoUs of 86.39% and 88.10%, respectively, compared to NOPS's 77.22% and 86.65%.

However, across all four splits, our method underperforms in the road marking and fence classes. These classes consistently achieve lower IoU scores, indicating a persistent challenge across different splits. The low performance in the road marking and fence classes can be attributed to several factors. First, these classes are typically underrepresented in the dataset, leading to insufficient training data for the model to learn distinguishing features effectively. Additionally, road markings are often confused with the road itself due to their close spatial proximity and similar appearance, making it difficult for the model to differentiate between them. Furthermore, road markings, being flat and often very thin, may not form distinct clusters as effectively as more volumetric objects. Fences, often thin and elongated, present similar challenges. To address these issues, future research could focus on integrating more sophisticated feature extraction techniques that capture the fine-grained details necessary for distinguishing these classes. Furthermore, the NSPM could be refined to better account for the unique characteristics of thin and elongated structures, ensuring more accurate segmentation.

The segmentation results for the Toronto-3D dataset are visualized in Figs. 11, 12, 13 and 14. These figures illustrate the segmentation performance on different splits. In these visualizations, different colors represent various semantic classes, with 'wrong' indicating areas of segmentation errors and 'right' indicating correctly segmented regions. Overall, the visualizations confirm the quantitative improvements observed in Table 2, showcasing the method's robustness and accuracy in 3D semantic segmentation across various challenging classes.

4.7. Evaluation and visualization results on SemanticSTF

The evaluation of the proposed method on the SemanticSTF dataset demonstrates significant improvements across several classes, as shown in Table 3. The pink cells in the table denote novel classes. The following examples illustrate the key areas of improvement and their underlying causes. In the SemanticSTF-5^0 split, the proposed method significantly outperforms NOPS in the road and building classes. For example, the IoU for the road class is 60.65%, compared to NOPS's 47.40%. Similarly, the building class shows an IoU of 26.68%, illustrating the method's effectiveness in handling detailed geometric structures. For the SemanticSTF-5^1 split, the proposed method shows a strong performance in the other-vehicle and fence classes, achieving IoUs of 33.55% and 37.04%, respectively, compared to NOPS's 26.72% and 30.40%. In the SemanticSTF-5^2 split, the proposed method excels in the bicyclist and bicycle classes, achieving IoUs of 64.36% and 32.38%, respectively, compared to NOPS's 41.92% and 22.51%. The SemanticSTF-4^3 split further demonstrates the method's capabilities, particularly in the building and other-ground classes, where it achieves IoUs of 72.79% and 30.31%, compared to NOPS's 67.57% and 23.83%.

However, the performance in specific classes such as motorcycle in SemanticSTF-5^0, trunk in SemanticSTF-5^1, other-vehicle in Semantic

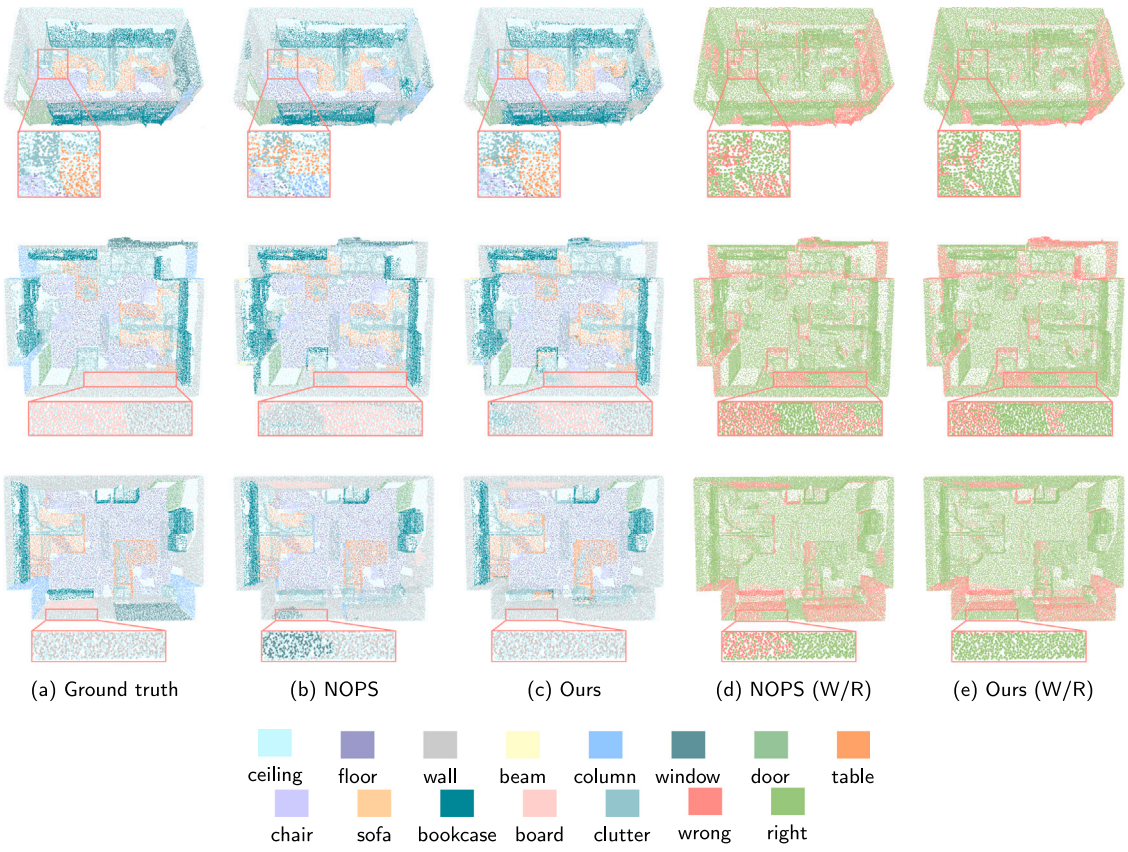

Fig. 9. Segmentation results on the S3DIS-3^2 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. Zoom-in views detailedly show the representative segmentation results for classes such as chair, table, and clutter, highlighting the improved performance of our method compared to NOPS in these classes.

STF-5^2, and motorcyclist in SemanticSTF-4^3 highlights significant challenges for the model. The lower performance in these classes can be attributed to several key factors. Firstly, the scarcity of these classes in the dataset makes it difficult for the model to learn and recognize them accurately. Secondly, the geometries of these objects are inherently smaller and visually similar, which can complicate segmentation. Moreover, a potential issue arises from the NSPM. The hashing process can lead to an over-simplification of spatial relationships, which may cause critical details to be lost. It might not capture the fine-grained spatial relationships necessary for distinguishing smaller or more similar objects. To address these limitations, future research could explore adaptive methods for the hashing process in the NSPM. By dynamically adjusting the hashing strategy based on the complexity of the local geometry, it may be possible to retain more critical spatial details. Additionally, incorporating techniques such as transfer learning from pre-trained models on larger, more diverse datasets could enhance the model's generalization capabilities.

The segmentation results for the SemanticSTF dataset are visualized in four sets of images, corresponding to the SemanticSTF-5^0, SemanticSTF-5^1, SemanticSTF-5^2, and SemanticSTF-4^3 splits (Figs. 15, 16, 17, and 18). These visualizations help illustrate the qualitative performance differences between the proposed method and NOPS. In each set of visualizations, different colors represent different semantic classes, while regions marked as "wrong" indicate areas where the segmentation was incorrect, and "right" indicates areas where the segmentation was correct.

4.8. Evaluation and visualization results on SemanticPOSS

Table 4 highlights the performance differences between the proposed method and NOPS across various splits in the SemanticPOSS dataset. Notable improvements are observed in several classes. In the SemanticPOSS-4^0 split, the proposed method significantly outperforms NOPS in the building and plant classes. The IoU for the building class is 54.83%, compared to NOPS's 30.35%, and for the plant class, it is 59.90%, compared to NOPS's 42.07%. These improvements highlight the effectiveness of the proposed method in capturing fine spatial details and complex structures. For the SemanticPOSS-3^1 split, the proposed method shows strong performance in the cone and fence classes, achieving IoUs of 28.55% and 23.54%, respectively, compared to NOPS's 12.21% and 3.94%. In the SemanticPOSS-3^2 split, the method excels in the building and people classes, achieving IoUs of 78.32% and 55.81%, respectively, compared to NOPS's 71.81% and 52.09%. This demonstrates the method's capability to effectively segment objects with intricate shapes and varying geometries. The SemanticPOSS-3^3 split further demonstrates the method's capabilities, particularly in the cone and traffic-sign classes, where it achieves IoUs of 13.66% and 30.52%, compared to NOPS's 0.00% and 26.15%.

In terms of overall performance, the proposed method consistently shows superior results. For instance, in the SemanticPOSS-4^0 split, the mIoU of novel classes is 43.68%, compared to NOPS's 35.70%. The mIoU of base classes is also higher at 28.41%, compared to NOPS's 26.57%, leading to an mIoU of all classes at 33.11%, surpassing NOPS's 29.38%. This demonstrates that the proposed method performs well not only on novel classes but also maintains high accuracy on base classes. Similarly, in the SemanticPOSS-3^1 split, the mIoU of base classes reaches 41.02%, compared to NOPS's 38.24%, showing a balanced improvement across base classes. In the SemanticPOSS-3^2 split, the mIoU of all classes is 39.44%, compared to NOPS's 36.04%, with a significant improvement in the mIoU of novel classes at 10.30%, compared to NOPS's 8.95%, and a higher mIoU of base classes at 48.18%, compared to NOPS's 44.17%. For the SemanticPOSS-3^3 split, the mIoU

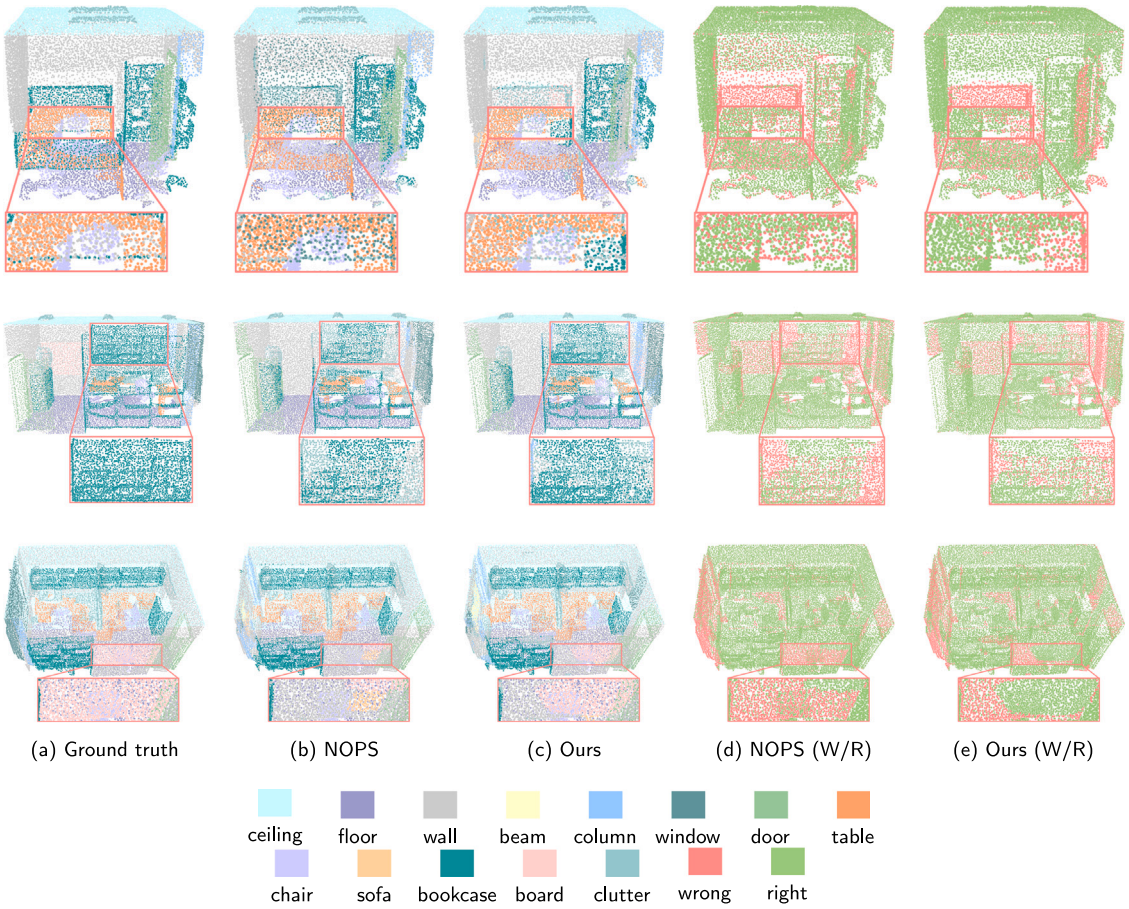

Fig. 10. Segmentation results on the S3DIS-3[3] split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views provide detailed segmentation results for classes such as table, chair, bookcase, floor, and board. It can be observed that NOPS missegments portions of the wall as bookcase in the first row, bookcase as clutter in the second row, and board as wall or sofa in the third row. Similar errors also appear in our method, but the occurrences are significantly reduced.

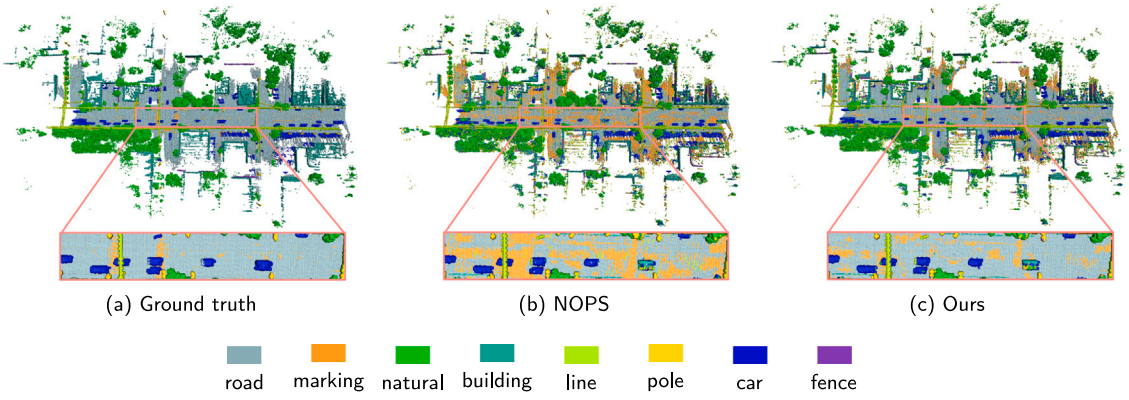

Fig. 11. Segmentation results on the Toronto-3D-2[0] split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a–c), different colors represent different semantic classes. Zoom-in views show detailed segmentation results of the road surface. It can be observed that our method more closely matches the ground truth, while NOPS incorrectly segments a larger portion of the road surface as road markings.

of all classes is 38.71%, compared to NOPS's 36.32%, demonstrating superior generalization across the majority of classes.

However, analyzing the results from the SemanticPOSS dataset, we observe specific classes where the model's performance is suboptimal, notably in the trashcan class across all splits. This discrepancy provides insights into the model's limitations and areas for future improvement. Trashcans are smaller objects with less distinctive features, making accurate segmentation challenging. Additionally, the limited number of trashcan instances in the dataset provides fewer examples for the model to learn from. A critical examination of our approach reveals that the SFAM, while effective in many scenarios, might contribute to the observed limitations. The SFAM currently employs a straightforward method for integrating localized and global features. This integration might oversimplify the complex spatial relationships necessary for accurately distinguishing smaller or more nuanced objects. To address these limitations, future research could explore more advanced techniques for feature integration within the SFAM. By incorporating

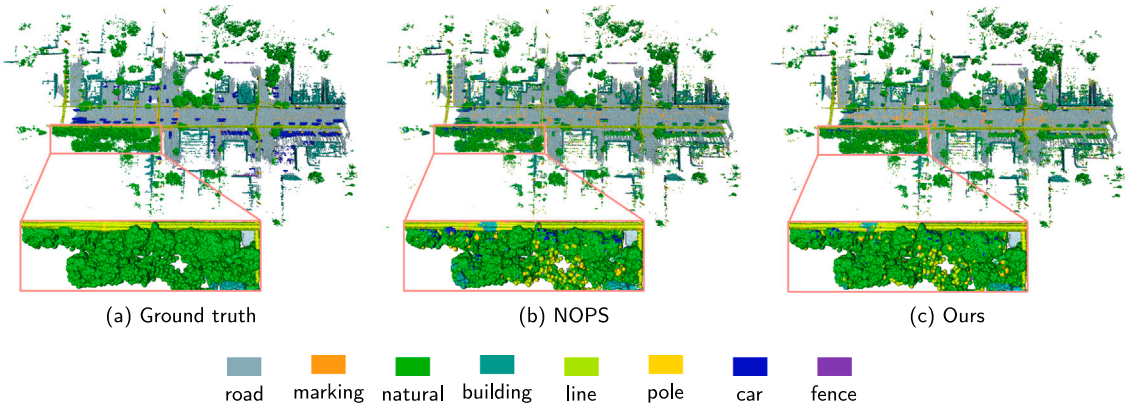

Fig. 12. Segmentation results on the Toronto-3D-2[1] split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a–c), different colors represent different semantic classes. Zoom-in views show detailed segmentation results of the natural class. It can be observed that both our method and NOPS perform well overall, but NOPS incorrectly segments more detailed areas of the natural class as buildings, utility lines, and cars.

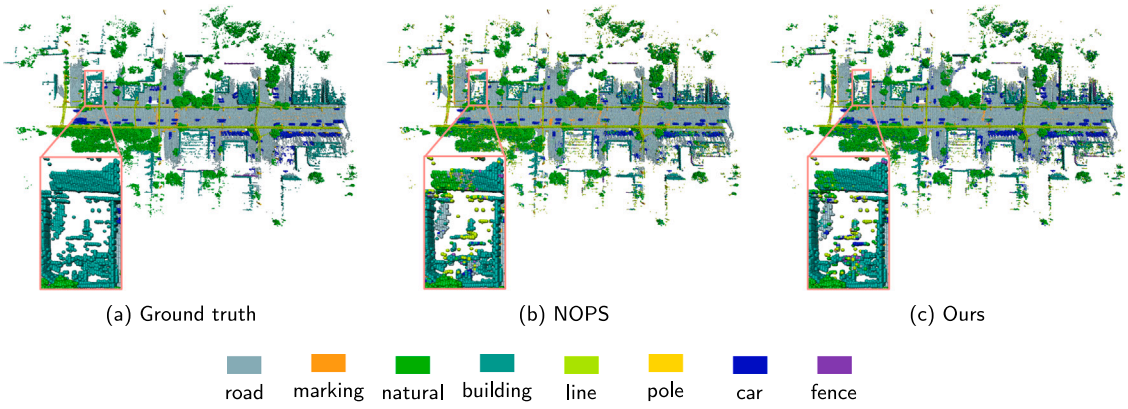

Fig. 13. Segmentation results on the Toronto-3D-2[2] split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a–c), different colors represent different semantic classes. The zoom-in views highlight the detailed segmentation results of the building class. Both our method and NOPS show good overall performance, with NOPS segmenting relatively more detailed areas of the building class as natural elements, utility lines, cars, and fences.

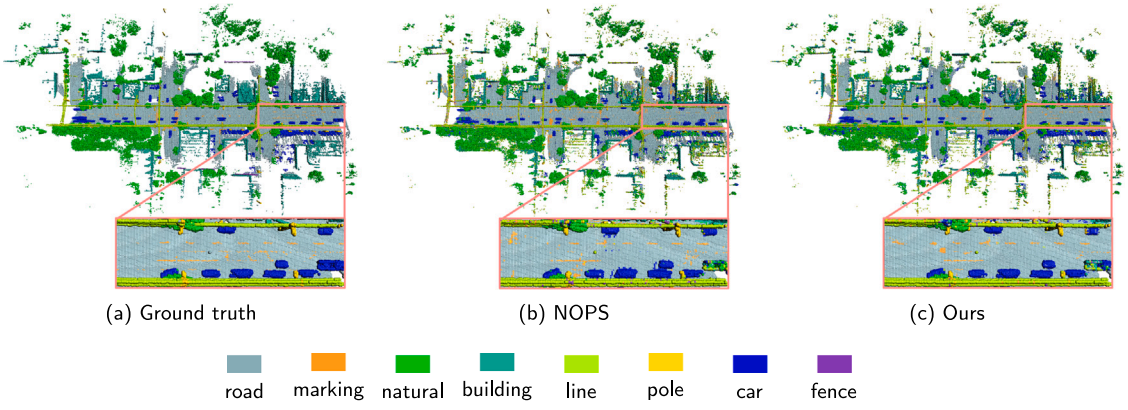

Fig. 14. Segmentation results on the Toronto-3D-2[3] split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a–c), different colors represent different semantic classes. The zoom-in views highlight the detailed segmentation results of the utility line class. Both segmentation methods show comparable overall performance; however, closer inspection reveals that our method achieves higher accuracy in finer details.

more sophisticated attention mechanisms or hierarchical feature fusion strategies, it might be possible to retain more critical spatial details.

The visualizations for the SemanticPOSS dataset across various splits, SemanticPOSS-4[0], and SemanticPOSS-3[1], are displayed in figures sequentially labeled from Figs. 19 to 20. These images serve to illustrate the nuanced differences in segmentation accuracy between the method presented and the NOPS baseline. Each visualization set provides a color-coded depiction of the segmented scenes. Colors distinctly mark different semantic classes, offering a clear visual reference to the segmentation capabilities of each approach. Incorrectly segmented areas are indicated with specific markings that highlight discrepancies, whereas correctly segment-ed areas confirm the precision of the segmentation method. These visual contrasts not only emphasize the enhanced ability of the proposed method to segment complex urban landscapes but also underline the practical implications of improved segmentation in real-world applications.

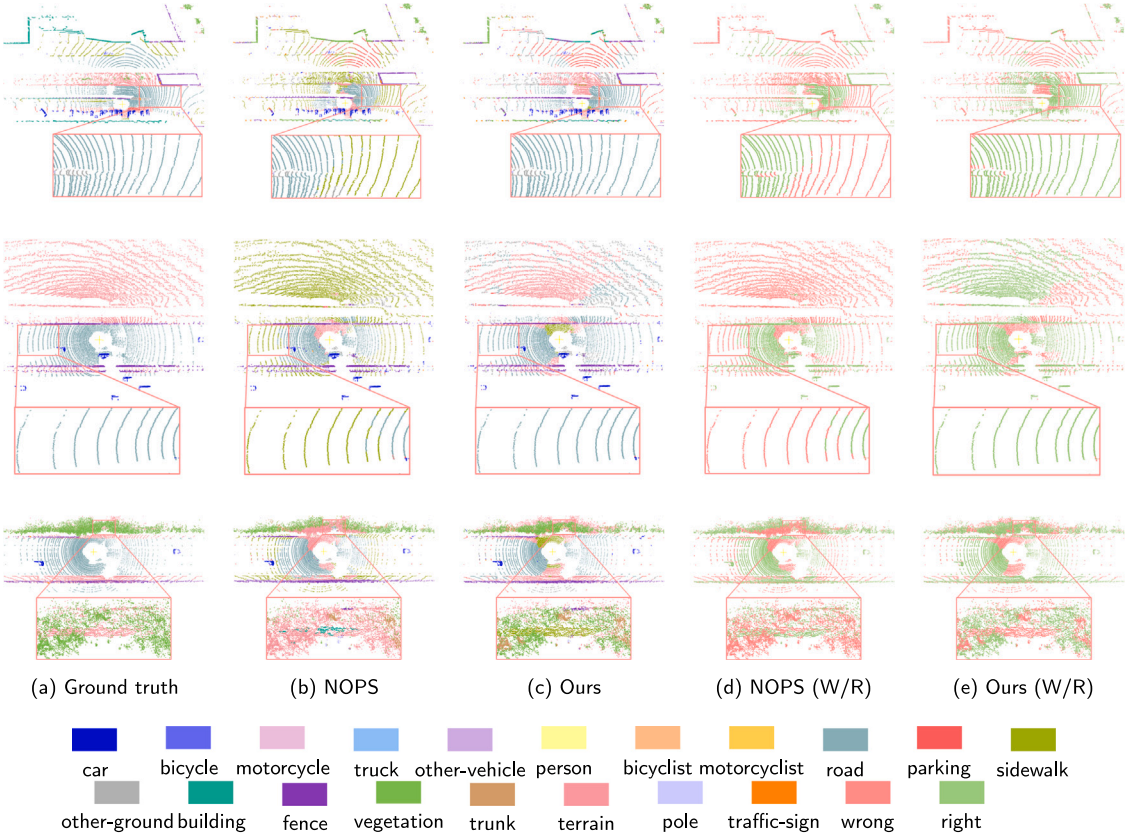

Fig. 15. Segmentation results on the SemanticSTF-5^0 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views provide detailed segmentation results for classes such as road, vegetation, and terrain. It can be observed that NOPS missegments portions of the road as sidewalk, vegetation as terrain, and terrain as building. In comparison, our method demonstrates relatively better performance, though it also incorrectly segments some terrain as sidewalk.

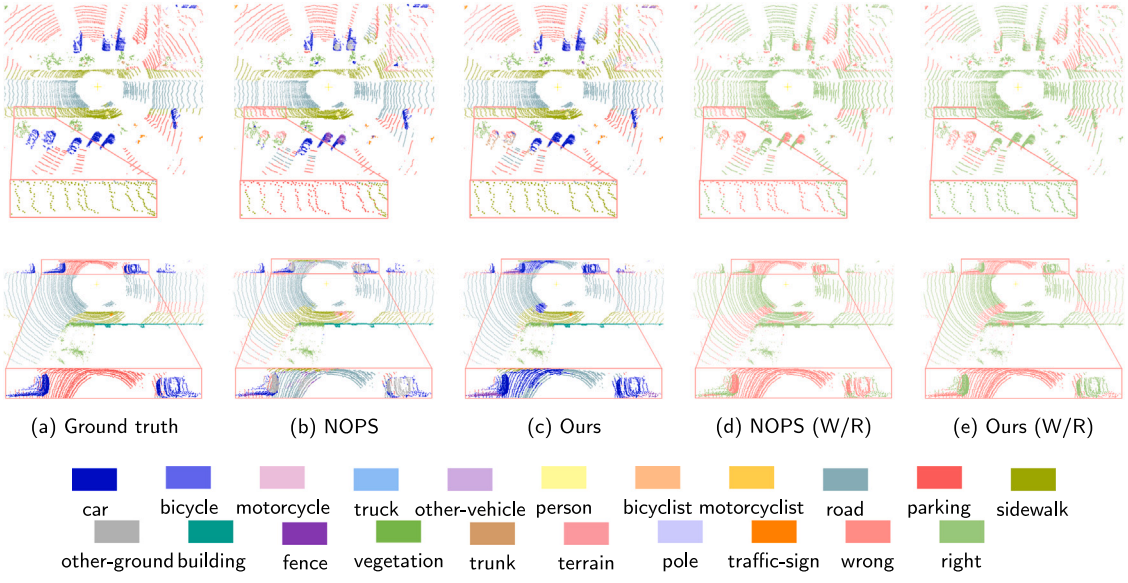

Fig. 16. Segmentation results on the SemanticSTF-5^1 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views illustrate the segmentation results for classes such as sidewalk, parking, and car. It can be observed that NOPS confuses sections of sidewalk with parking, parking with road, and car with other-ground. On the other hand, our method performs better overall, though it also makes some errors in segmenting parking as car or road.

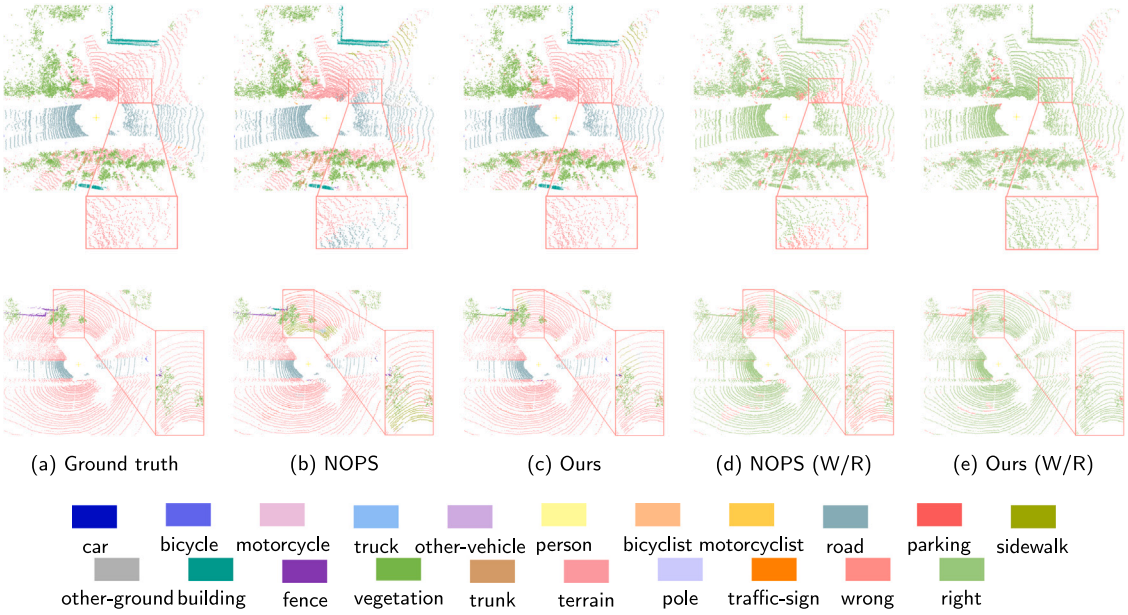

Fig. 17. Segmentation results on the SemanticSTF-5[2] split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views illustrate the segmentation outcomes for classes such as terrain and vegetation. NOPS incorrectly segments portions of terrain as road in the first row and as parking or vegetation in the second row. Our method addresses some of these segmentation errors, though it still occasionally segments portions of terrain as vegetation.

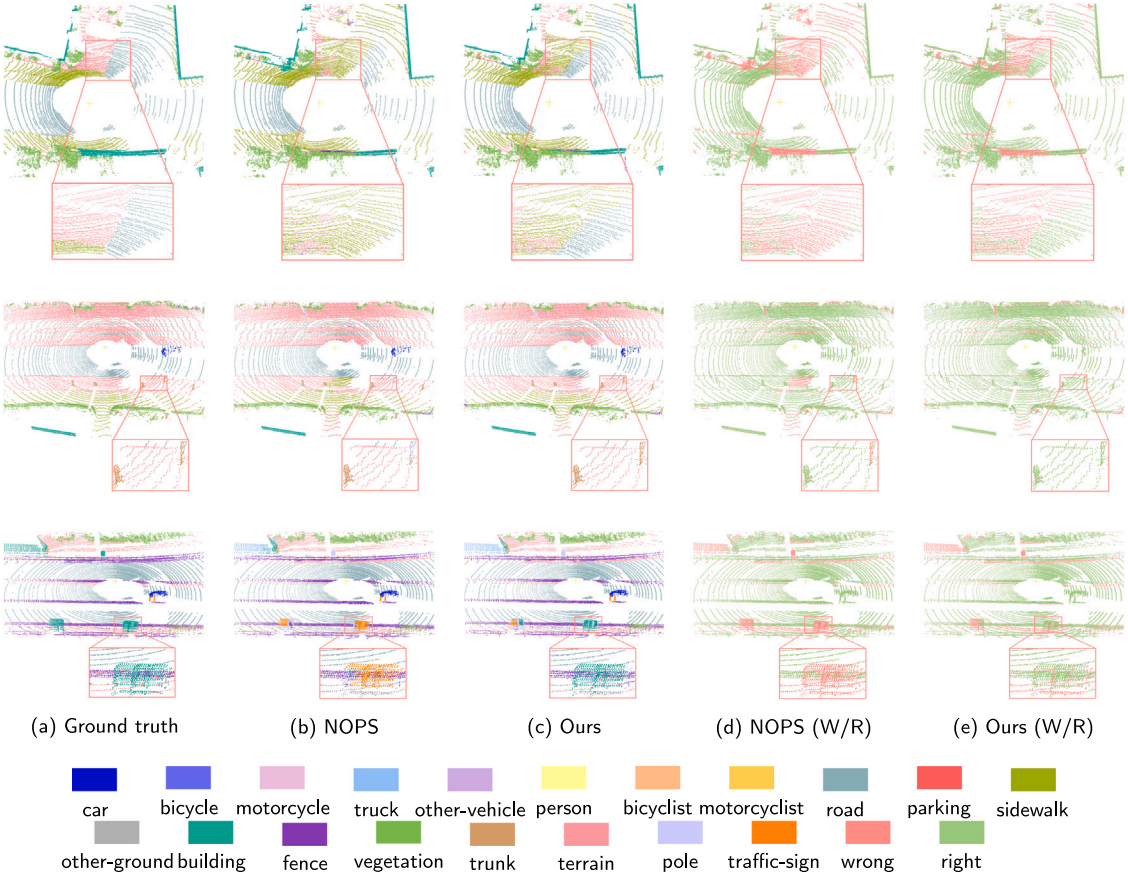

Fig. 18. Segmentation results on the SemanticSTF-4[3] split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views provide the segmentation outcomes for classes such as terrain, road, trunk, fence, and building. NOPS incorrectly segments parts of terrain and road as sidewalk in the first row, trunk as pole in the second row, and building as traffic sign in the third row. Our method addresses some of these segmentation issues, but still missegments portions of terrain as sidewalk in the first row or as other-ground in the third row.

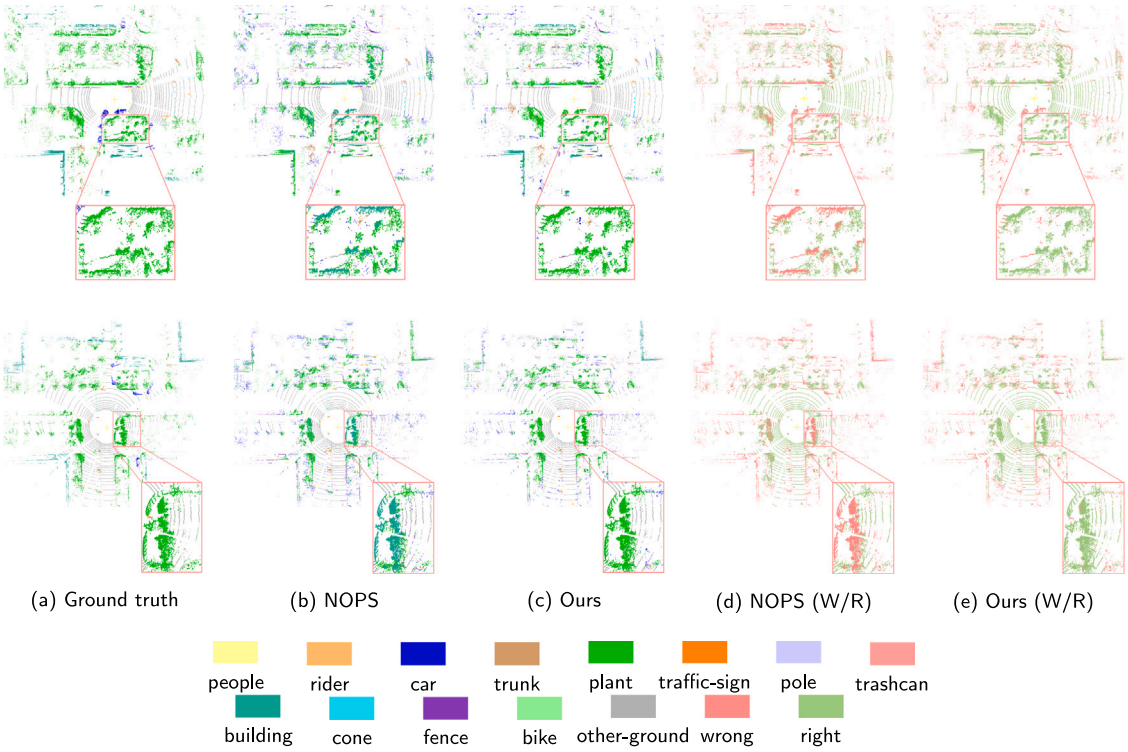

Fig. 19. Segmentation results on the SemanticPOSS-4^0 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views show detailed segmentation results. It can be observed that our method achieves better segmentation for plants, while NOPS incorrectly segments portions of plants as buildings.

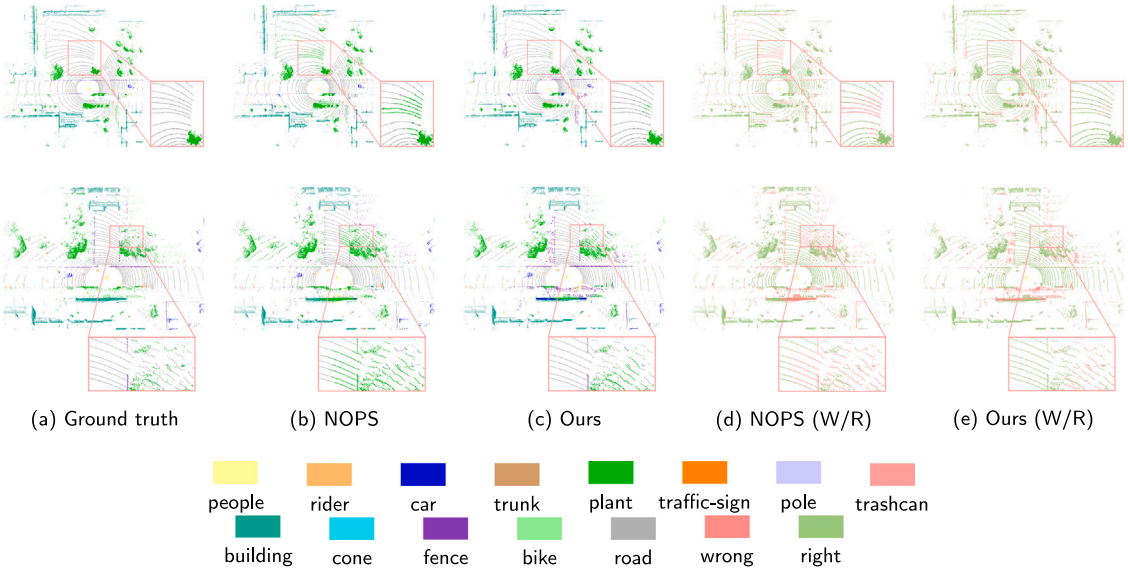

Fig. 20. Segmentation results on the SemanticPOSS-3^1 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a–c), different colors represent different semantic classes. In images (d–e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views present detailed segmentation results. It can be observed that NOPS tends to missegment the road adjacent to plants as part of the plants, while our method reduces this type of error.

4.9. Ablation study

To provide a comprehensive comparison, we conducted validation experiments on the Toronto-3D dataset with offset to evaluate the robustness and effectiveness of our proposed method in Table 5. In accordance with the Toronto-3D dataset's official guidelines, an offset was applied to prevent potential loss of detail during point cloud processing. We compared our method with NOPS across four different splits, analyzing the performance in terms of mIoU for novel, base, and all classes.

In the Toronto-3D-2^0 split, our method achieved a notable improvement in novel class performance, with the mIoU of novel classes increasing from 64.78% (NOPS) to 71.45%. Similarly, the mIoU of base classes rose from 54.00% to 58.52%, resulting in an overall increase in mIoU of all classes from 56.69% to 61.75%. Notably, in the road and building novel classes, our method outperformed NOPS with mIoU scores of 87.10% and 55.80%, compared to 79.68% and 49.88%,

Table 5
Novel class discovery results on the Toronto-3D dataset with offset (%). Pink highlighted values are the novel classes in each split. "Novel" denotes the mIoU of novel classes, "Base" indicates the mIoU of non-novel classes, and "All" shows the mIoU of all classes.

Split	Model	road	road marking	natural	building	utility line	pole	car	fence	Novel	Base	All
Toronto-3D-2^0	NOPS	79.68	11.08	91.76	49.88	71.70	64.75	63.78	20.91	64.78	54.00	56.69
	Ours	87.10	10.39	94.32	55.80	78.21	70.84	67.44	29.92	71.45	58.52	61.75
Toronto-3D-2^1	NOPS	87.92	11.26	46.96	60.14	74.67	73.34	17.17	23.65	32.07	55.16	49.39
	Ours	91.49	14.82	55.49	61.75	74.51	76.50	6.82	29.47	31.16	58.09	51.36
Toronto-3D-2^2	NOPS	91.02	4.16	93.92	57.87	74.28	56.50	70.06	28.65	30.33	69.30	59.56
	Ours	91.78	10.53	94.31	61.34	70.61	55.17	70.65	39.02	32.85	71.29	61.68
Toronto-3D-2^3	NOPS	91.03	12.51	94.75	61.86	48.61	74.16	65.92	8.08	28.35	66.71	57.12
	Ours	91.09	14.11	94.44	62.46	49.53	75.82	68.66	14.52	32.03	67.76	58.83

Table 6
Ablation study on the number of buckets m for novel class discovery on the S3DIS-4^0 split (%).

Split	m	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter	Novel	Base	All
S3DIS-4^0	32	80.36	84.66	35.21	0.00	24.45	35.37	52.58	64.19	73.66	50.22	58.62	26.74	12.63	53.22	42.87	46.05
S3DIS-4^0	64	81.46	88.12	47.83	0.00	28.33	23.89	54.83	62.24	72.28	36.34	51.88	19.74	16.30	58.42	38.84	44.86
S3DIS-4^0	128	78.75	63.92	36.84	0.00	31.26	28.07	54.11	67.07	80.40	41.76	60.75	10.73	4.08	45.90	41.57	42.90
S3DIS-4^0	256	51.17	71.37	29.28	0.00	32.47	24.80	49.82	61.44	72.04	41.03	53.25	25.29	16.69	42.13	40.02	40.67

respectively. For the Toronto-3D-2^1 split, the mIoU for base classes improved from 55.16% to 58.09%. In the natural and pole classes, our method significantly outperformed NOPS with mIoU values of 55.49% and 76.50%, compared to NOPS's 46.96% and 73.34%. In the Toronto-3D-2^2 split, our method achieved an mIoU of 32.85% for novel classes, surpassing NOPS's 30.33%, and also demonstrated better performance in base classes, with an increase in mIoU from 69.30% to 71.29%. Finally, in the Toronto-3D-2^3 split, our method continued to outperform NOPS in both novel and base classes, achieving an mIoU of 32.03% for novel classes and 67.76% for base classes, compared to 28.35% and 66.71% for NOPS. Significant improvements were observed in the utility line and fence classes, with our method achieving mIoU scores of 49.53% and 14.52%, compared to NOPS's 48.61% and 8.08%. Overall, our method consistently outperformed NOPS across different splits of the Toronto-3D dataset with offset, particularly in the novel classes and base classes. These results demonstrate the robustness and generalization capability of our approach, confirming its effectiveness in novel class discovery across diverse scenarios.

Moreover, the ablation study results in Table 6 show that increasing the number of buckets in the NSPM leads to distinct effects on novel classes and base classes. When increasing from 32 to 64 buckets, the mIoU of novel classes improves from 53.22% to 58.42%, while the mIoU of base classes drops from 42.87% to 38.84%. This suggests that finer partitioning through more buckets allows the model to better capture nuanced features in previously unseen classes, helping to distinguish novel classes. However, the reduction in the mIoU of base classes appears to result from excessive fragmentation of the feature space, leading to a loss of broader, more generalized features that are beneficial for base class recognition. The S3DIS dataset, with its rich indoor scenes, is highly structured, and the increase in buckets introduces challenges in maintaining coherent representations for the base classes.

Further increasing the bucket count to 128 and 256 demonstrates diminishing returns across both novel and base classes. For instance, the mIoU of novel classes decreases to 45.90% with 128 buckets and drops further to 42.13% with 256 buckets, suggesting over-segmentation. Similarly, the mIoU of base classes exhibits a slight recovery to 41.57% at 128 buckets, but this trend does not hold at 256 buckets. The decline in performance for the mIoU of all classes (from 46.05% at 32 buckets to 40.67% at 256 buckets) indicates that excessive granularity impairs the model's ability to capture both global and local contextual information effectively.

This phenomenon is also related to the unique characteristics of the S3DIS dataset, which represents complex indoor environments that include densely populated areas (e.g., cluttered offices or furniture-heavy rooms) as well as relatively sparse spaces (e.g., open corridors or large walls). The indoor setting is typically constrained, with many small, enclosed regions that require precise spatial partitioning for accurate class discovery. As the number of buckets increases, it appears that partitioning becomes too fine-grained, leading to reduced coherence in spatial relationships. Although increasing the number of buckets could intuitively improve fine-grained representation, the results show diminishing returns for novel class discovery beyond 64 buckets, with significant drops in performance at 128 and 256 buckets. This fragmentation disrupts the spatial coherence necessary for effective feature aggregation across both novel and base classes. Consequently, the model struggles to generalize effectively, leading to a decline in overall performance, particularly in the mIoU of all classes. This underscores the importance of carefully selecting the number of buckets to maintain a balance between granularity and spatial coherence for robust performance across all classes.

In addition, we explore the impact of varying the scaling factor γ on the performance of NCD on the SemanticPOSS dataset. The scaling factor γ controls the neighborhood size around representative points, directly affecting the segmentation processes. From the results in Table 7, we observe that increasing γ from 5 to 10 leads to an improvement in both mIoU of novel and base classes. The mIoU of novel classes increases from 34.15% to 43.68%, while the mIoU of base classes slightly rises from 27.69% to 28.41%. This suggests that enlarging the neighborhood size captures more local context, which enhances the identification of novel classes that rely on detailed spatial relationships. However, further increases in γ to 15 and 20 show diminishing returns, with both mIoU of novel and base classes experiencing fluctuations. For instance, the mIoU of novel classes decreases to 34.57% at $\gamma = 15$ and continues to drop to 30.71% at $\gamma = 20$, while the mIoU of base classes exhibits only a modest recovery.

This trend can be attributed to the characteristics of the SemanticPOSS dataset, which features large open spaces and varying object distributions. A smaller neighborhood size may fail to capture sufficient spatial context in these wide regions, while overly large neighborhoods can result in feature dilution, where important local details are lost. The fluctuation in the mIoU of all classes, from 29.68% at $\gamma = 5$, peaking at 33.11% for $\gamma = 10$, and then dropping to 30.23% at $\gamma = 20$, highlights the challenge of determining the optimal neighborhood size for both novel and base class discovery in diverse outdoor scenes. Overall, the results suggest that a moderate neighborhood size offers the best compromise between capturing features of novel classes and maintaining generalization across base classes.

Table 8 provides insight into how each module in the proposed method contributes to overall performance, particularly in the context of NCD on the SemanticSTF-5^0 dataset. We systematically analyzed the effects of modifying four key components to understand how these changes influence segmentation accuracy across novel, base, and all classes.

Firstly, replacing the original spatial coordinates in the VGDI module with voxel coordinates results in a decline in performance across all classes. The mIoU of base classes decreases from 33.64% to 30.41%,

Table 7
Ablation study on the scaling factor γ for novel class discovery on the SemanticPOSS-4^0 split (%). γ adjusts the neighborhood size around representative points.

Split	γ	bike	building	car	cone	fence	road	people	plant	pole	rider	traffic-sign	trashcan	trunk	Novel	Base	All
SemanticPOSS-4^0	5	41.78	22.40	1.25	16.57	29.17	72.46	41.93	40.49	20.24	39.86	24.56	12.12	22.99	34.15	27.69	29.68
SemanticPOSS-4^0	10	42.44	54.83	1.67	17.59	32.20	58.33	44.51	59.90	25.98	42.33	25.73	4.56	20.32	43.68	28.41	33.11
SemanticPOSS-4^0	15	42.40	27.85	2.86	25.15	32.23	69.26	43.85	38.29	19.79	44.81	30.07	4.37	23.20	34.57	29.54	31.09
SemanticPOSS-4^0	20	42.42	18.15	3.17	23.28	31.26	69.33	47.06	32.18	18.71	47.52	26.84	9.39	23.73	30.71	30.02	30.23

Table 8
Ablation study on different modules for novel class discovery on the SemanticSTF-5^0 split (%). Each row presents results for the full model and three ablated versions of the model, where key components are replaced or modified. VGDI: Voxel Coord refers to replacing the original point-based spatial coordinates with voxel coordinates in the VGDI. CRSM: Random Sampling involves replacing the representative point sampling in the CRSM with random sampling, disregarding the spatial and feature coherence of clusters. NSPM: Random Neighbors indicates that in the NSPM, the neighboring points are randomly selected instead of being chosen based on their proximity to the representative point.

Split	Model	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	Novel	Base	All
SemanticSTF-5^0	VGDI: Voxel Coord	68.75	13.60	18.60	25.05	19.24	48.12	49.08	11.61	38.58	19.19	9.67	18.12	36.62	42.28	37.72	20.62	23.59	38.02	33.40	29.24	30.41	30.10
SemanticSTF-5^0	CRSM: Random Sampling	71.35	17.59	20.34	32.11	23.39	46.65	56.18	11.16	42.42	19.07	5.00	17.76	40.60	41.75	24.68	23.10	25.92	37.18	31.21	27.72	32.06	30.92
SemanticSTF-5^0	NSPM: Random Neighbors	67.87	18.89	19.32	35.08	24.02	42.90	49.66	15.06	45.78	20.63	0.11	18.62	26.68	39.39	35.54	23.23	14.48	29.68	37.28	24.50	31.54	29.70
SemanticSTF-5^0	Original Model	59.38	23.70	14.00	37.31	35.64	43.22	61.02	22.10	60.65	23.04	13.33	23.01	26.68	42.42	23.73	23.02	23.66	28.11	35.00	29.61	33.64	32.58

Table 9
Ablation study on SFAM for novel class discovery on the SemanticSTF-5^0 split (%). Five distinct configurations are presented to clarify how each design choice in SFAM influences segmentation performance: SFAM: Single-Head Only removes multi-head attention to evaluate whether multiple heads are necessary for capturing diverse spatial features; SFAM: No $\sqrt{d_k}$ Scaling omits the usual normalization in dot-product attention to examine the importance of controlling attention magnitudes; SFAM: Average-Pooling discards the learned attention-based weighting in favor of uniformly averaging neighbor features, probing the benefit of adaptive attention; SFAM: Max-Pooling similarly replaces attention weighting but selects the maximum feature value among neighbors rather than averaging.

Split	Model	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	Novel	Base	All
SemanticSTF-5^0	SFAM: Single-Head Only	68.56	17.51	16.51	28.32	20.28	51.17	54.26	14.87	58.46	18.75	12.22	20.00	34.36	44.49	36.73	20.78	21.05	30.86	37.59	32.56	31.71	31.94
SemanticSTF-5^0	SFAM: No $\sqrt{d_k}$ Scaling	71.58	17.58	18.99	26.41	26.10	42.63	51.56	12.81	44.43	20.37	0.70	22.02	38.24	43.21	47.06	24.55	20.24	34.61	39.73	30.13	32.30	31.73
SemanticSTF-5^0	SFAM: Average-Pooling	64.64	14.59	17.64	16.97	14.48	39.67	78.79	12.73	53.48	18.04	13.70	18.37	40.20	39.41	42.35	19.68	23.93	29.80	27.95	34.73	29.48	30.86
SemanticSTF-5^0	SFAM: Max-Pooling	65.18	14.50	14.62	35.14	19.19	44.29	60.04	15.65	45.93	17.45	7.02	15.64	27.89	40.60	31.17	18.91	8.85	37.27	36.11	24.17	31.04	29.23
SemanticSTF-5^0	Original Model	59.38	23.70	14.00	37.31	35.64	43.22	61.02	22.10	60.65	23.04	13.33	23.01	26.68	42.42	23.73	23.02	23.66	28.11	35.00	29.61	33.64	32.58

and the mIoU of all classes drops from 32.58% to 30.10%. This outcome illustrates the importance of using spatial coordinates from the original point cloud. The downstream processes, such as representative point selection in CRSM, neighborhood selection in NSPM, and attention computation in SFAM, rely on the spatial accuracy provided by these coordinates. Replacing them with voxel coordinates affects these operations, leading to a noticeable impact on overall segmentation performance.

When random sampling is used to replace representative point sampling in the CRSM, the results show a decrease in the mIoU of novel classes from 29.61% to 27.72%. This drop is largely attributed to the loss of spatial coherence within clusters when points are selected randomly, diminishing the network's ability to capture the spatial relationships that are essential for identifying novel classes. The mIoU of all classes decreases from 32.58% to 30.92%, reflecting the broader impact. CRSM's effectiveness relies on selecting representative points that preserve spatial consistency, and random sampling disrupts this process.

Similarly, the NSPM also experiences performance degradation when neighboring points are selected randomly rather than based on proximity to the representative point. The mIoU of novel classes decreases from 29.61% to 24.50%, while the mIoU of all classes drops to 29.70%. Random selection disrupts the model's ability to preserve meaningful spatial relationships, which is particularly important for novel classes, where precise spatial context is needed for effective segmentation. NSPM is responsible for constructing accurate local neighborhoods, and randomization reduces the model's understanding of spatial proximity, impacting segmentation performance.

Table 9 presents the performance of four SFAM variants for novel class discovery on the SemanticSTF-5^0 split, highlighting how specific architectural or operational changes affect IoU across novel and base classes. Compared with the original model, the single-head only setting removes multi-head attention and yields moderate performance reductions across several classes. The mIoU of novel classes increases slightly from 29.61% to 32.56%, whereas the mIoU of base classes decreases from 33.64% to 31.71%, indicating that multiple heads facilitate broader base-class coverage but are not strictly necessary for novel-class improvement. The no scaling variant omits the normalization factor in dot-product attention, generally leading to a modest gain in the mIoU of novel classes (29.61% to 30.13%) but leads to a drop in the mIoU of base classes (33.64% to 32.30%), suggesting that normalization helps maintain stable representations across frequently occurring base classes. In the average-pooling condition, learned attention weights are replaced by a mean over all neighbor features; while this configuration can achieve competitive results for select novel classes, it often reduces overall segmentation quality, highlighting the benefit of adaptive weighting. The max-pooling approach exhibits a more pronounced decline, especially in the mIoU of novel classes (from 29.61% to 24.17%), suggesting that discarding the nuanced variability in spatial relationships severely hampers the model's ability to focus on subtle, fine-grained features. Collectively, these findings confirm that multi-head attention, proper scaling, and learned feature weighting each play a vital role in SFAM's effectiveness for novel class discovery.

In summary, Tables 8 and 9 demonstrate the importance of each module in the proposed method. The VGDI module's use of original spatial coordinates is crucial for maintaining spatial accuracy throughout the network, as indicated by the drop in mIoU when voxel coordinates are used. The CRSM and NSPM modules are critical for preserving spatial coherence and context, and their ablation shows how random selection impacts the network's ability to capture meaningful spatial features. Finally, the SFAM module's attention mechanism plays a vital role in the final stage of feature refinement.

5. Conclusion

Novel class discovery is an emerging research direction in 3D semantic segmentation, focusing on the segmentation of unseen classes

in point cloud data. This is crucial for applications such as autonomous navigation for vehicles and drones, where dynamically recognizing and segmenting novel classes is essential for operational safety, as well as for urban planning, where it supports more efficient city management and design. NCD enhances the robustness and adaptability of segmentation models, enabling effective processing of new and evolving classes in dynamic real-world environments.

This study introduces a novel framework for 3D semantic segmentation, specifically targeting the discovery of novel classes in complex and dynamic environments. By integrating voxel-geometry data with spatial coordinates, the proposed approach mitigates the limitations of traditional voxel-based and point-based methods, preserving fine-grained spatial details essential for accurate segmentation. The experimental results demonstrate significant improvements over the baseline method, NOPS, across various benchmark data-sets, including S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS.

The framework integrates several innovative modules, including the Voxel-Geometry Data Integration module, Cluster-based Representative Sampling module, Neighborhood Spatial Partitioning module, and Spatial Feature Attention Mechanism. These components enhance the method's capability to handle intricate spatial structures and large-scale point clouds. The experiments validate the effectiveness of these modules, showcasing notable enhancements in IoU for novel and base classes. For instance, in the S3DIS-3[1] split, the proposed method achieves an mIoU of 51.46% for novel classes, significantly surpassing NOPS's 31.53%. Similarly, in the Toronto-3D splits, the method consistently outperforms NOPS in multiple classes, demonstrating its robustness and adaptability.

The visualizations further corroborate the quantitative results, illustrating clear qualitative improvements in segmentation performance. The proposed method effectively segments complex scenes, particularly in classes such as road, car, and utility lines. These visual results provide a deeper insight into the method's performance and practical applicability, enhancing the overall understanding of its capabilities.

CRediT authorship contribution statement

Jing Du: Writing – review & editing, Writing – original draft, Visualization, Methodology, Funding acquisition. **Linlin Xu:** Writing – review & editing, Validation, Resources. **Lingfei Ma:** Writing – review & editing, Validation, Funding acquisition. **Kyle Gao:** Writing – review & editing, Validation. **John Zelek:** Writing – review & editing, Validation, Supervision, Methodology. **Jonathan Li:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42101451, and the China Scholarship Council under PhD Scholarship 202208350003.

References

Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I.K., Fischer, M., Savarese, S., 2016. 3D semantic parsing of large-scale indoor spaces. In: Proc. CVPR. IEEE, pp. 1534–1543. http://dx.doi.org/10.1109/CVPR.2016.170.

Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F., 2020. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: Proc. CVPR. IEEE, pp. 11679–11689.

Chi, H., Liu, F., Yang, W., Lan, L., Liu, T., Han, B., Niu, G., Zhou, M., Sugiyama, M., 2022. Meta discovery: Learning to discover novel classes given very limited data. In: Proc. ICLR. OpenReview.net, URL https://openreview.net/forum?id=MEpKGLsY8f.

Choy, C.B., Gwak, J., Savarese, S., 2019. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: Proc. CVPR. IEEE, pp. 3075–3084. http://dx.doi.org/10.1109/CVPR.2019.00319.

Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., Ricci, E., 2021. A unified objective for novel class discovery. In: Proc. ICCV. IEEE, pp. 9264–9272. http://dx.doi.org/10.1109/ICCV48922.2021.00915.

Gu, P., Zhang, C., Xu, R., He, X., 2023. Class-relation knowledge distillation for novel class discovery. In: Proc. ICCV. IEEE, pp. 16428–16437. http://dx.doi.org/10.1109/ICCV51070.2023.01510.

Han, K., Rebuffi, S., Ehrhardt, S., Vedaldi, A., Zisserman, A., 2020. Automatically discovering and learning new visual categories with ranking statistics. In: Proc. ICLR. OpenReview.net, URL https://openreview.net/forum?id=BJl2_nVFPB.

Han, K., Vedaldi, A., Zisserman, A., 2019. Learning to discover novel visual categories via deep transfer clustering. In: Proc. ICCV. IEEE, pp. 8400–8408. http://dx.doi.org/10.1109/ICCV.2019.00849.

Hsu, Y., Lv, Z., Kira, Z., 2018. Learning to cluster in order to transfer across domains and tasks. In: Proc. ICLR. OpenReview.net, URL https://openreview.net/forum?id=ByRWCqvT-.

Jia, X., Han, K., Zhu, Y., Green, B., 2021. Joint representation learning and novel category discovery on single- and multi-modal data. In: Proc. ICCV. IEEE, pp. 590–599. http://dx.doi.org/10.1109/ICCV48922.2021.00065.

Joseph, K.J., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., Balasubramanian, V.N., 2022a. Novel class discovery without forgetting. In: Proc. ECCV. 13684, Springer, pp. 570–586. http://dx.doi.org/10.1007/978-3-031-20053-3_33.

Joseph, K.J., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., Balasubramanian, V.N., 2022b. Spacing loss for discovering novel categories. In: Proc. CVPR. IEEE, pp. 3760–3765. http://dx.doi.org/10.1109/CVPRW56347.2022.00420.

Li, W., Fan, Z., Huo, J., Gao, Y., 2023. Modeling inter-class and intra-class constraints in novel class discovery. In: Proc. CVPR. IEEE, pp. 3449–3458. http://dx.doi.org/10.1109/CVPR52729.2023.00336.

Liu, Z., Tang, H., Lin, Y., Han, S., 2019. Point-voxel CNN for efficient 3D deep learning. In: NeurIPS. pp. 963–973, URL https://proceedings.neurips.cc/paper/2019/hash/5737034557ef5b8c02c0e46513b98f90-Abstract.html.

Pan, Y., Gao, B., Mei, J., Geng, S., Li, C., Zhao, H., 2020. SemanticPOSS: A point cloud dataset with large quantity of dynamic instances. In: Proc. IV. IEEE, pp. 687–693. http://dx.doi.org/10.1109/IV47402.2020.9304596.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. CVPR. IEEE, pp. 77–85. http://dx.doi.org/10.1109/CVPR.2017.16.

Riz, L., Saltori, C., Ricci, E., Poiesi, F., 2023. Novel class discovery for 3D point cloud semantic segmentation. In: Proc. CVPR. IEEE, pp. 9393–9402. http://dx.doi.org/10.1109/CVPR52729.2023.00906.

Roy, S., Liu, M., Zhong, Z., Sebe, N., Ricci, E., 2022. Class-incremental novel class discovery. In: Proc. ECCV. In: Lecture Notes in Computer Science, 13693, Springer, pp. 317–333. http://dx.doi.org/10.1007/978-3-031-19827-4_19.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H., 2020. PV-RCNN: point-voxel feature set abstraction for 3D object detection. In: Proc. CVPR. IEEE, pp. 10526–10535. http://dx.doi.org/10.1109/CVPR42600.2020.01054.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J., 2020. Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways. In: Proc. CVPR. IEEE, pp. 797–806. http://dx.doi.org/10.1109/CVPRW50498.2020.00109.

Troisemaine, C., Lemaire, V., Gosselin, S., Reiffers-Masson, A., Flocon-Cholet, J., Vaton, S., 2023. Novel class discovery: an introduction and key concepts. CoRR http://dx.doi.org/10.48550/ARXIV.2302.12028, arXiv:2302.12028.

Xiao, A., Huang, J., Xuan, W., Ren, R., Liu, K., Guan, D., El-Saddik, A., Lu, S., Xing, E.P., 2023. 3D semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In: Proc. CVPR. IEEE, pp. 9382–9392. http://dx.doi.org/10.1109/CVPR52729.2023.00905.

Yang, M., Wang, L., Deng, C., Zhang, H., 2023. Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In: Proc. CVPR. IEEE, pp. 3459–3468. http://dx.doi.org/10.1109/CVPR52729.2023.00337.

Yang, M., Zhu, Y., Yu, J., Wu, A., Deng, C., 2022. Divide and conquer: Compositional experts for generalized novel class discovery. In: Proc. CVPR. IEEE, pp. 14248–14257. http://dx.doi.org/10.1109/CVPR52688.2022.01387.

Zang, Z., Shang, L., Yang, S., Wang, F., Sun, B., Xie, X., Li, S.Z., 2023. Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In: Proc. ICCV. IEEE, pp. 11824–11833. http://dx.doi.org/10.1109/ICCV51070.2023.01089.

Zhao, B., Han, K., 2021. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In: NeurIPS. pp. 22982–22994.

Zhao, H., Jiang, L., Jia, J., Torr, P.H.S., Koltun, V., 2021. Point transformer. In: Proc. ICCV. IEEE, pp. 16239–16248. http://dx.doi.org/10.1109/ICCV48922.2021.01595.

Zhao, Y., Zhong, Z., Sebe, N., Lee, G.H., 2022. Novel class discovery in semantic segmentation. In: Proc. CVPR. IEEE, pp. 4330–4339. http://dx.doi.org/10.1109/CVPR52688.2022.00430.

Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N., 2021a. Neighborhood contrastive learning for novel class discovery. In: Proc. CVPR. IEEE, pp. 10867–10875. http://dx.doi.org/10.1109/CVPR46437.2021.01072.

Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N., 2021b. OpenMix: Reviving known knowledge for discovering novel visual categories in an open world. In: Proc. CVPR. IEEE, pp. 9462–9470. http://dx.doi.org/10.1109/CVPR46437.2021.00934.

Zhou, Y., Tuzel, O., 2018. VoxelNet: End-to-end learning for point cloud based 3D object detection. In: Proc. CVPR. IEEE, pp. 4490–4499.

Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D., 2021. Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation. In: Proc. CVPR. IEEE, pp. 9939–9948.