Novel Class Discovery for 3D Point Cloud Semantic Segmentation in Large-Scale Environments

by

Jing Du

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2026

© Jing Du 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Naser El-Sheimy, Professor Dept. of Geomatics Engineering, University of Calgary
Supervisor(s):	Dr. Jonathan Li, Professor Dept. of Geography & Environmental Management, University of Waterloo
Supervisor(s):	Dr. John Zelek, Professor Dept. of Systems Design Engineering, University of Waterloo
Internal Member:	Dr. David Clausi, Professor Dept. of Systems Design Engineering, University of Waterloo
Internal Member:	Dr. Nasser Lashgarian Azad, Professor Dept. of Systems Design Engineering, University of Waterloo
Internal-External Member:	Dr. Michael A. Chapman, Professor Dept. of Civil Engineering (Adjunct GEM), Toronto Metropolitan University

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This doctoral thesis is accomplished in the manuscript option, following the graduation guidelines administered by the Graduate Program in Systems Design Engineering. It consists of three required manuscripts, which form the main body of Chapters 3, 4 and 5 of this thesis. These manuscripts contain material from three multi-authored papers, where the manuscripts are minimally changed for formatting consistency. As the lead author of these three papers, I was responsible for the conceptualization, methodology design, code implementation, validation experiments, data curation, investigation, formal analysis, visualization, and writing of both the original draft and subsequent revisions. My supervisors, Prof. John Zelek and Prof. Jonathan Li, provided overall supervision, methodological guidance, and constructive feedback on the manuscripts throughout the research and writing process. My co-authors contributed to the manuscript review. The references for the three papers are provided below:

(1) **Jing Du**, Linlin Xu, Lingfei Ma, Kyle Gao, John Zelek, Jonathan Li. 3D semantic segmentation: Cluster-based sampling and proximity hashing for novel class discovery[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 223: 274-295.

(2) **Jing Du**, John Zelek, Dedong Zhang, Jonathan Li. Gated multi-source fusion with geometric sequence modeling for novel urban structure discovery[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 230: 495-523.

(3) **Jing Du**, John Zelek, Dedong Zhang, Jonathan Li. 3DLCDM: Hybrid supervision for land cover discovery mapping of emerging urban structures in 3D remote sensing[J]. *Remote Sensing of Environment*, 2025, 331: 115018.

In addition to the three key manuscripts highlighted, my other publications have also provided a solid foundation and valuable experimental experience for this research topic. These works have not only contributed to the development of essential methodologies and insights but have also served as important references for understanding the background and related work, enriching the contextual framework of this thesis. The references are provided below:

(1) **Jing Du**, John Zelek, Jonathan Li. Weather-aware autopilot: Domain generalization for point cloud semantic segmentation in diverse weather scenarios[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 218: 204-219.

(2) **Jing Du**, Lingfei Ma, Jing Li, Nannan Qin, John Zelek, Haiyan Guan, Jonathan Li. RdmkNet & Toronto-Rdmk: Large-scale datasets for road marking classification and segmentation[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2024, 25(10): 13467-13482.

(3) **Jing Du**, Guorong Cai, Zongyue Wang, Jinhe Su, Min Huang, John Zelek, José Marcato Junior, Jonathan Li. MTCloud: Multi-type convolutional linkage network for point cloud instance segmentation[J]. *Expert Systems with Applications*, 2025, 270: 126432.

I understand that my thesis may be made electronically available to the public.

Abstract

Modern urban environments undergo continuous transformation as emerging infrastructure appears worldwide. Traditional semantic segmentation methods for 3D point clouds operate on fixed taxonomies, producing static representations that cannot adapt to novel categories. This dissertation addresses novel class discovery (NCD) in large-scale 3D point cloud segmentation through geometry-aware mechanisms, adaptive multi-source fusion, and hybrid supervision frameworks.

The first study establishes geometric foundations through voxel-geometry integration with region-centric organization, termed CHNCD. The framework couples voxel representations with original spatial coordinates via index mapping, identifies semantically informative points within clusters, accelerates neighbor retrieval through proximity hash mapping, and consolidates localized features with global context via spatial attention. Experiments on S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS demonstrate consistent improvements over discovery baselines.

The second study deepens representation through adaptive geometric sequence modeling, dynamic Gaussian embeddings, and gated multi-source fusion, termed AGDNet. Adaptive geometric sequence modeling employs learnable dimension weighting and dynamic grouping adjusted to local point density. Dynamic Gaussian embeddings represent point clouds as 3D Gaussians and compute Mahalanobis distances to generate multi-scale spatial embeddings. Gated multi-source fusion intelligently weights features through context-aware mechanisms. Three knowledge-transfer objectives operate at category, instance, and distribution levels to bridge semantic gaps. Evaluation on Toronto-3D, SemanticSTF, and SemanticPOSS demonstrates substantial improvements.

The third study integrates discovery with operational land cover mapping, termed 3DLCDM. The framework processes features through a supervised head for established categories and a dual unsupervised head comprising a primary branch with fixed prototypes and an over-segmentation branch with progressive scheduling. Temporal Sinkhorn-Knopp normalization with adaptive temperature scheduling stabilizes pseudo-labels, while dynamic weighting combines per-batch and global frequency statistics to address class imbalance. Evaluation on DALES and H3D datasets demonstrates substantial improvements for continuous land cover discovery mapping.

Taken together, the three studies advance a progressive research agenda unifying discovery and 3D segmentation for large-scale point cloud scenes. The dissertation demonstrates consistent gains across six benchmark datasets, exhibits generalization across sensors and acquisition geometries, and provides a principled route to maintain updated urban maps as new structures emerge.

Acknowledgements

Completing my doctoral studies has been a long and transformative journey, and I am deeply grateful to the many people who have supported and inspired me throughout this time.

I wish to express my deepest gratitude to my supervisors, Professor Jonathan Li and Professor John Zelek, for their generous guidance, encouragement, and support throughout my doctoral studies. Their expertise, patience, and dedication to research excellence have profoundly shaped both this dissertation and my development as a researcher. I am truly grateful for their insights, high academic standards, and the trust they have placed in me.

I am deeply grateful to my thesis committee members: Professor Naser El-Sheimy, Professor David Clausi, Professor Nasser Lashgarian Azad, and Professor Michael A. Chapman, for the time they devoted to reviewing my dissertation and supporting my doctoral examination, as well as for the thoughtful comments and suggestions they provided. Their expertise and perspectives have significantly strengthened the clarity, depth, and rigor of this work.

My heartfelt thanks go to my colleagues and friends in the laboratory, including William Xu, Willow, Jingxiang Li, Hongjie He, Kyle Gao, Liangzhi Li, Hongzhang Xu, Weikai Tan, Lanying Wang, Dedong Zhang, and Wentao Sun. I truly appreciate the encouragement and support they have offered throughout my research journey. Working alongside them has made the challenges of doctoral study both more meaningful and more enjoyable.

I wish to acknowledge Dening Lu for his assistance during the concluding phase of my Ph.D. His help with several practical matters allowed me to concentrate more fully during this demanding period. I am sincerely thankful for his reliability and support.

I am profoundly grateful to my grandparents, whose love, sacrifice, and guidance shaped the foundation of my life. Their dedication to raising me is a kindness I will carry with me forever. I am deeply thankful to my mother for her many years of hard work, strength, and unwavering support. I would also like to thank my younger sister, and I sincerely hope that we will both continue to move toward a bright and fulfilling future.

As I reach the end of my doctoral journey, I am grateful for the many remarkable individuals whose support has guided me throughout this long path.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract vi

Acknowledgements vii

List of Figures xiii

List of Tables xvi

List of Abbreviations xviii

1 Introduction 1

1.1 Background and Motivation 1

1.2 Challenges and Research Questions 3

1.3 Research Objectives and Approach 4

1.4 Thesis Structure 5

2 Novel Class Discovery and Land Cover Mapping **7**
2.1 Novel Class Discovery . . . 7
2.1.1 Early Foundations of Novel Class Discovery . . . 7
2.1.2 Architectures and Learning Strategies for NCD . . . 8
2.1.3 Contrastive Objectives and Unified Label-Space Formulations . . . 9
2.1.4 Meta-learning and Incremental Discovery . . . 9
2.1.5 Extension to Complex Domains . . . 10
2.1.6 Novel Class Discovery for 3D Point Clouds . . . 10
2.2 Deep Learning for Land Cover Mapping . . . 11
2.2.1 2D Semantic Segmentation for Land Cover Mapping . . . 11
2.2.2 3D Point Cloud Segmentation for Land Cover Mapping . . . 13
2.3 Chapter Summary . . . 13

3 CHNCD: Geometry-aware Novel Class Discovery **15**
3.1 Introduction . . . 15
3.2 Method . . . 17
3.2.1 Voxel-geometry Data Integration . . . 17
3.2.2 Cluster-based Representative Sampling . . . 20
3.2.3 Neighborhood Spatial Partitioning . . . 21
3.2.4 Spatial Feature Attention Mechanism . . . 24
3.3 Experiments . . . 27
3.3.1 Datasets . . . 27
3.3.2 Dataset Preparation . . . 28
3.3.3 Experimental Setup . . . 30
3.3.4 Evaluation and Visualization Results on S3DIS . . . 30
3.3.5 Evaluation and Visualization Results on Toronto-3D . . . 38
3.3.6 Evaluation and Visualization Results on SemanticSTF . . . 41
3.3.7 Evaluation and Visualization Results on SemanticPOSS . . . 45
3.3.8 Ablation Study . . . 50
3.4 Chapter Summary . . . 55

4 AGDNet: Gated Multi-Source Fusion and Semantic Transfer 56
4.1 Introduction . . . 56
4.2 Method . . . 59
4.2.1 Method Overview . . . 59
4.2.2 Adaptive geometric sequence modeling . . . 60
4.2.3 Dynamic Gaussian Embedding . . . 67
4.2.4 Gated Multi-source Feature Fusion . . . 69
4.2.5 Novel Class Discovery Implementation . . . 70
4.2.6 Multi-objective Learning for Novel Class Discovery . . . 72
4.3 Experiments . . . 74
4.3.1 Datasets . . . 74
4.3.2 Implementation Details . . . 76
4.3.3 Training Settings . . . 76
4.3.4 Evaluation and Visualization Results on Toronto-3D . . . 77
4.3.5 Evaluation and Visualization Results on SemanticSTF . . . 84
4.3.6 Evaluation and Visualization Results on SemanticPOSS . . . 89
4.3.7 Ablation Study on Framework Modules . . . 97
4.3.8 Ablation Study on AGSM . . . 97
4.3.9 Ablation Study on DGE . . . 99
4.3.10 Ablation Study on Knowledge Transfer Objectives . . . 101
4.3.11 Feature Discriminability Analysis . . . 102
4.3.12 Computational Efficiency Analysis . . . 102
4.4 Discussion . . . 105
4.4.1 Framework Performance Analysis . . . 105
4.4.2 Practical Deployment Considerations . . . 106
4.5 Chapter Summary . . . 106

5 3DLCDM: Hybrid Supervision for Land Cover Discovery Mapping 108
5.1 Introduction . . . 108
5.2 Study Area and Dataset Processing . . . 112
5.2.1 Study Area . . . 112
5.2.2 Data Preprocessing and Block Partitioning for DALES . . . 115
5.2.3 Data Preprocessing and Block Partitioning for H3D . . . 117
5.2.4 Base-novel Class Splitting . . . 121
5.3 Method . . . 123
5.3.1 Network Architecture . . . 123
5.3.2 Progressive Dual-branch Prototype Learning . . . 125
5.3.3 Temporal Sinkhorn-Knopp Balancing . . . 128
5.3.4 Dynamic Weighting Loss . . . 130
5.4 Experiments . . . 133
5.4.1 Experimental Configuration and Train-test Partitioning . . . 133
5.4.2 Experimental Protocol and Real-world Equivalence . . . 134
5.4.3 Evaluation and Visualization Results on DALES . . . 135
5.4.4 Evaluation and Visualization Results on H3D . . . 142
5.4.5 Comparison with Traditional Semantic Segmentation Methods . . . 146
5.4.6 Impact of Backbone Architecture on LCDM Performance . . . 147
5.4.7 Evaluation on SemanticPOSS Dataset with Alternative Splitting Strategies . . . 148
5.4.8 Ablation Study on Key Components . . . 149
5.4.9 Evaluation of Alternative Base-novel Class Splitting Strategies . . . 151
5.4.10 Evaluation of Dataset-specific Partitioning Strategies . . . 153
5.4.11 Computational Efficiency Analysis . . . 154
5.5 Discussion . . . 155
5.5.1 Temporal Implications for LCDM . . . 155
5.5.2 Practical Applications in Environmental Remote Sensing . . . 156
5.6 Chapter Summary . . . 157

6 Conclusions and Future Work **158**

6.1 Conclusions 158

6.1.1 Contributions 160

6.2 Future Work 160

6.2.1 Large Language Model Reasoning for Contextual Scene Understanding 160

6.2.2 Multimodal Fusion for Robust Discovery Under Heterogeneous Data 161

6.2.3 Agent-Based Discovery for Continuous Urban Intelligence 162

6.2.4 Scalability to Real-Time Systems and Deployment in Dynamic Environments 162

References **164**

APPENDICES **176**

A Evaluation metrics **177**

B Hyperparameter Configurations **179**

C Copyright Acknowledgment for IEEE Materials **181**

List of Figures

1.1 Organization of the thesis. 5

3.1 Framework of the proposed network. 18

3.2 Illustration of the Proximity Hash Mapping (PHM) process. 21

3.3 Histograms illustrating the number of points for each class on the S3DIS dataset. 29

3.4 Histograms illustrating the number of points for each class on the Toronto-3D dataset. 30

3.5 Histograms illustrating the number of points for each class on the SemanticSTF dataset. 31

3.6 Histograms illustrating the number of points for each class on the SemanticPOSS dataset. 32

3.7 Segmentation results on the S3DIS-4^0 split. 33

3.8 Segmentation results on the S3DIS-3^1 split. 35

3.9 Segmentation results on the S3DIS-3^2 split. 36

3.10 Segmentation results on the S3DIS-3^3 split. 37

3.11 Segmentation results on the Toronto-3D-2^0 split. 39

3.12 Segmentation results on the Toronto-3D-2^1 split. 39

3.13 Segmentation results on the Toronto-3D-2^2 split. 40

3.14 Segmentation results on the Toronto-3D-2^3 split. 40

3.15 Segmentation results on the SemanticSTF-5^0 split. 42

3.16 Segmentation results on the SemanticSTF-5^1 split. 44

3.17 Segmentation results on the SemanticSTF-5^2 split. 45
3.18 Segmentation results on the SemanticSTF-4^3 split. 46
3.19 Segmentation results on the SemanticPOSS-4^0 split. 48
3.20 Segmentation results on the SemanticPOSS-3^1 split. 49

4.1 Overview of the proposed 3D novel class discovery network AGDNet. . . . 59
4.2 Framework of the proposed Adaptive Geometric Sequence Modeling module. 61
4.3 Framework of the proposed Dynamic Gaussian Embedding module. 66
4.4 Framework of the proposed GMSFF module. 68
4.5 Segmentation results on Toronto-3D-2^0 split. 78
4.6 Segmentation results on Toronto-3D-2^1 split. 79
4.7 Segmentation results on Toronto-3D-2^2 split. 80
4.8 Segmentation results on Toronto-3D-2^3 split. 81
4.9 Segmentation results on SemanticSTF-5^0 split. 87
4.10 Segmentation results on SemanticSTF-5^1 split. 88
4.11 Segmentation results on SemanticPOSS-4^0 split. 90
4.12 Segmentation results on SemanticPOSS-3^1 split. 91
4.13 Segmentation results on SemanticPOSS-3^2 split. 92
4.14 Segmentation results on SemanticPOSS-3^3 split. 93
4.15 Feature discriminability comparison using t-SNE visualization on SemanticPOSS-4^0 split. 103

5.1 Comparison of traditional semantic segmentation methods and our unified end-to-end framework for land cover discovery mapping. 109
5.2 Geographic distribution and dataset characteristics of the DALES and H3D datasets used in this study. 113
5.3 Illustration of the block partitioning results for the DALES dataset. 114
5.4 Illustration of the block partitioning results for the H3D dataset. 116
5.5 Point count distribution for each class in the processed DALES dataset. . . 117
5.6 Point count distribution for each class in the processed H3D dataset. . . . 118

5.7 Class splitting in the DALES dataset. 119
5.8 Class splitting in the H3D dataset. 120
5.9 Network architecture of the proposed 3DLCDM framework for land cover discovery mapping. 122
5.10 Dual-branch prototype learning architecture with progressive prototype scheduling. 124
5.11 Illustration of how the prototype dimension $\rho(t)$ evolves over 100 epochs. . 125
5.12 Temporal temperature scheduling mechanism for robust Sinkhorn-Knopp pseudo-label generation. 127
5.13 Temporal schedule of the temperature parameter $\mathcal{T}(t)$ used in the Sinkhorn–Knopp balancing module. 128
5.14 Dynamic weighting strategy for addressing class imbalance in urban point cloud analysis. 131
5.15 Visualization of land cover discovery mapping results on the DALES-2^0 split, where ground and pole are defined as the novel classes. 137
5.16 Visualization of land cover discovery mapping results on the DALES-2^1 split, where vegetation and truck are defined as the novel classes. 138
5.17 Visualization of land cover discovery mapping results on the DALES-2^2 split, where building and power line are defined as the novel classes. 139
5.18 Visualization of land cover discovery mapping results on the DALES-2^3 split, where car and fence are defined as the novel classes. 140
5.19 Visualization of land cover discovery mapping results on the H3D dataset for four splits. 144

List of Tables

1.1 Comparison of learning paradigms for handling novel categories. 2

3.1 Novel class discovery results on the S3DIS dataset (%). 31
3.2 Novel class discovery results on the Toronto-3D dataset without offset (%). 38
3.3 Novel class discovery results on the SemanticSTF dataset (%). 43
3.4 Novel class discovery results on the SemanticPOSS dataset (%). 47
3.5 Novel class discovery results on the Toronto-3D dataset with offset (%). . . 50
3.6 Ablation study on the number of buckets m for novel class discovery on the S3DIS-4^0 split (%). 51
3.7 Ablation study on the scaling factor γ for novel class discovery on the SemanticPOSS-4^0 split (%). 52
3.8 Ablation study on different modules for novel class discovery on the SemanticSTF-5^0 split (%). 52
3.9 Ablation study on SFAM for novel class discovery on the SemanticSTF-5^0 split (%). 54

4.1 Detailed network architecture of the proposed framework. 75
4.2 Novel class discovery results on Toronto-3D dataset with offset(%). 77
4.3 Novel class discovery results on SemanticSTF dataset (%). 84
4.4 Novel class discovery results on SemanticPOSS dataset (%). 89
4.5 Ablation study on Toronto-3D-2^0. 96
4.6 Ablation study of AGSM module components on Toronto-3D-2^1 split (%) . 97
4.7 Ablation study of DGE module components on SemanticPOSS-3^2 split (%). 99

4.8 Ablation study of knowledge transfer objectives on SemanticPOSS-3^3 split (%). 100
4.9 Comprehensive computational efficiency comparison across three datasets. 104

5.1 Land cover discovery mapping results on DALES dataset (%). 135
5.2 Land cover discovery mapping results on H3D dataset (%). 142
5.3 Performance analysis on DALES dataset (%). 146
5.4 Land cover discovery mapping results on SemanticPOSS dataset (%). . . . 148
5.5 Ablation study of key components in 3DLCDM on DALES-2^0 split (%). . . 149
5.6 Ablation study on base-novel class splitting strategies using DALES dataset (%). 150
5.7 Ablation study on partitioning strategies for DALES-2^0 split (%). 152
5.8 Ablation study on partitioning strategies for H3D-3^0 split (%). 152
5.9 Comprehensive computational efficiency comparison on DALES and H3D datasets with different partitioning strategies. 154

B.1 Hyperparameter configuration for CHNCD. 179
B.2 Hyperparameter configuration for AGDNet. 180
B.3 Hyperparameter configuration for 3DLCDM. 180

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AIO	All-in-One
AGSM	Adaptive Geometric Sequence Modeling
ALS	Airborne Laser Scanning
BYOP	Bring Your Own Prior
C2Seg	Cross-City Semantic Segmentation
CASID	Climate-Aware Satellite Image Dataset
CNN	Convolutional Neural Network
ComEx	Compositional Experts
CRF	Conditional Random Field
CRSM	Cluster-based Representative Sampling Module
DALES	Dayton Annotated LiDAR Earth Scan
DEC	Deep Embedded Clustering
DGE	Dynamic Gaussian Embedding
DTC	Deep Transfer Clustering
EUMS	Entropy-based Uncertainty Modeling and Self-training
FRoST	Feature Replay and Distillation with Self-Training
GNCD	Generalized Novel Class Discovery
GMSFF	Gated Multi-Source Feature Fusion
GPS	Global Positioning System
GPU	Graphics Processing Unit
HighDAN	High-Resolution Domain Adaptation Network
IMU	Inertial Measurement Unit
IoU	Intersection over Union
kNN	k-Nearest Neighbors
LCDM	Land Cover Discovery Mapping

LiDAR	Light Detection and Ranging
MAML	Model-Agnostic Meta-Learning
MLP	Multi-Layer Perceptron
MLS	Mobile Laser Scanning
mIoU	Mean Intersection over Union
NCD	Novel Class Discovery
NCDSS	Novel Class Discovery for Semantic Segmentation
NCDwF	Novel Class Discovery without Forgetting
NCL	Neighborhood Contrastive Learning
NOPS	NOvel Point Segmentation
NSPM	Neighborhood Spatial Partitioning Module
ODA	Open-set Domain Adaptation
PHM	Proximity Hash Mapping
ProtoNet	Prototypical Networks
PV-RCNN	PointVoxel-RCNN
PVCNN	Point-Voxel Convolutional Neural Network
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
SAR	Synthetic Aperture Radar
S3DIS	Stanford Large-Scale 3D Indoor Spaces
SCL	Soft Contrastive Learning
SFAM	Spatial Feature Attention Mechanism
sKLD	Symmetric Kullback–Leibler Divergence
UAV	Unmanned Aerial Vehicle
VGDI	Voxel-Geometry Data Integration
VHR	Very High Resolution
WTA	Winner-Take-All

Chapter 1

Introduction

Real-world 3D environments evolve continually, with new structures and object categories emerging across autonomous driving, robotics, urban planning, and environmental monitoring. Conventional 3D semantic segmentation assumes a fixed label taxonomy, yielding static semantic maps that cannot adapt when previously unseen categories are encountered at deployment [68, 109, 110, 104, 17, 15, 14, 16, 62]. This dissertation develops a unified research program for NCD in 3D point-cloud semantic segmentation, linking geometry-aware organization, adaptive multi-source modeling, and hybrid supervision to support continuously updated mapping in large-scale urban settings. The chapter motivates the problem, articulates the challenges and research questions, outlines the research objectives and overall approach, and concludes with the thesis structure.

1.1 Background and Motivation

Traditional 3D semantic segmentation pipelines presuppose that all semantic categories are known a priori [68, 109, 110, 104]. NCD challenges this assumption by classifying samples from unlabeled data into novel classes using knowledge transferred from labeled *base* classes [32]. In this setting, base classes refer to annotated categories available during training, whereas *novel* classes arise only in unlabeled data. For 3D point clouds, NCD assigns each point either a base or a novel label, enabling adaptive semantic mapping of evolving environments [71]. NCD is essential where exhaustive annotation is not feasible: urban scenes change over time, new structures appear, and manual relabeling at city scale is costly [45, 9]. By discovering and learning novel classes from unlabeled observations, NCD improves adaptability and long-horizon generalization with limited human

supervision [106].

The limitations of fixed taxonomies are particularly acute in rapidly transforming urban settings. Emerging infrastructure [2, 47, 24], innovative architectural elements [80, 40], adaptive building surfaces [38], and other novel urban features [28, 108, 22] continuously reshape the built environment while affecting local climate dynamics [97, 93, 60], hydrological patterns [84, 23], and ecosystem services. Within environmental remote sensing and land cover mapping, recent deep learning advances have improved urban density mapping [6], built-up area delineation [89], urban functional zone extraction [21], multi-modal fusion across hyperspectral, multispectral, and SAR data [37, 88], weakly supervised refinement of coarse products [8], multi-temporal change analysis [26, 58], and coastal wetland mapping [67]. Despite these advances, most pipelines still operate within fixed label spaces [86, 34, 85, 44, 55], producing inherently static maps that cannot accommodate emerging elements such as renewable energy installations, adaptive green infrastructure, smart-city fixtures, temporary urban forms, and novel façades [33, 78, 82, 48, 65]. Such features alter surface thermal properties and imperviousness, with downstream implications for heat, runoff, and ecosystem functions, underscoring the need for discovery-capable mapping.

Table 1.1 contrasts NCD with related learning paradigms. Unlike supervised learning, which requires exhaustive annotation of all categories, or open-set recognition, which rejects novel instances as a single "unknown" class, NCD actively discovers and clusters novel categories into semantically distinct groups.

Table 1.1: Comparison of learning paradigms for handling novel categories.

Aspect	**Supervised**	**Open-set**	**NCD**
Novel classes at training	None	None	Unlabeled
Novel classes at inference	Misclassified	Rejected as "unknown"	Assigned distinct labels

Progress on NCD has been substantial in 2D vision, including class-incremental methods [73], forgetting prevention [46], segmentation-oriented extensions [105], compositional experts [96], soft contrastive learning [98], class-relation modeling [29], and distribution-agnostic formulations [95]. Open-vocabulary approaches such as CLIP and ViLD further leverage image–text alignment [70, 30]. However, these advances rely on semantic anchors from large-scale image–text corpora that are not natively available in 3D point clouds, which are geometric measurements without paired textual descriptions. Extending NCD to 3D therefore introduces distinct challenges: unstructured data organization, pronounced non-uniform density and scale variation, and the integration of fine-grained geometry with broader semantic context under realistic computational constraints.

NOPS pioneered 3D NCD by combining online clustering, prototype updates, and queuing strategies, thus accommodating multiple novel classes per scene without foreground–background assumptions [71]. While establishing a strong baseline, it remains challenging to preserve fine geometry under voxel discretization, to define neighborhoods stably in irregular or overlapping regions, and to fuse local geometric cues with global context in a manner that remains robust near class boundaries. These limitations motivate the framework developed in this dissertation.

1.2 Challenges and Research Questions

We consider three challenges that constrain 3D NCD performance and scalability and hinder operational deployment:

(C1) Spatial heterogeneity and geometric organization. City-scale point clouds exhibit non-uniform density and multi-scale structure. Fixed schemes for sampling, grouping, and feature aggregation either lose fine geometry or incur prohibitive cost, leading to reduced boundary fidelity and unstable discovery.

(C2) Adaptive integration of complementary representations. Reliable discovery requires both local geometric precision and global contextual awareness. Predominantly static fusion strategies underuse complementary encoders and fail to adapt weighting to scene characteristics, sensing conditions, and viewpoints, limiting transfer from known to novel categories.

(C3) Operational deployment for evolving land cover. Practical systems must bridge the semantic gap between known and novel classes using unlabeled data, remain robust under severe class imbalance, and scale computationally to large areas with continual updates. In remote-sensing practice, fully supervised methods are confined to fixed label spaces [12, 100]; weakly and semi-supervised pipelines reduce annotation cost but still cannot capture truly novel structures [102, 13]; and unsupervised clustering lacks semantic guidance, often yielding groupings misaligned with meaningful land cover categories [57, 83].

These challenges give rise to the following research questions:

- **RQ1:** How can spatial–geometric organization effectively preserve fine-grained detail and boundary fidelity to support stable novel class discovery in large-scale point clouds?
- **RQ2:** How can complementary geometric and semantic representations be fused adaptively, with principled knowledge transfer from known to novel categories?
- **RQ3:** How can supervised segmentation for established classes be coupled with discovery mechanisms for emerging classes to support continuously updated land cover mapping?

1.3 Research Objectives and Approach

The overarching objective is to enable dense 3D semantic segmentation that discovers previously unseen categories while remaining tractable at city scale, generalizing across sensors and acquisition geometries, and delivering mapping products that keep pace with urban evolution. The approach develops along three tightly connected directions.

First, the dissertation establishes geometric foundations for efficient spatial organization (realized in *CHNCD* [18]). Representations and neighborhood construction are designed to maintain explicit links between voxel-level context and original spatial coordinates, to emphasize semantically representative regions via cluster-driven selection in feature space, and to stabilize local structures through efficient neighbor construction. Attention-based refinement mitigates artifacts from non-uniform density and overlapping regions. This geometry-aware organization preserves fine detail and improves boundary fidelity while remaining computationally practical.

Second, the dissertation deepens representation and transfer by integrating complementary feature families and making fusion context-aware (realized in *AGDNet* [20]). Density-aware geometric sequence modeling captures multi-scale spatial relations; dynamic probabilistic embeddings provide continuous fields that accommodate varying sparsity patterns; and gated multi-source fusion adapts the weighting of complementary encoders to scene statistics and sensing conditions. Category-, instance-, and distribution-level objectives support principled knowledge transfer from well-represented known classes to underrepresented novel ones.

Third, the dissertation bridges research frameworks with operational deployment for land cover discovery mapping (realized in *3DLCDM* [19]). Hybrid supervision couples full supervision for established land cover types with prototype-driven discovery for emerging

classes, stabilizes pseudo-labels through temporal normalization, and mitigates imbalance via frequency-aware dynamic weighting that combines per-iteration statistics with long-term momentum. A shared sparse-convolutional backbone supports efficient large-scale processing and continuity with existing remote-sensing production pipelines.

1.4 Thesis Structure

This thesis comprises six chapters, summarized in Figure 1.1.

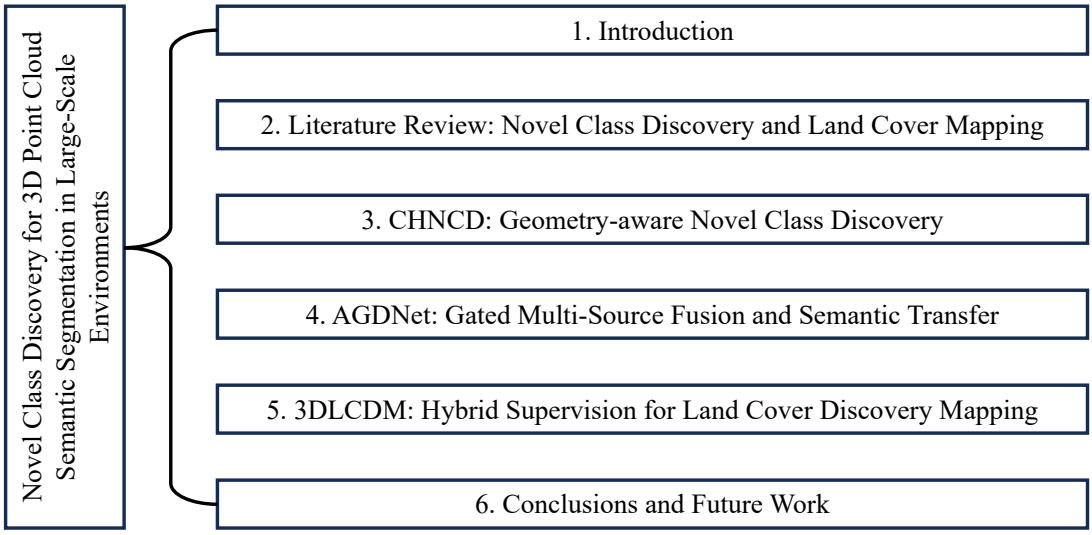

Figure 1.1: Organization of the thesis.

Chapter 1 introduces the problem setting, motivates NCD for 3D point clouds, formulates the challenges and research questions, and outlines the overall approach and contributions.

Chapter 2 reviews NCD across 2D and 3D, covering early transfer and contrastive formulations, extensions to dense prediction, and point cloud pipelines, with an application focus on land cover mapping through recent 2D segmentation and selected 3D LiDAR literature.

Chapter 3 presents *CHNCD*, a geometry-aware discovery framework that organizes spatial information and stabilizes neighborhood construction for large-scale point clouds.

Chapter 4 introduces *AGDNet*, which advances adaptive geometric modeling, probabilistic embeddings, context-aware fusion, and multi-level semantic transfer.

Chapter 5 proposes *3DLCDM*, a hybrid-supervision formulation for land cover discovery mapping and demonstrates operational relevance on high-resolution LiDAR datasets.

Chapter 6 concludes the dissertation and outlines directions for future work.

Chapter 2

Novel Class Discovery and Land Cover Mapping

Urban infrastructure evolves continuously, yet traditional 3D mapping relies on exhaustive manual annotation that cannot keep pace with construction cycles. This chapter establishes the theoretical foundations for discovering novel infrastructure categories in multi-sensor point clouds. Section 2.1 surveys novel class discovery methods, identifying limitations in geometry-aware processing and spatial coherence. Section 2.2 reviews land cover mapping applications, exposing a gap: most methods assume fixed taxonomies and cannot accommodate emerging categories. Section 2.3 synthesizes three research needs and previews how our studies address them through cluster-based sampling, geometric sequence modeling, and hybrid-supervision protocols.

2.1 Novel Class Discovery

2.1.1 Early Foundations of Novel Class Discovery

NCD has become increasingly important for urban scene understanding and infrastructure monitoring applications. Hsu et al. [41] established the theoretical framework by introducing transferable pairwise similarity information, demonstrating that categorical knowledge could be effectively transferred across domains through learnable clustering objectives. Their approach showed remarkable robustness through dense pairwise constraints within mini-batches, though computational overhead increased quadratically with batch size and

performance remained dependent on similarity prediction quality. Building upon this foundation, Han et al. [32] introduced Deep Transfer Clustering (DTC), extending Deep Embedded Clustering with transfer learning capabilities through representational bottlenecks and temporal ensembling. DTC demonstrated solid theoretical foundations and achieved substantial improvements over existing methods, though the multi-stage training pipeline increased computational requirements. Han et al. [31] advanced the field by proposing ranking statistics for novel category discovery, utilizing rank-based pseudo-labeling to handle noisy descriptors in high-dimensional spaces. While their three-stage pipeline combining self-supervised pretraining, supervised fine-tuning, and joint optimization showed significant improvements, the approach involved computational overhead from rank-based comparisons and operated under assumptions about visual similarity between known and novel classes.

These foundational approaches established solid theoretical frameworks for novel class discovery and demonstrated broad applicability across domains. However, they commonly faced challenges including computational scalability issues, dependence on similarity prediction quality, and the need for multi-stage training processes, highlighting the requirement for more efficient clustering mechanisms.

2.1.2 Architectures and Learning Strategies for NCD

The development of multi-branch architectures was explored by Zhao et al. [103], who introduced dual ranking statistics with mutual knowledge distillation. Their approach incorporated local part-level information through dynamically maintained object dictionaries, generating more robust pseudo-labels by combining local and global features. While demonstrating superior performance on fine-grained recognition tasks, the method necessitated increased computational resources from dual-branch training. Jia et al. [45] presented a unified framework eliminating complex multi-stage pipelines through end-to-end optimization, combining category discrimination with cross-modal contrastive objectives and Winner-Take-All hashing. Despite eliminating external clustering stages and achieving strong results, the method required substantial computational resources and empirically tuned hyperparameters. The integration of supervised knowledge during clustering was addressed by Zhong et al. [107] through OpenMix, which introduced tailored data augmentation mechanisms inspired by MixUp. Their approach effectively reduced overfitting to noisy pseudo-labels through joint distributions, though it involved multiple augmentation parameters requiring careful tuning.

Advanced learning strategies in this category demonstrated progress toward more sophisticated feature integration and end-to-end optimization. These methods achieved im-

proved performance through multi-branch architectures and unified frameworks, yet commonly required substantial computational resources and extensive hyperparameter optimization, indicating the need for more parameter-efficient approaches.

2.1.3 Contrastive Objectives and Unified Label-Space Formulations

Zhong et al. [106] introduced Neighborhood Contrastive Learning (NCL), exploiting locality properties in embedding spaces by treating k-nearest neighbors as pseudo-positives. Their Hard Negative Generation synthesized informative negatives by mixing labeled and unlabeled features, capitalizing on the disjoint nature of class sets. Despite solid mathematical foundations, NCL required considerable computational resources due to large memory banks and showed sensitivity to initial feature quality. A unified objective approach was proposed by Fini et al. [25], treating pseudo-labels homogeneously with ground truth labels to enable single cross-entropy loss operation on both labeled and unlabeled data. This elegant solution eliminated self-supervised pretraining and reduced computational complexity while achieving substantial improvements, though it remained dependent on related labeled data availability and required careful hyperparameter tuning.

Contrastive learning approaches offered strong mathematical foundations and demonstrated impressive empirical results. However, they typically required substantial computational resources, showed sensitivity to initialization quality, and remained dependent on memory-intensive mechanisms, motivating the development of more resource-efficient alternatives with better adaptation capabilities.

2.1.4 Meta-learning and Incremental Discovery

Chi et al. [9] established the connection between novel class discovery and meta-learning, formulating NCD as a sampling problem and proving necessary conditions for solvability when known and novel classes share high-level semantic features. Their approach adapted MAML and ProtoNet through clustering-rule-aware task sampling, demonstrating effectiveness with limited data but involving considerable computational complexity from meta-learning procedures. Class-incremental learning was addressed by Roy et al. [73] through FRoST, preventing catastrophic forgetting via dual classifiers with feature replay and knowledge distillation. The method demonstrated superior balance between old and new class accuracy, though it involved increased computational considerations from

prototype storage. Joseph et al. [46] extended NCD to practical scenarios through pseudo-latent generation and mutual information regularization, enabling deployment without labeled data storage but involving computational complexity and generalization challenges.

Meta-learning and incremental approaches provided theoretical insights and addressed practical deployment scenarios. These methods established important connections between novel class discovery and meta-learning principles while tackling catastrophic forgetting. Nevertheless, they generally involved considerable computational complexity and faced challenges in generalization across domains with significant semantic gaps.

2.1.5 Extension to Complex Domains

The extension to semantic segmentation was pioneered by Zhao et al. [105], formalizing NCDSS with a two-stage framework combining basic clustering with entropy-based uncertainty modeling. While demonstrating solid theoretical foundations, the approach relied on accurate saliency maps and involved multi-stage training overhead. Recent advances include distribution-agnostic approaches by Yang et al. [95], who relaxed uniform distribution assumptions through iterative prior estimation and dynamic temperature adjustment. Their method, BYOP, demonstrated robust performance across various scenarios though encountered challenges with minority classes in complex datasets. Gu et al. [29] leveraged class-relation knowledge distillation to preserve semantic relationships, demonstrating natural capture of inter-class relationships though relying on fixed supervised models. Li et al. [54] modeled disjoint characteristics through symmetric KLD constraints, offering theoretical foundations based on information theory but involving computational overhead from pairwise computations.

Extensions to complex domains advanced novel class discovery beyond image classification to more challenging scenarios. These approaches demonstrated progress in handling distribution imbalances and preserving semantic relationships, yet commonly relied on auxiliary components and incurred computational overhead from specialized constraint mechanisms, suggesting the need for more integrated solutions.

2.1.6 Novel Class Discovery for 3D Point Clouds

The transition to 3D point cloud processing was initiated by Riz et al. [71], addressing fundamental differences from 2D domains where background-foreground assumptions do not hold. Their approach employed online clustering with uncertainty quantification and optimal transport formulation, achieving computational efficiency and handling multiple

novel classes simultaneously. The extension to 3D point clouds represents a significant advancement in novel class discovery, addressing unique challenges of geometric data processing. While these methods demonstrated computational efficiency and the ability to handle multiple novel classes, they face several limitations in complex urban environments.

2.2 Deep Learning for Land Cover Mapping

2.2.1 2D Semantic Segmentation for Land Cover Mapping

In recent years, the application of deep learning in environmental remote sensing has achieved remarkable results. Numerous studies have leveraged semantic segmentation techniques to perform detailed and diverse environmental analyses in tasks such as urban density mapping [6], urban built-up area mapping [89], multi-scale mapping of urban functional zones [21], and multi-modal urban land cover mapping [37]. In addition, some works have refined low-resolution land cover products through weakly supervised approaches [8]. These methods have demonstrated outstanding performance in enhancing segmentation accuracy and spatial detail representation, thereby providing a solid foundation for land cover mapping.

Several studies have focused on leveraging convolutional neural network (CNN)-based segmentation for diverse environmental tasks. Chai et al. [5] employed a multi-layer CNN-based semantic segmentation approach to detect clouds and shadows in Landsat imagery for environmental analyses, effectively leveraging multi-level spectral–spatial features and surpassing threshold-based methods, especially under thin-cloud and complex-surface conditions, thus enhancing large-scale land monitoring. Further extending the scope to urbanization impacts, Wu et al. [89] proposed a deep-learning-based semantic segmentation method to map built-up areas from GF-3 (Gaofen-3) Synthetic Aperture Radar images across China, aiming to evaluate environmental influences of urbanization. By introducing residual blocks for robust training, it reduces misclassification in complex terrains and delivers a nationwide high-resolution dataset for analyzing urban expansion's ecological impacts. Moreover, multi-scale approaches have been developed to enhance spatial precision in urban analyses. Du et al. [21] integrated multi-scale semantic segmentation with an object-based Conditional Random Field to map Urban Functional Zones (UFZs) in very-high-resolution images. This approach provides fine-grained, large-scale identification of built-up areas, farmland, and green spaces, aiding environment-related analyses, such as monitoring urbanization and planning sustainable landscapes.

Multi-modal approaches have further enhanced land cover mapping robustness. Hong

et al. [37] introduced a multimodal dataset, Cross-City Semantic Segmentation (C2Seg), and a High-Resolution Domain Adaptation Network (HighDAN) to enhance semantic segmentation in urban land cover mapping. By integrating hyperspectral, multispectral, and Synthetic Aperture Radar (SAR) data, it handles domain shifts across cities, enabling robust environmental applications such as large-scale monitoring of diverse urban landscapes. To overcome climate-induced domain gaps, Liu et al. [58] proposed a novel Climate-Aware Satellite Image Dataset (CASID) for land cover semantic segmentation. They enable robust cross-zone building, forest, road, and water mapping, advancing generalization and improving climate-sensitive environmental monitoring via domain adaptation techniques that accommodate diverse climatic conditions, fostering reliable vegetation variation analysis for broader large-scale operational applications. Additional efforts have focused on refining coarse land cover products with minimal labeling. Chen et al. [8] introduced a two-stage weakly supervised semantic segmentation framework (WESUP-LCP) to refine low-resolution Land Cover Products (LCPs). By transferring reliable samples and leveraging dynamic label propagation, it achieves high-resolution, area-specific maps, aiding environmental tasks like urban expansion monitoring, forest detection, and agricultural management without additional labeling costs.

For dynamic urban environments, temporal segmentation methods have emerged to capture land cover transitions. He et al. [35] integrated temporal semantic segmentation with monthly Sentinel-2 data, enabling precise segmentation of land cover transitions (e.g., forest to impervious surface) for sustainable urban development. By jointly modeling "where, when, and what" changes occur, it addresses critical environmental monitoring needs, demonstrating improved accuracy over traditional methods through direct classification of spectral sequences into semantic labels and capturing multi-type changes. In a broader environmental context, Saputra et al. [74] combined deep learning-based semantic segmentation with multispectral satellite imagery to map diverse mining footprints globally, thereby enabling comprehensive environmental assessments. Their approach classifies both mining and non-mining land covers, providing automated, scalable, and accurate methods for monitoring ecological and social impacts across varied climate zones and commodities.

Despite these advances, 2D methods face two fundamental limitations. First, they rely on fixed semantic taxonomies defined during training. Introducing new land cover types such as green infrastructure, modular construction, or adaptive facades requires complete dataset re-annotation and model retraining. Second, 2D imagery lacks vertical geometric information essential for distinguishing structures with similar spectral signatures but different heights and configurations. These limitations constrain the ability of 2D methods to characterize complex urban morphologies and adapt to evolving land cover patterns.

2.2.2 3D Point Cloud Segmentation for Land Cover Mapping

The transition from 2D satellite imagery to 3D point clouds addresses the geometric information gap. Airborne and terrestrial LiDAR systems enable detailed characterization of urban vertical structures and fine-grained surface geometries. Recent advances have demonstrated potential for automated land cover extraction from 3D data. Xiang et al. [92] integrated a 3D deep-learning pipeline for high-density airborne LiDAR forest scans, combining semantic segmentation and instance segmentation. Their method automatically extracts tree components and structural attributes, enabling precise environmental assessments (e.g., canopy characterization, biomass estimation), and thus advances forest inventory and management for sustainable resource monitoring. Further advancing 3D LiDAR applications in environmental monitoring, Lu et al. [63] introduced Sen-net, an end-to-end method for semantic segmentation in forest data. By integrating specialized modules for distinct vegetation components, it robustly extracts key ecological structures (e.g., foliage, trunks) across diverse forest conditions. The results demonstrate improved forest ecosystem interpretation, aiding further environmental monitoring and long-term sustainable management.

In urban contexts, 3D segmentation enables precise delineation of categories ambiguous in 2D imagery. Buildings can be distinguished by height and architectural complexity; roads and parking lots separated by elevation differences; vegetation classified by canopy structure rather than spectral response alone. Multi-temporal acquisitions further support change detection for urban growth monitoring, infrastructure development tracking, and green space evolution analysis. However, the computational cost of processing large-scale urban point clouds and the scarcity of annotated training data remain significant barriers to operational deployment.

Despite these advances, 3D methods inherit the fixed-taxonomy constraint from 2D approaches while introducing additional challenges specific to point cloud geometry. Semantic categories must be predefined and exhaustively annotated before training. Models cannot autonomously discover emerging urban features such as rooftop installations, solar panel arrays, or temporary construction elements without new manual annotations, limiting their applicability for continually evolving urban environments.

2.3 Chapter Summary

This chapter reviews NCD from early theory to dense prediction and 3D point cloud segmentation, with an application context on land cover mapping. Foundational studies show

that transferable similarity learning, deep transfer clustering, and ranking-based pseudo-labeling can organize unlabeled samples into coherent categories while leveraging limited supervision, yet they introduce computational–statistical trade-offs through dense pairwise constraints and multi-stage training and are sensitive to pseudo-label quality. Subsequent work expands the design space via multi-branch distillation, end-to-end objectives that unify labeled and unlabeled signals, and contrastive formulations that reduce reliance on external clustering; these gains often depend on memory banks and delicate temperature or margin schedules, and generalization weakens when novel categories diverge semantically from the labeled set. Extensions from recognition to segmentation couple clustering with uncertainty modeling and distribution-aware objectives, improving applicability to structured outputs but frequently relying on auxiliary components and incurring added cost. The transition to 3D point clouds exposes a representation gap: methods conceived for images rarely account for sparsity, non-uniform sampling, and topology; pipelines may rely on fixed backbones, heuristic imbalance handling, or static fusion/attention and thus adapt poorly when viewpoint, density, or sensing conditions shift. In environmental remote sensing, 2D CNN and multimodal approaches, as well as 3D LiDAR pipelines, have improved semantic interpretation across cities, climates, and sensors, yet most frameworks assume a fixed label space and periodic reannotation, yielding static maps that struggle to reflect emerging urban structures in a timely manner.

Three needs follow from this synthesis. First, efficient neighborhood construction and geometry-aware feature organization are required to handle high-dimensional point clouds at scale while preserving fine-grained spatial details crucial for novel class detection. Second, multi-source fusion should be integrated with scene-dependent gating rather than treated as a static concatenation, so that complementary sensing modalities and geometric sequences contribute according to context. Third, discovery must be coupled to practical mapping protocols to enable continually updated land cover products without exhaustive relabeling, balancing supervised learning for established classes with principled mechanisms that introduce new classes as they appear.

The thesis addresses these needs through three studies: cluster-based sampling with proximity hashing for geometry-aware grouping, adaptive multi-source fusion through context-aware gating, and hybrid supervision coupling discovery with land cover mapping. The following three chapters detail each study, presenting the methodology, comprehensive evaluations, and ablation studies that quantify the contributions of key components.

Chapter 3

CHNCD: Geometry-aware Novel Class Discovery[1]

3.1 Introduction

NCD in 3D semantic segmentation presents unique challenges due to the high dimensionality, unstructured nature, and large volume of point cloud data. The complexity of high-dimensional data is a major hurdle, as 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. Furthermore, efficient determination of local neighborhoods and effective feature representation in high-dimensional spaces are problematic, with standard nearest-neighbor searches being computationally costly. Processing all points in large point clouds is computationally prohibitive, and indiscriminate downsampling risks losing critical information. Moreover, real-world 3D environments are dynamic, with constantly evolving objects and conditions, necessitating methods capable of generalizing to unseen classes and adapting to changing environments.

While NOPS [71] pioneered 3D NCD by extending the framework from 2D domains and introducing online clustering with continuous prototype updates and queuing strategies to handle multiple novel classes per scene, several key limitations constrain its ability to preserve fine-grained geometry and organize neighborhoods effectively in complex ur-

[1]The content of the chapter has been published in ISPRS Journal of Photogrammetry and Remote Sensing with the paper entitled "3D semantic segmentation: cluster-based sampling and proximity hashing for novel class discovery".

ban scenes. 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.

To address these challenges, 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, termed CHNCD, is proposed to address these questions, as shown in Figure 3.1. The framework integrates voxel-based representations with original spatial coordinates using an index mapping mechanism. Although existing methods such as PV-RCNN [75] and PVCNN [59] have effectively combined voxel- and point-based feature extraction for 3D data analysis, 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 PVCNN, 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.

Additionally, the proposed method employs a Cluster-based Representative Sampling module (CRSM) to reduce computational complexity by identifying key representative

points within clusters in the high-dimensional feature space, a Neighborhood Spatial Partitioning module (NSPM) utilizing Proximity Hash Mapping (PHM) to map dense feature vectors to a lower-dimensional hash space for improved neighborhood searches, and a Spatial Feature Attention Mechanism (SFAM) to integrate localized feature vectors from neighboring points with global feature context. 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-dimensional 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 while preserving essential spatial and semantic features, thereby 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 Method

3.2.1 Voxel-geometry Data Integration

In the context of 3D point cloud semantic segmentation, we present a novel methodology, the 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 when discovering 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

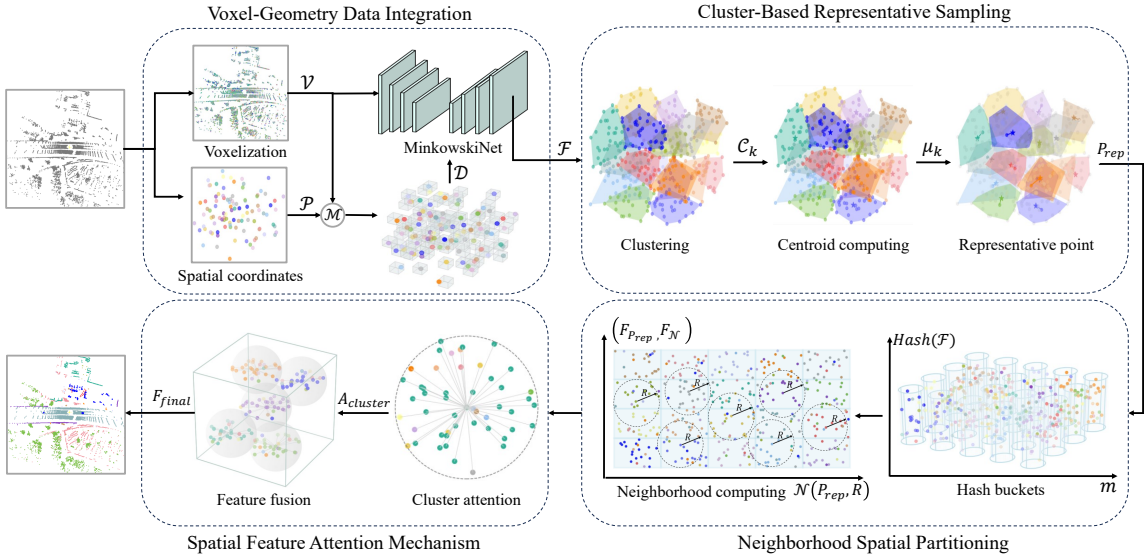

Figure 3.1: Framework of the proposed network. The proposed network consists of four functional modules: Voxel-Geometry Data Integration module, Cluster-based Representative Sampling module, Neighborhood Spatial Partitioning module, and Spatial Feature Attention Mechanism. $\mathcal{V}$: Voxelized representation 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. $\left(F_{P_{\text{rep}}}, F_{\mathcal{N}}\right)$: Localized feature vectors of representative points and their neighbors. $\mathcal{N}\left(P_{\text{rep}}, R\right)$: 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.

computational models. To this end, we define a dual-mapping mechanism that transforms $\mathcal{P}$ into a voxelized representation $\mathcal{V}$ and preserves the original point coordinates via 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:

$$\mathcal{V} = f_{\text{vox}}\ (\mathcal{P})\,. \tag{3.1}$$

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:

$$\mathcal{M} : \mathcal{V} \leftrightarrow \{\mathcal{P}_j\}_{j=1}^{M}\,. \tag{3.2}$$

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:

$$f_{\text{int}}\ (\mathcal{V}, \mathcal{M}) \rightarrow \mathcal{D}, \tag{3.3}$$

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 [10], a widely recognized and well-established method for voxel-based feature extraction [71]. 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:

$$\mathcal{F} = \text{MinkowskiNet}(\mathcal{D}) = \{f_i\}_{i=1}^{N}, \tag{3.4}$$

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

3.2.2 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:

$$C_1, C_2, \ldots, C_k = \text{MiniBatchKMeans}(\mathcal{F}, k), \tag{3.5}$$

where C_k denotes the set of indices of points belonging to the k-th 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:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} f_i, \tag{3.6}$$

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:

$$P_{\text{rep}} = \underset{i \in C_k}{\text{argmin}} \, \| f_i - \mu_k \|_2 . \tag{3.7}$$

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

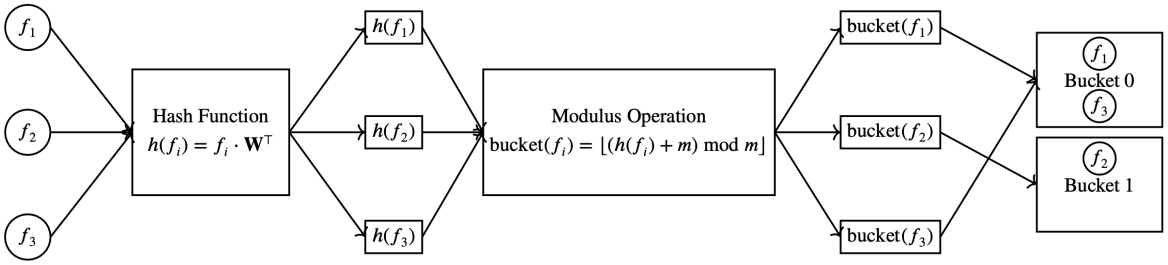

Figure 3.2: Illustration of the Proximity Hash Mapping (PHM) process. High-dimensional feature vectors f_i are projected using the hash function $h(f_i) = f_i \cdot \mathbf{W}^\top$, resulting in real-valued hash outputs $h(f_i)$. The modulus operation maps these values into discrete bucket indices $\text{bucket}(f_i)$. 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.

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 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.

3.2.3 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:

$$h(f_i) = f_i \cdot \mathbf{W}^\top, \tag{3.8}$$

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:

$$\text{bucket}(f_i) = \lfloor (h(f_i) + m) \bmod m \rfloor, \tag{3.9}$$

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:

$$a \bmod m = a - m \cdot \left\lfloor \frac{a}{m} \right\rfloor, \tag{3.10}$$

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:

$$\text{Hash}(\mathcal{F}) = \bigcup_{i=1}^{N} \text{bucket}(f_i), \tag{3.11}$$

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-dim-ensional 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 Figure 3.2, 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 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:

$$R = \gamma \cdot \rho_{\text{rep}}, \tag{3.12}$$

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:

$$\mathcal{N}\left(P_{\text{rep}}, R\right) = \left\{ p_j \in \mathcal{P} \mid \left\| p_j - P_{\text{rep}} \right\|_2 \leq R \right\}, \tag{3.13}$$

where $\mathcal{N}(P_{\text{rep}}, R)$ 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 $\|p - P_{\text{rep}}\|_2 \leq R$, 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.

3.2.4 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:

$$Q = W^Q \cdot F_{P_{rep}}, K = W^K \cdot F_{\mathcal{N}}, V = W^V \cdot F_{\mathcal{N}}, \tag{3.14}$$

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:

$$\alpha = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \tag{3.15}$$

$$\mathbf{A} = \alpha V, \tag{3.16}$$

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:

$$\mathbf{A}_{\text{cluster}} = \frac{1}{d}\sum_{i=1}^{d} \mathbf{A}_i, \tag{3.17}$$

where d is the feature dimension, and $\mathbf{A}_i$ denotes the i-th 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:

$$\mathbf{F}_{\text{updated}} = \mathcal{F}_{C_k} + \mathbf{A}_{\text{cluster}} , \tag{3.18}$$

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 $\mathbf{F}_{\text{final}}$:

$$\mathbf{F}_{\text{final}} = \mathbf{F}_{\text{global}} \oplus \mathbf{F}_{\text{updated}} , \tag{3.19}$$

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.

3.3 Experiments

3.3.1 Datasets

The S3DIS dataset [1] 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 6,000 square meters. The spaces covered range in size from approximately 450 to 1,900 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 [78] 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 [94] is a large-scale, point-wise annotated dataset designed for 3D semantic segmentation under various adverse weather conditions. It leverages the STF [3] benchmark, collected in Germany, Sweden, Denmark, and Finland, which captures LiDAR scans under conditions such as snow, dense fog, light fog, and rain. The dataset includes 2,076 scans from a Velodyne HDL64 S3D LiDAR sensor, carefully selected for geographic diversity, and is split into 1,326 training scans, 250 validation scans, and 500 test scans.

The SemanticPOSS dataset [66] 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 envi-

ronments. 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, 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.

3.3.2 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

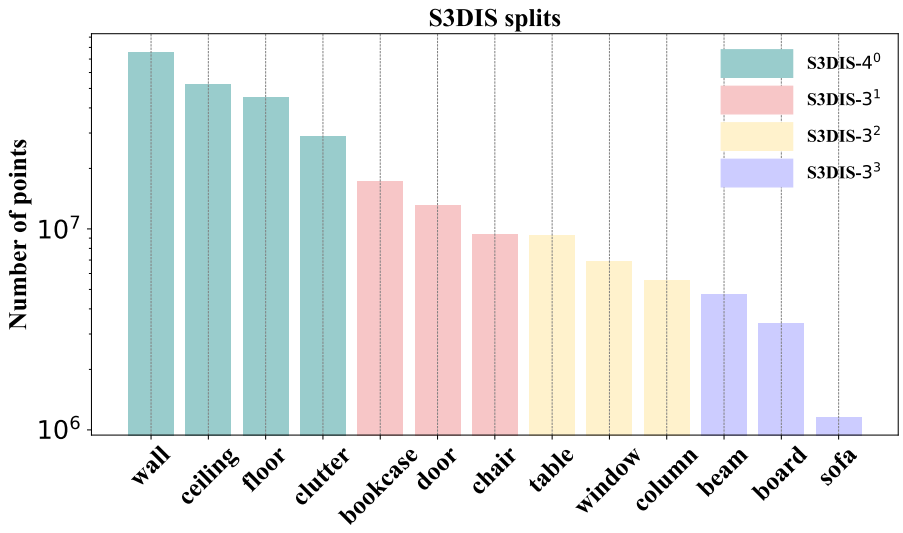

Figure 3.3: Histograms illustrating the number of points for each class on the S3DIS dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

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 [71]. 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

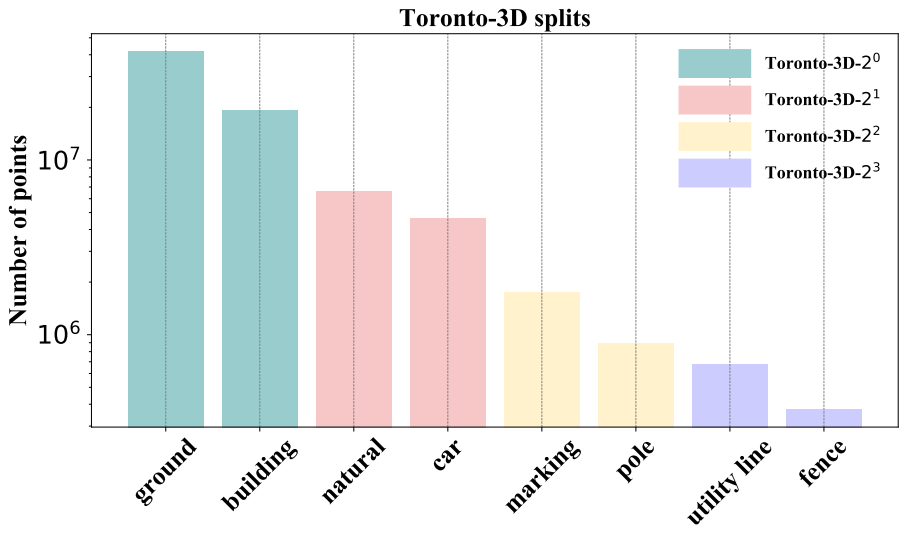

Figure 3.4: Histograms illustrating the number of points for each class on the Toronto-3D dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

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. Figures 3.3, 3.4, 3.5, and 3.6 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.

3.3.3 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.

3.3.4 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

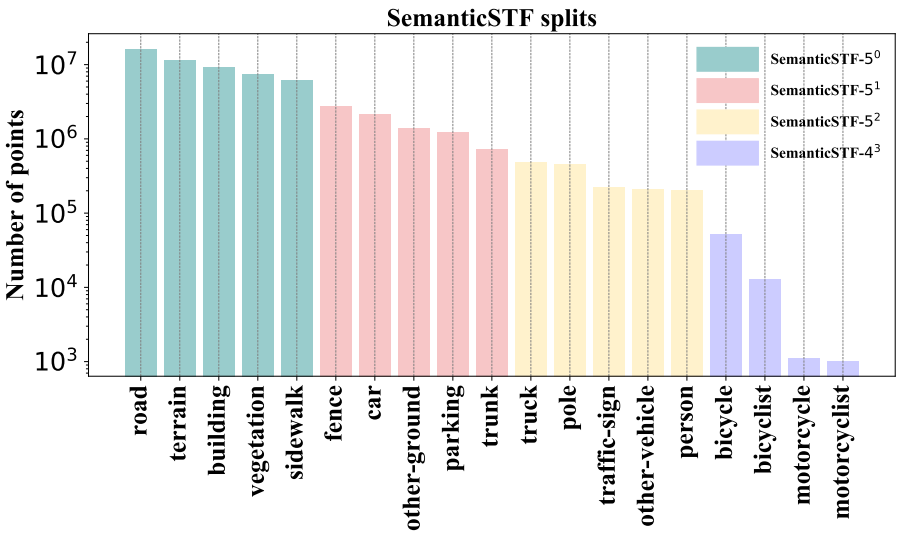

Figure 3.5: Histograms illustrating the number of points for each class on the SemanticSTF dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

Table 3.1: 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

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 3.1 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 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

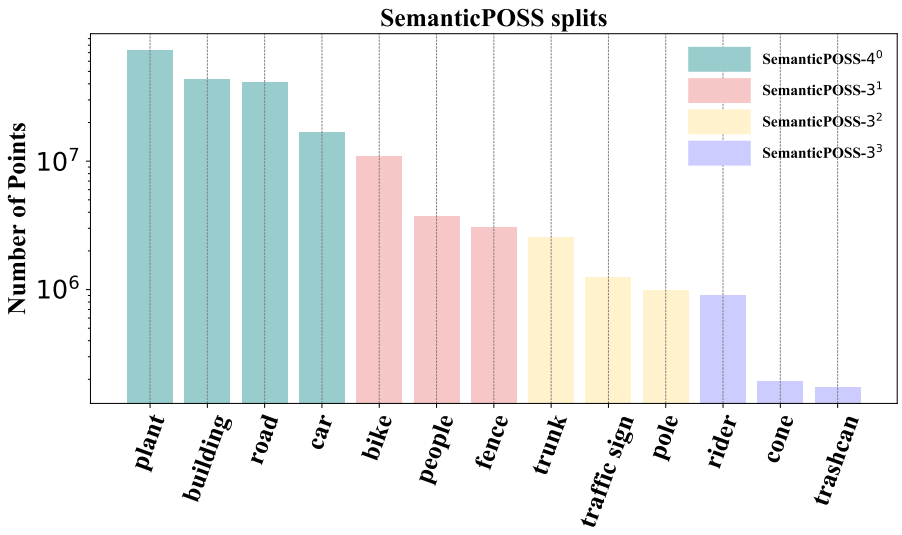

Figure 3.6: Histograms illustrating the number of points for each class on the SemanticPOSS dataset. The assigned color indicates the class is regarded as novel in the corresponding split.

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 per-

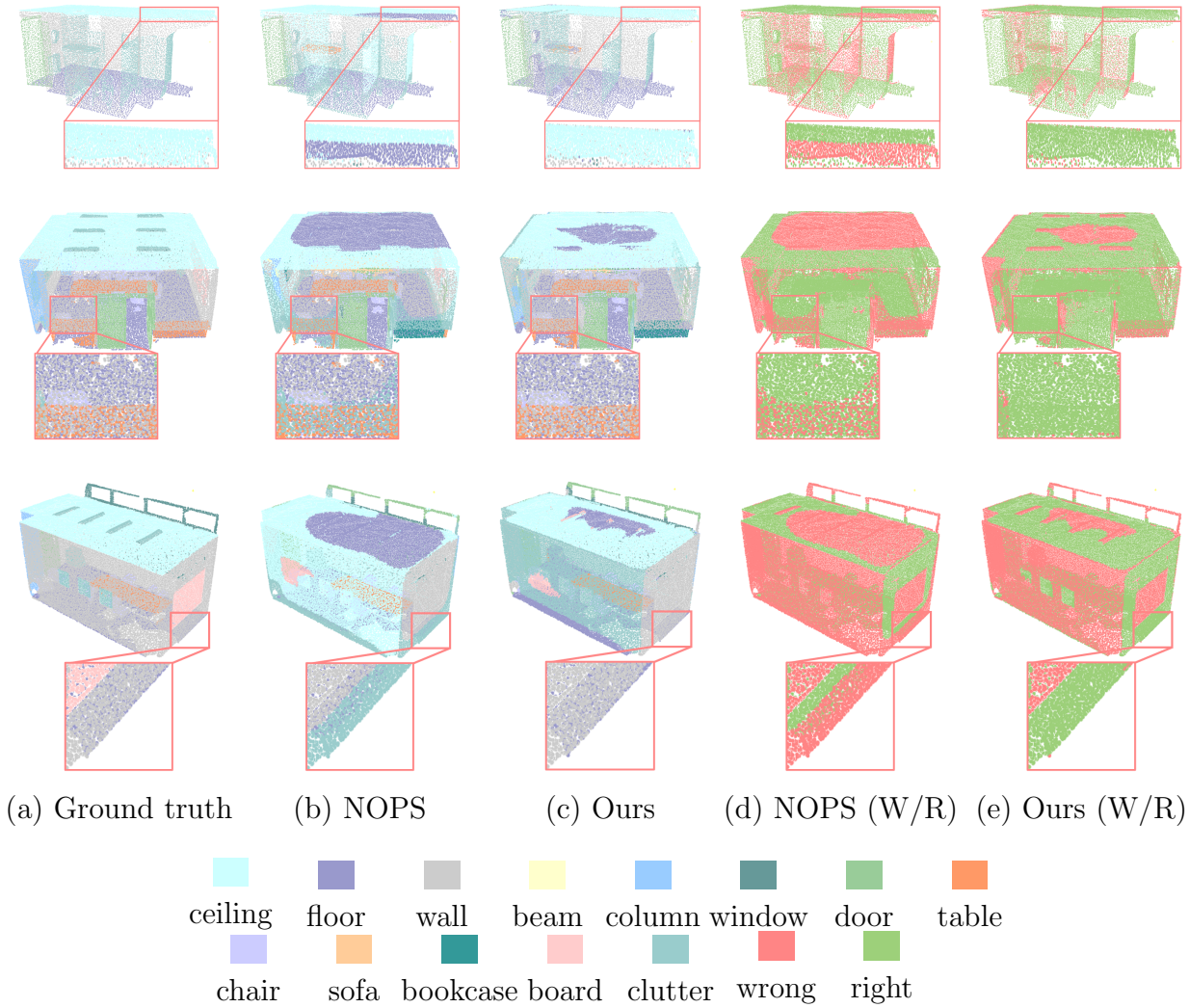

Figure 3.7: Segmentation results on the S3DIS-4^0 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes as indicated by the legend. In images (d-e), red represents segmentation errors, while green indicates 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.

formance 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 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.

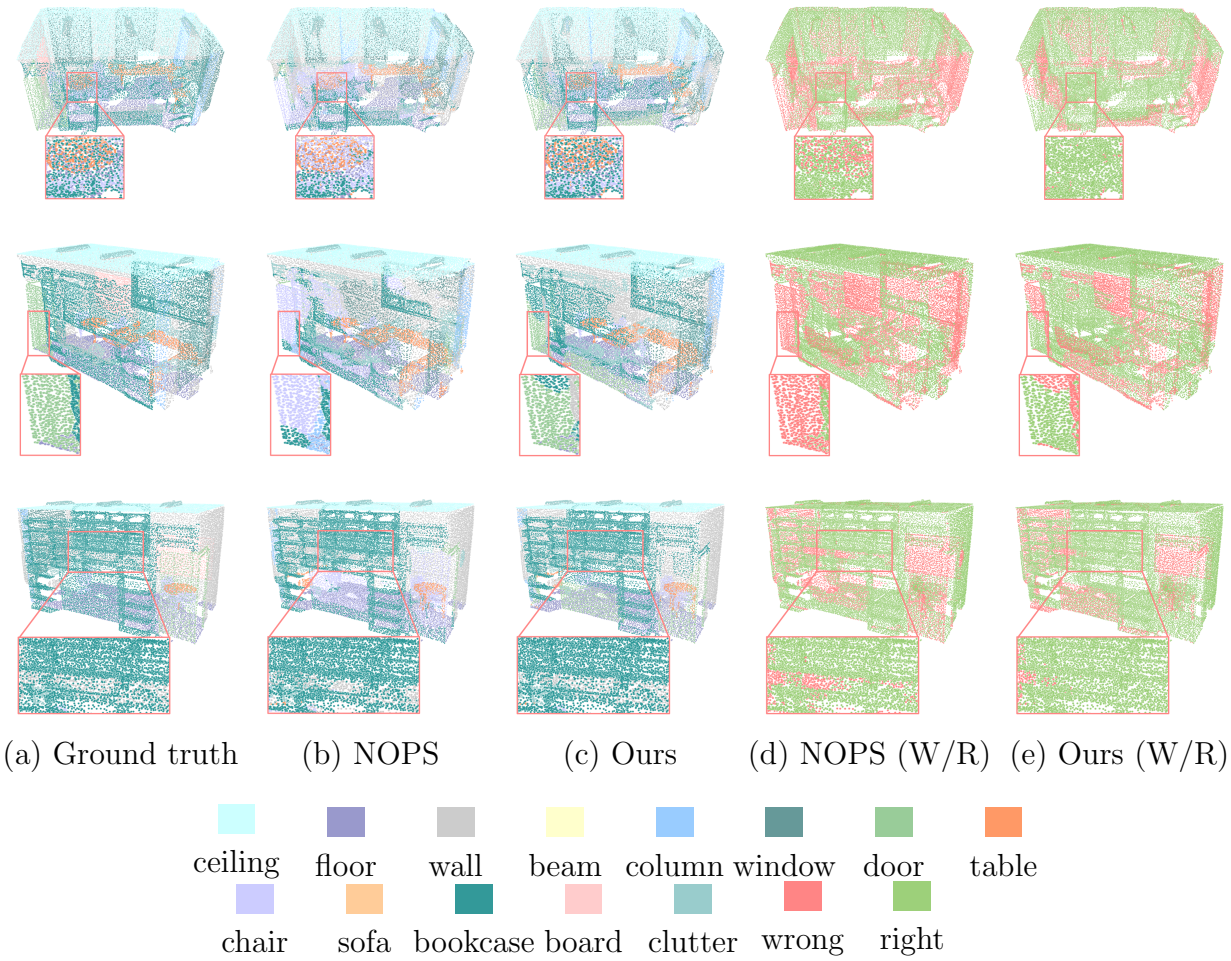

Figure 3.8: Segmentation results on the S3DIS-3[1] split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. The 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.

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.

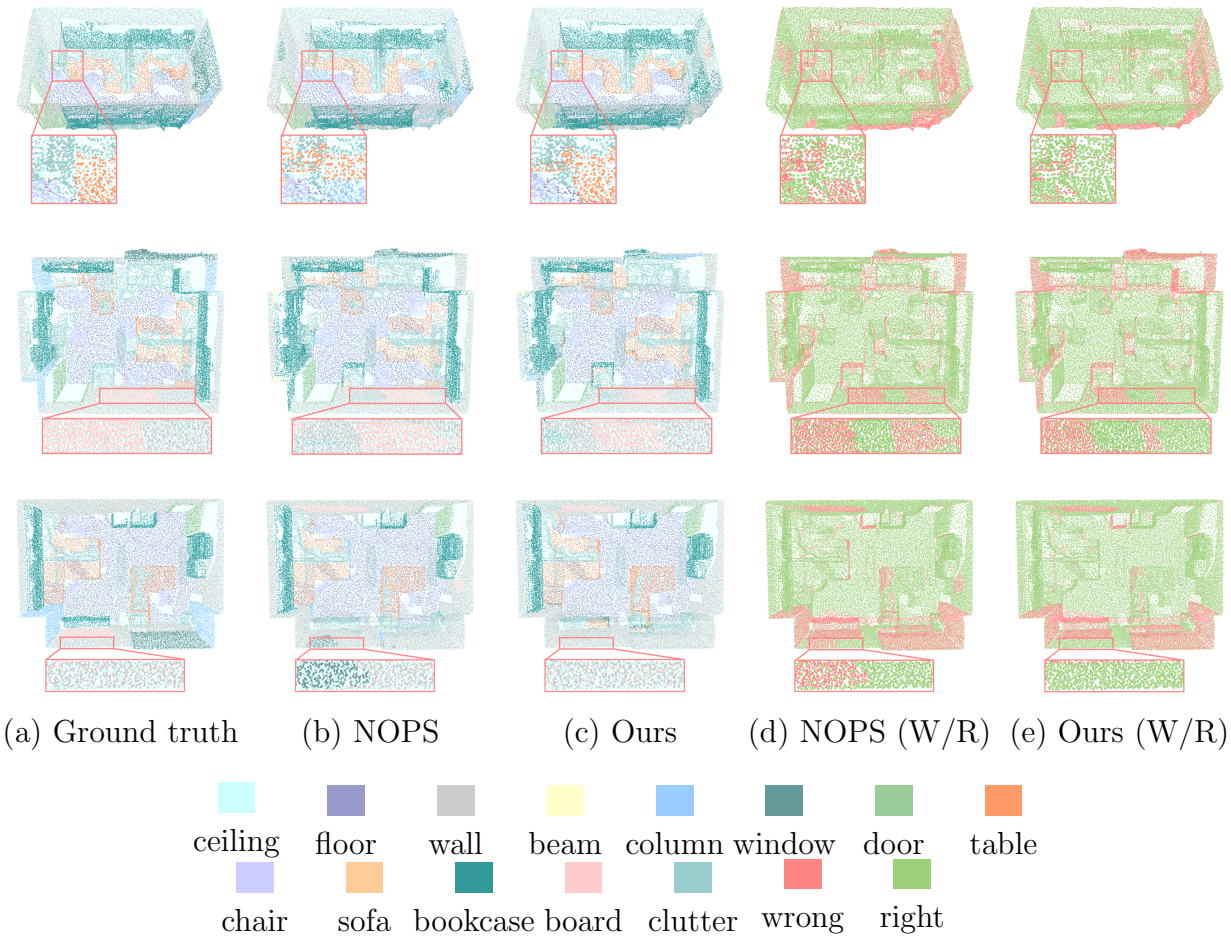

Figure 3.9: Segmentation results on the S3DIS-3^2 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. Zoom-in 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.

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 Figures 3.7, 3.8, 3.9, and 3.10 illustrate the segmentation

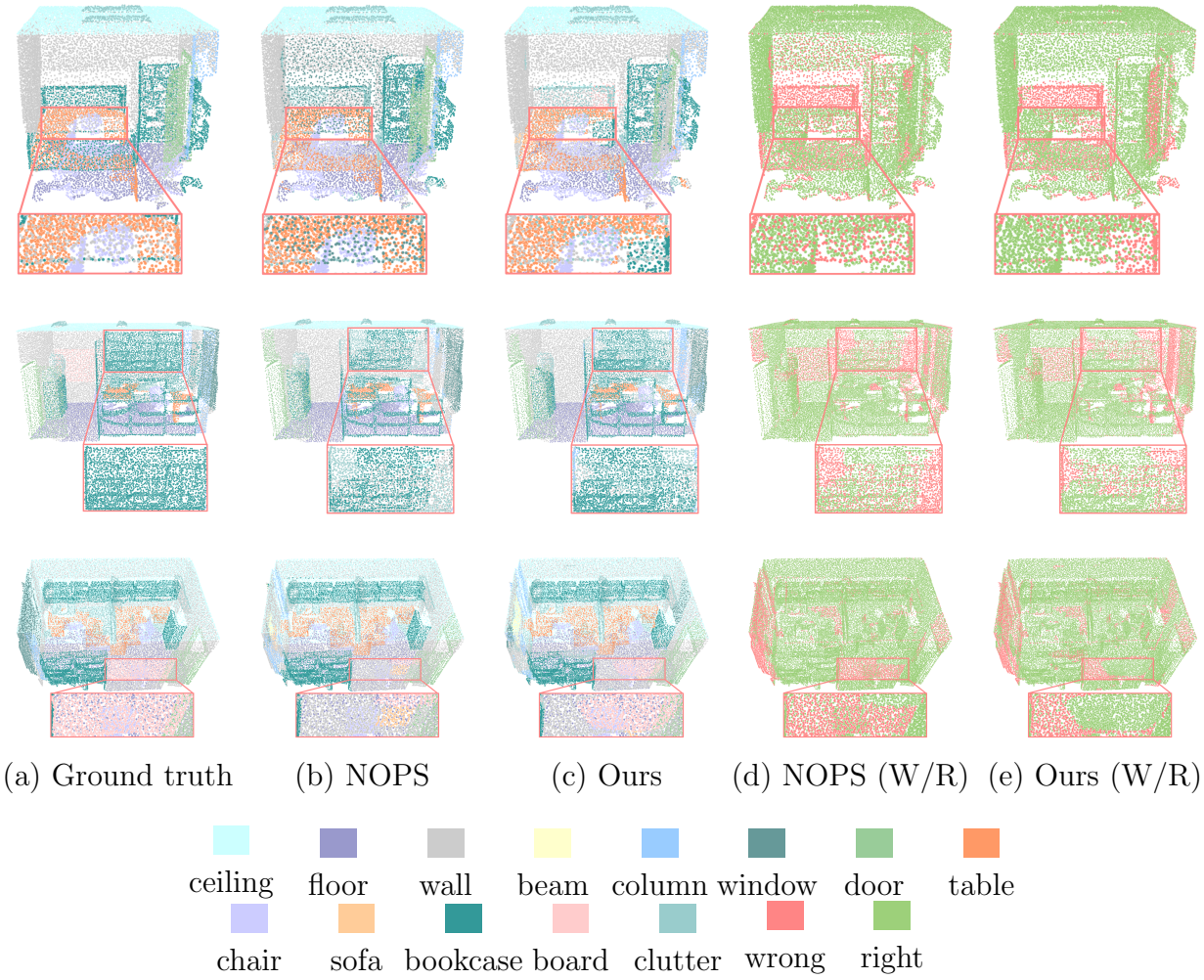

Figure 3.10: Segmentation results on the S3DIS-3^3 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. The 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.

Table 3.2: 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

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.

3.3.5 Evaluation and Visualization Results on Toronto-3D

The experimental results on the Toronto-3D dataset are detailed in Table 3.2. 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

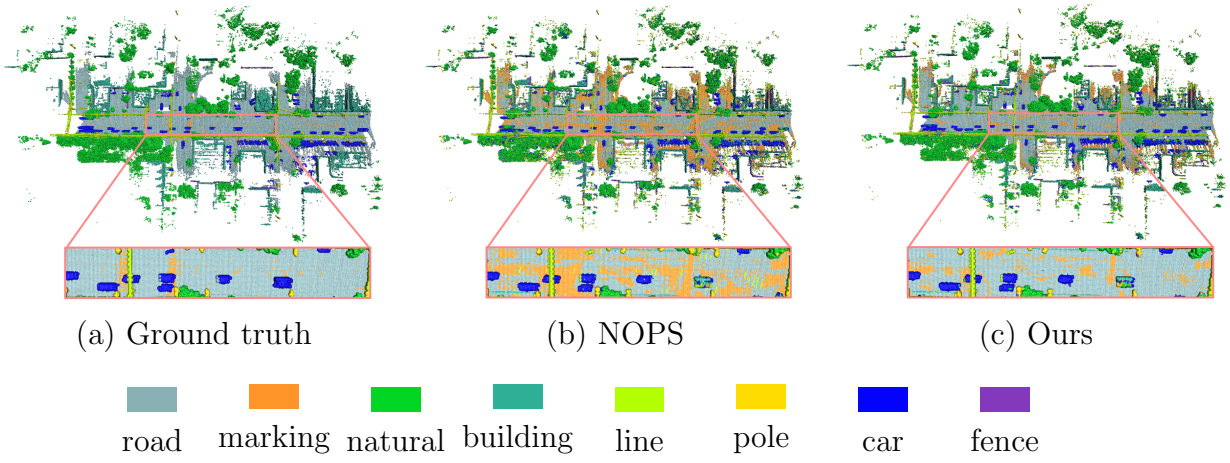

Figure 3.11: Segmentation results on the Toronto-3D-2^0 split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a-c), different colors represent different semantic classes. Zoom-in views show detailed segmentation results of the road surface. It can be observed that our method more closely matches the ground truth, while NOPS incorrectly segments a larger portion of the road surface as road markings.

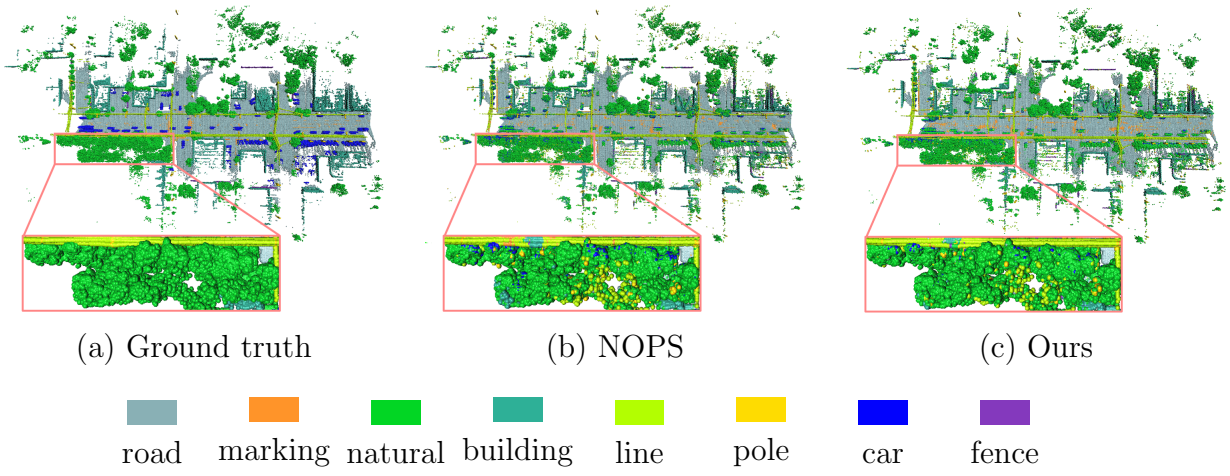

Figure 3.12: Segmentation results on the Toronto-3D-2^1 split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a-c), different colors represent different semantic classes. Zoom-in views show detailed segmentation results of the natural class. It can be observed that both our method and NOPS perform well overall, but NOPS incorrectly segments more detailed areas of the natural class as buildings, utility lines, and cars.

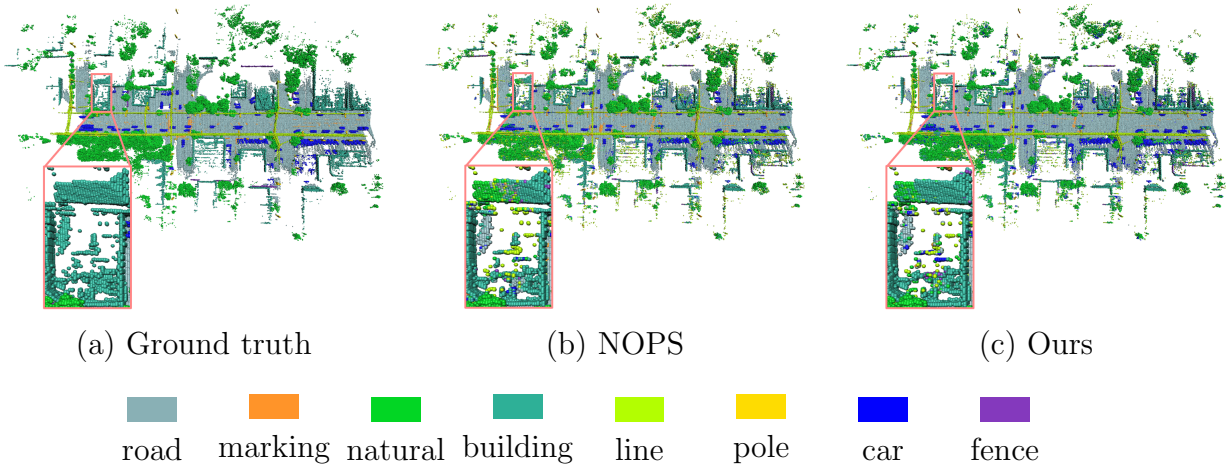

Figure 3.13: Segmentation results on the Toronto-3D-2^2 split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a-c), different colors represent different semantic classes. The zoom-in views highlight the detailed segmentation results of the building class. Both our method and NOPS show good overall performance, with NOPS segmenting relatively more detailed areas of the building class as natural elements, utility lines, cars, and fences.

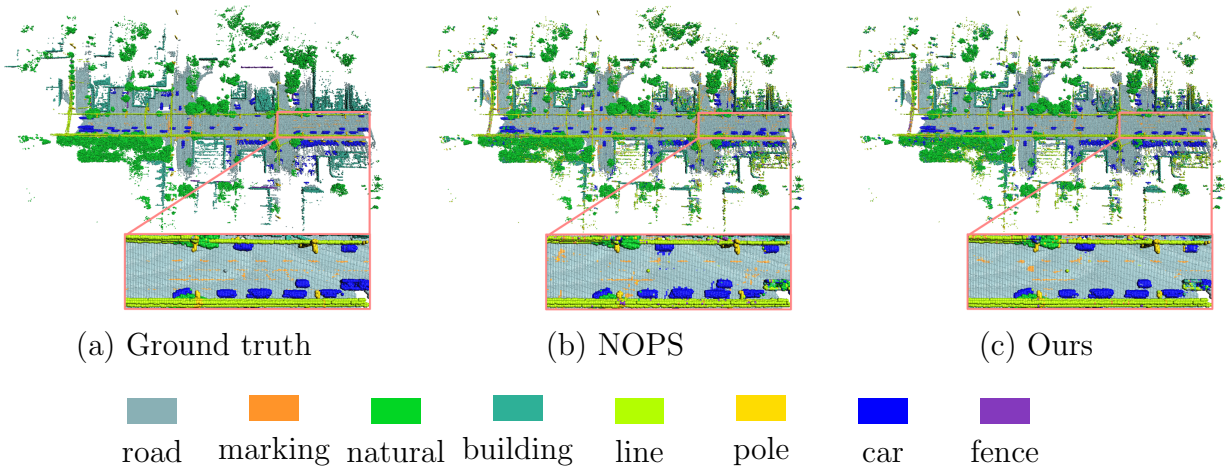

Figure 3.14: Segmentation results on the Toronto-3D-2^3 split. (a) Ground truth, (b) NOPS, (c) Ours. In images (a-c), different colors represent different semantic classes. The zoom-in views highlight the detailed segmentation results of the utility line class. Both segmentation methods show comparable overall performance; however, closer inspection reveals that our method achieves higher accuracy in finer details.

classes, achieving IoUs of 56.03% and 61.49%, respectively, compared to NOPS's 46.84% and 45.27%. The CRSM 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 Figures 3.11, 3.12, 3.13 and 3.14. 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 3.2, showcasing the method's robustness and accuracy in 3D semantic segmentation across various challenging classes.

3.3.6 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 3.3. 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

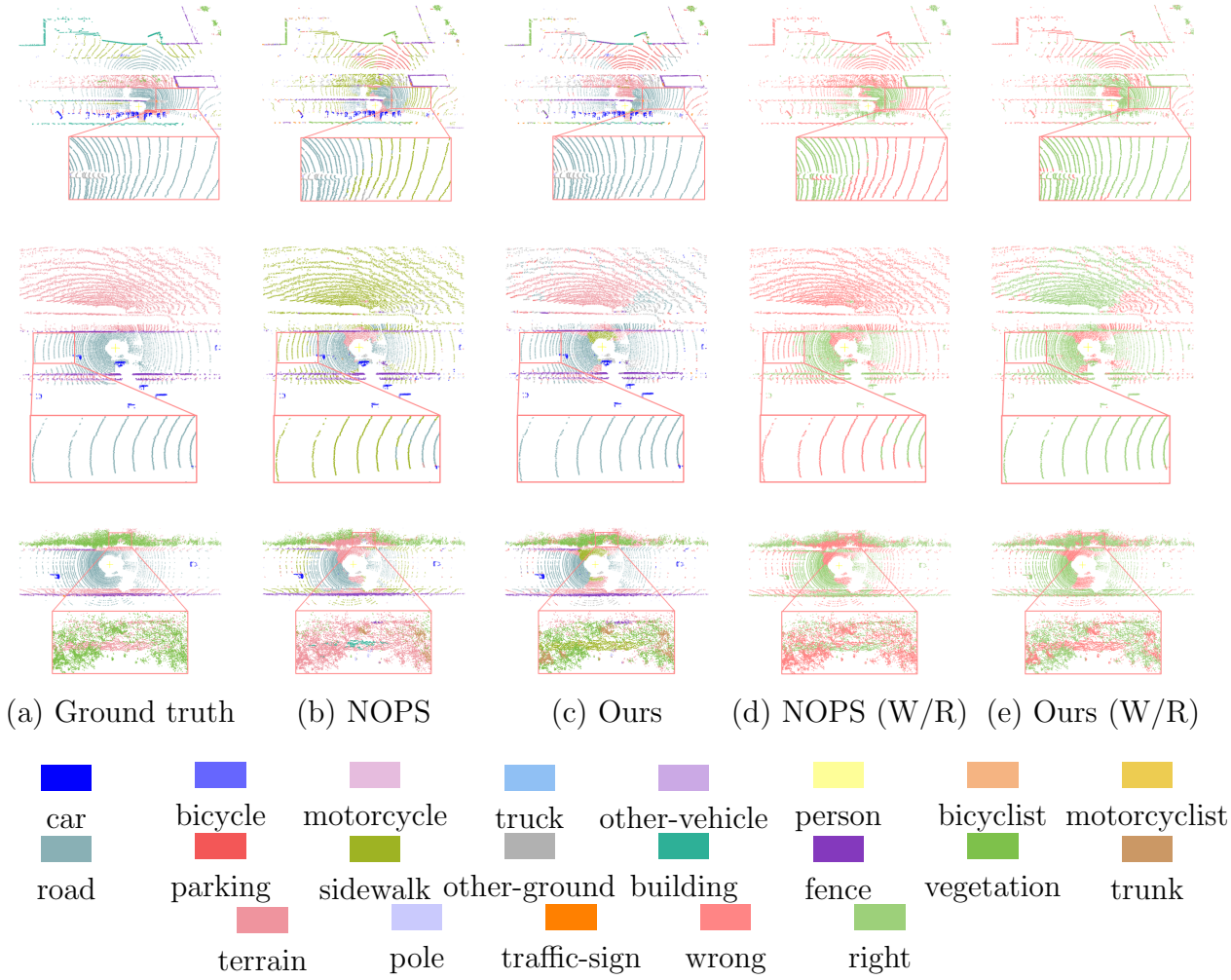

Figure 3.15: Segmentation results on the SemanticSTF-5^0 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. 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.

Table 3.3: 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

outperforms NOPS in the road and car classes. For example, the IoU for the road class is 60.65%, compared to NOPS's 47.40%. Similarly, the car class shows an IoU of 59.38%, 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 SemanticSTF-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.

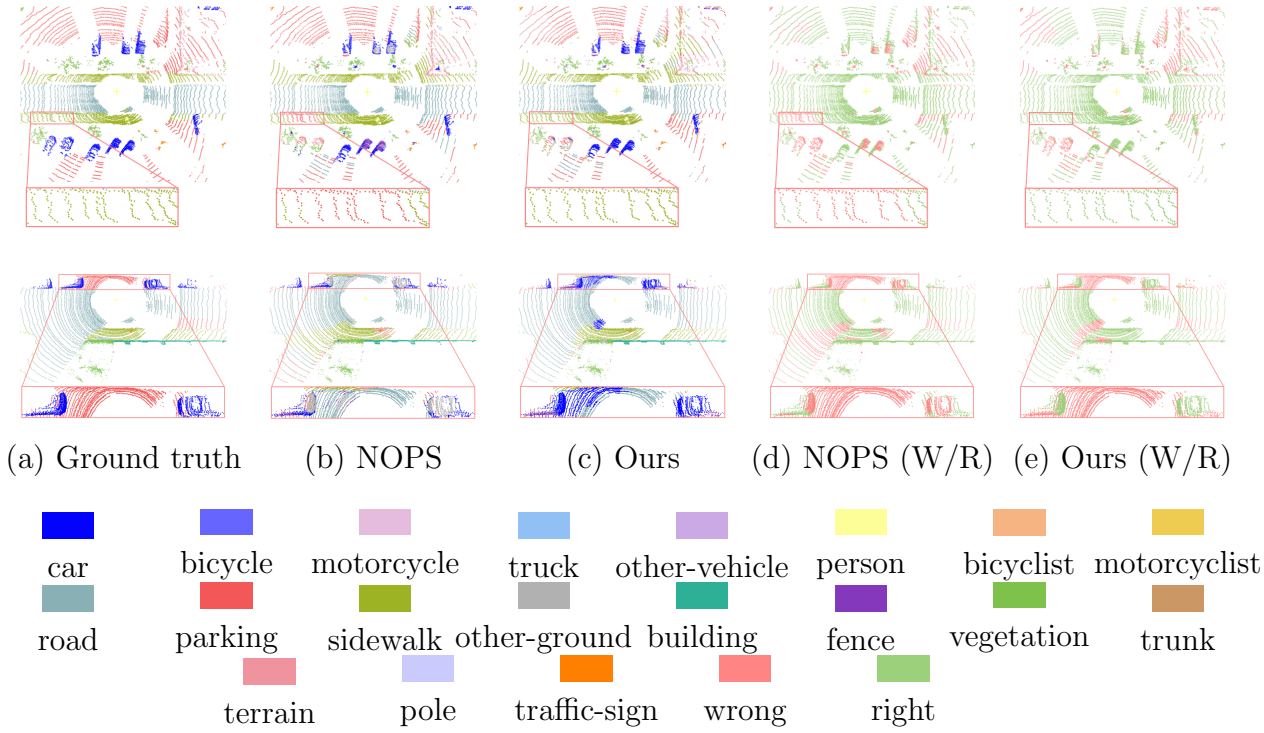

Figure 3.16: Segmentation results on the SemanticSTF-5^1 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. 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.

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 (Figures 3.15, 3.16, 3.17, and 3.18). 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.

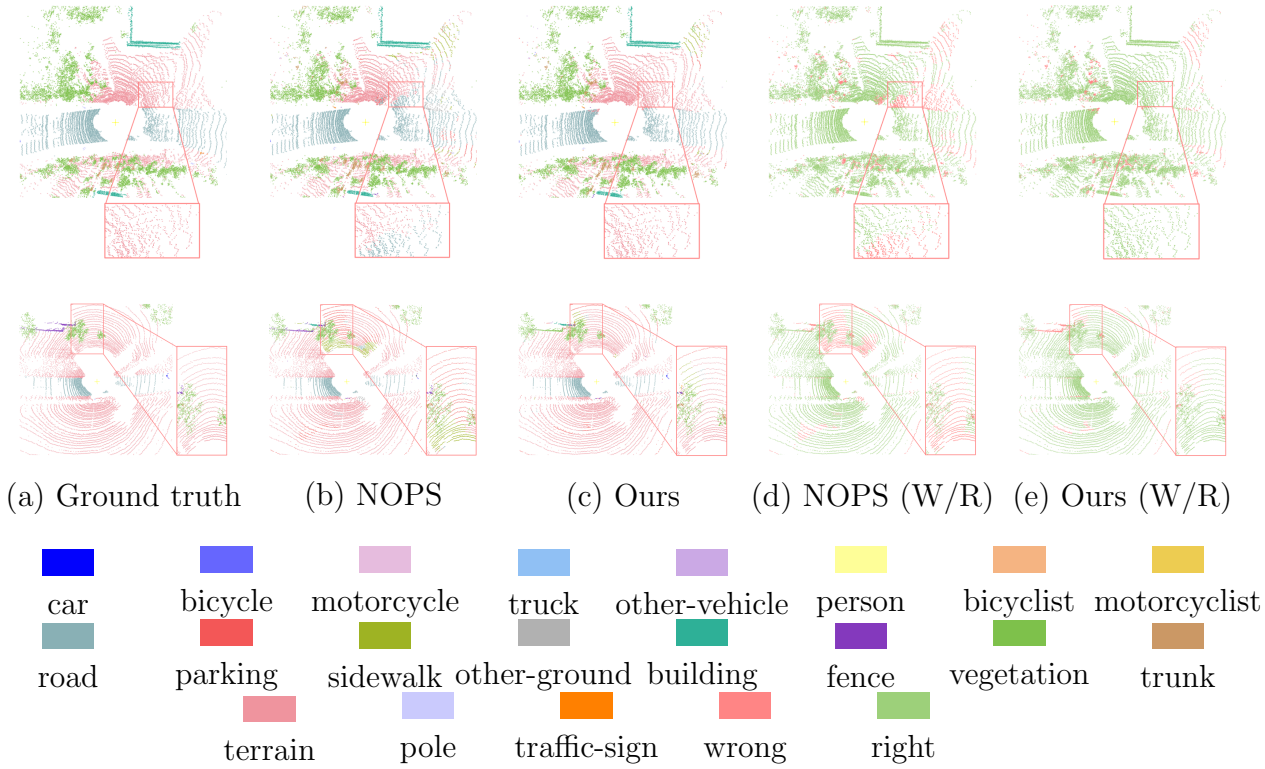

Figure 3.17: Segmentation results on the SemanticSTF-5^2 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. 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.

3.3.7 Evaluation and Visualization Results on SemanticPOSS

Table 3.4 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 cap-

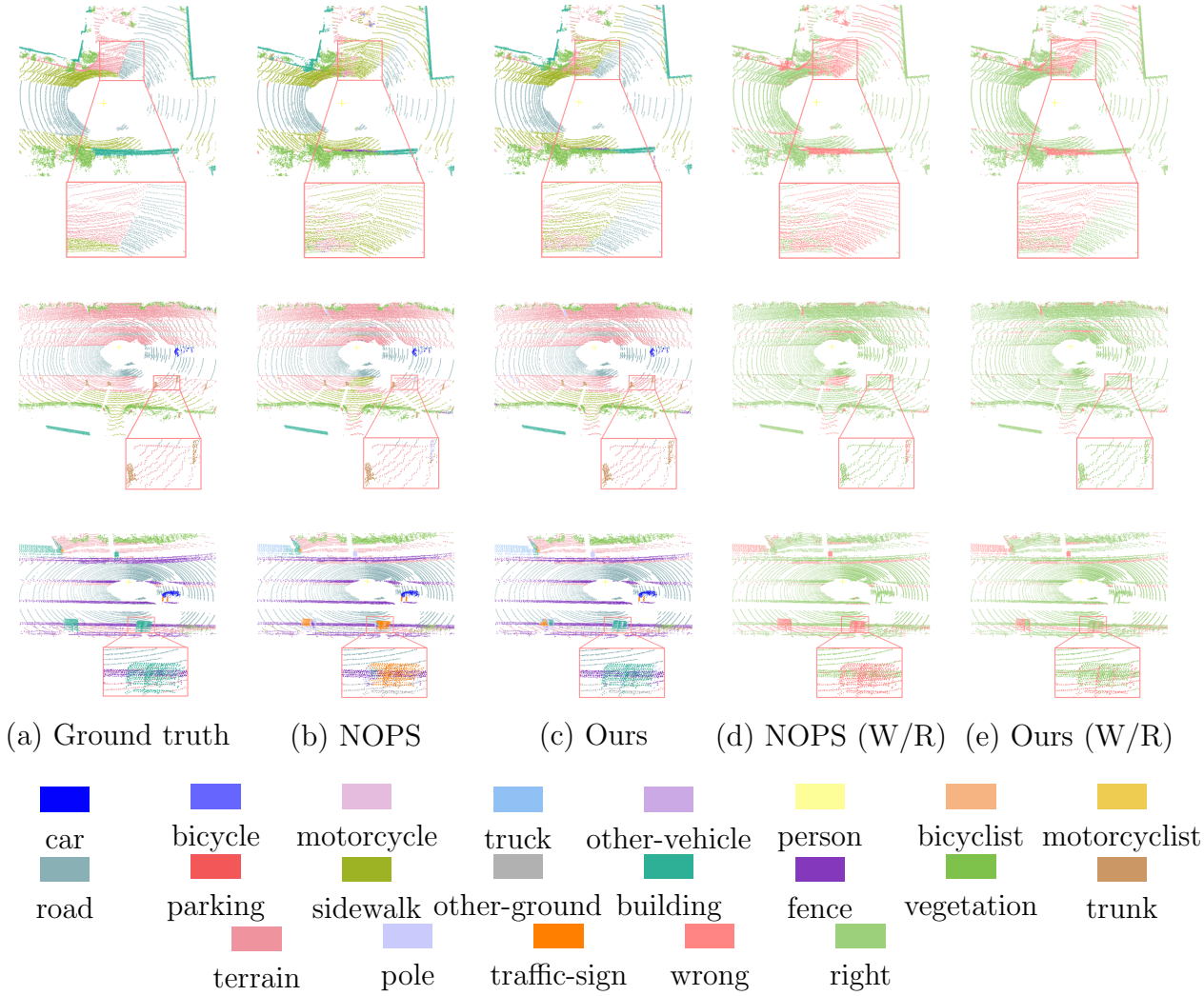

Figure 3.18: Segmentation results on the SemanticSTF-4[3] split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views show segmentation results 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.

Table 3.4: 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

turing 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 Sem-anticPOSS-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 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

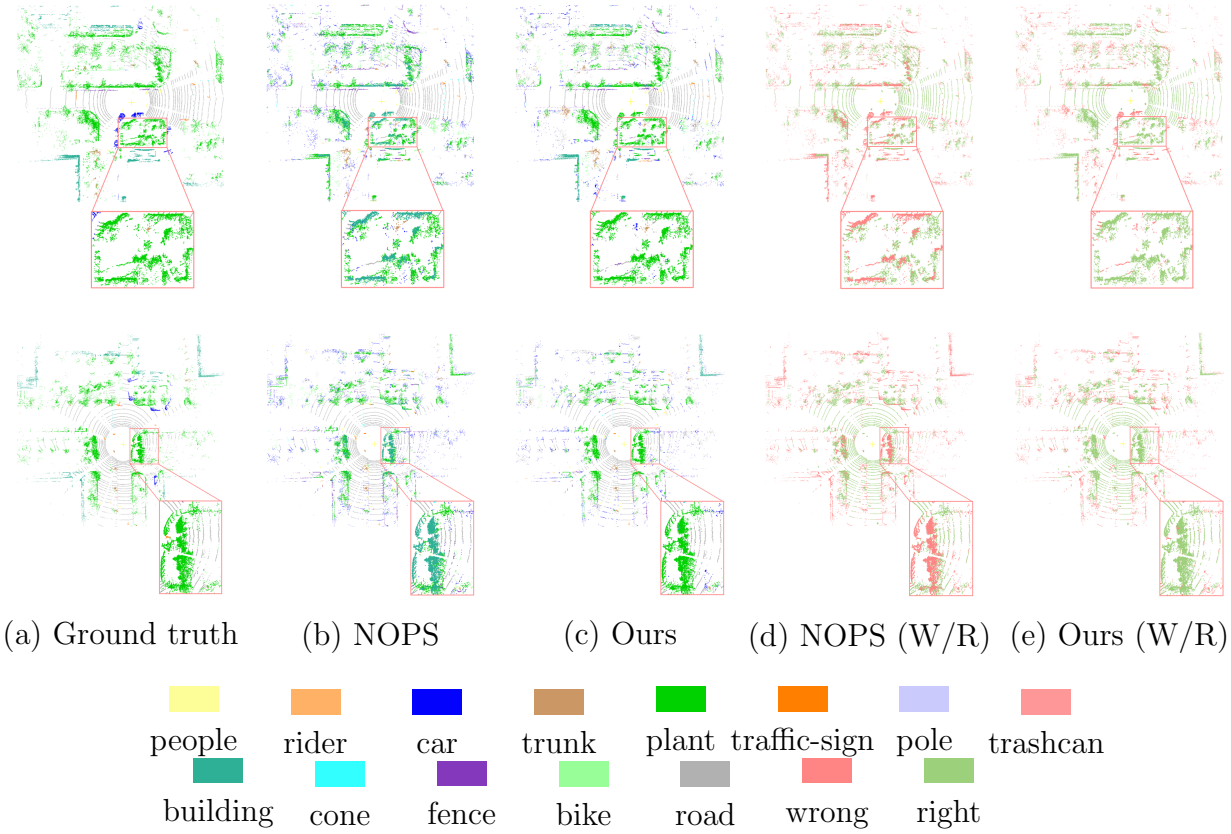

Figure 3.19: Segmentation results on the SemanticPOSS-4^0 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. The zoom-in views show detailed segmentation results. It can be observed that our method achieves better plant segmentation, whereas NOPS incorrectly segments plant parts as buildings.

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

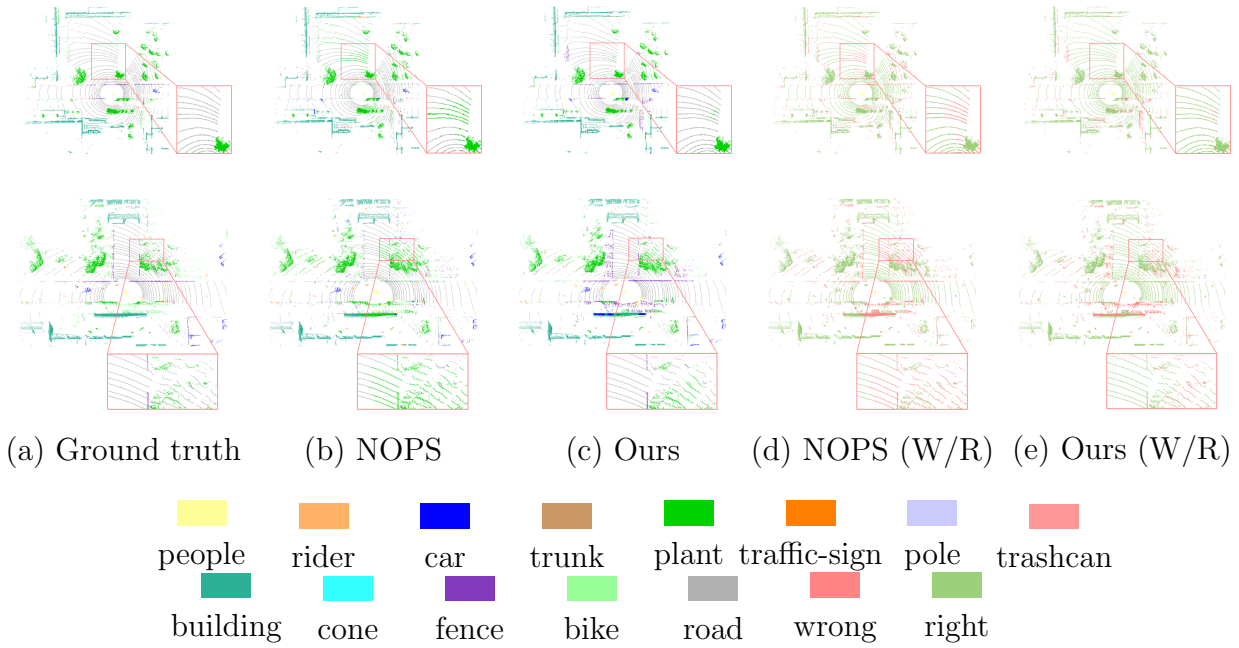

Figure 3.20: Segmentation results on the SemanticPOSS-3^1 split. (a) Ground truth, (b) NOPS, (c) Ours, (d) NOPS (W/R), (e) Ours (W/R). In images (a-c), different colors represent different semantic classes. In images (d-e), red represents segmentation errors, while green indicates correct segmentation. 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.

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 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 Figures 3.19 to 3.20. 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.

Table 3.5: 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

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.

3.3.8 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, as shown in Table 3.5. 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%, 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

Table 3.6: 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

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 3.6 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

Table 3.7: 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 3.8: 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

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 3.7, 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 3.8 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%, 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

Table 3.9: 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

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 3.9 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 nor-

malization 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 3.8 and 3.9 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.

3.4 Chapter Summary

This chapter introduces CHNCD, a geometry-aware framework for novel class discovery in 3D semantic segmentation. CHNCD addresses the challenge of discovering novel classes in large-scale point clouds through four integrated modules: VGDI couples voxel representations with original spatial coordinates via index mapping to preserve fine-grained geometry; CRSM reduces computational complexity by identifying semantically representative points in feature space; NSPM with PHM accelerates neighbor retrieval while maintaining spatial precision; and SFAM consolidates local geometric detail with global context to improve boundary fidelity. Comprehensive experiments on S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS demonstrate that CHNCD consistently outperforms the baseline NOPS method in both novel class discovery and overall segmentation. The experiments validate the effectiveness of these modules, showcasing notable enhancements in IoU for both novel and base classes. Qualitative visualizations further confirm enhanced segmentation performance in complex urban scenes, particularly for classes such as road, car, and utility lines. This work introduces key principles for geometry-aware grouping and scalable neighborhood construction. The following chapter advances representation and transfer through adaptive sequence modeling and multi-source fusion.

Chapter 4

AGDNet: Gated Multi-Source Fusion and Semantic Transfer[2]

4.1 Introduction

As established in Chapter 1, novel class discovery in 3D point cloud semantic segmentation addresses the fundamental challenge of adapting to previously unseen urban categories without exhaustive retraining. Chapter 3 introduced key principles for geometry-aware grouping and scalable neighborhood construction, demonstrating consistent improvements across S3DIS, Toronto-3D, SemanticSTF, and SemanticPOSS datasets. This chapter advances representation and transfer through adaptive sequence modeling and multi-source fusion.

Despite recent advancements in NCD for 3D point cloud semantic segmentation, several fundamental challenges persist that impede optimal performance in complex urban environments. First, the processing of point clouds with heterogeneous spatial characteristics remains problematic due to insufficient adaptive mechanisms. Fixed strategies for sampling, grouping, and feature aggregation fail to adequately capture the multi-scale geometric patterns and variable point distributions inherent in real-world urban scenes. This limitation impairs the capacity to adaptively represent complex structures across different scales and densities, thereby affecting segmentation performance across diverse urban elements.

[2]The content of the chapter has been published in ISPRS Journal of Photogrammetry and Remote Sensing with the paper entitled "Gated multi-source fusion with geometric sequence modeling for novel urban structure discovery".

Second, achieving an effective equilibrium between local detail preservation and global context integration constitutes a significant challenge in 3D point cloud analysis. The coherent integration of fine-grained geometric details with broader structural contexts represents a critical requirement, particularly for novel class discovery, where both localized geometric characteristics and their contextual relationships substantially contribute to the differentiation of previously unseen urban structures.

Third, feature fusion strategies in existing methods predominantly employ static schemes rather than context-aware adaptive mechanisms that can intelligently weight and combine complementary representations according to the specific characteristics of different urban elements.

Beyond these technical challenges, the semantic gap between known and novel classes introduces additional complexities. Establishing effective relationships between known and novel classes to enable knowledge transfer remains particularly challenging in 3D scenarios where models must rely primarily on geometric features in the absence of color and texture information. Additionally, class imbalance between well-represented known classes and limited novel class examples can lead to biased learning, making it difficult to achieve balanced performance across both category types. Current approaches lack explicit mechanisms for transferring semantic structure from base to novel classes, reducing their effectiveness in discovering rare or emerging urban elements.

To address these challenges, the following research questions (RQ) are formulated:

RQ1: How can geometric sequence modeling be adapted to handle heterogeneous spatial distributions in large-scale urban point clouds?

RQ2: How can local and global geometric context be continuously represented to capture fine structural variations for novel class discovery?

RQ3: How can multi-source features be dynamically fused through learnable gating to emphasize the most discriminative spatial or semantic information?

RQ4: How can semantic relations between known and novel classes be effectively modeled to bridge the semantic gap and mitigate class imbalance?

A comprehensive deep learning framework, termed AGDNet, is proposed to address these questions, as shown in Figure 4.1. The framework addresses each research question as follows:

- **Adaptive Geometric Sequence Modeling (RQ1):** The AGSM module employs learnable dimension weighting in adaptive farthest point sampling to automatically

prioritize geometrically discriminative spatial directions, ensuring representative coverage of structural patterns essential for novel class differentiation. Dynamic grouping adjusts neighborhood sizes according to local point density, providing appropriate geometric context across varying urban densities while avoiding the fixed neighborhood pitfalls encountered by prior methods.

- **Continuous Local-to-Global Geometric Encoding (RQ2):** The DGE module represents point clouds as learnable 3D Gaussians parameterized by position, scale, orientation, and features. This continuous probabilistic representation captures both local geometric details and global spatial contexts, providing a parametric field that ties local samples to global structure while retaining fine geometry and smoothing noise. The DGE module computes Mahalanobis distances from each point to learned Gaussian distributions, generating spatial embeddings that encode geometric relationships at multiple scales through the parameterized covariance structure.

- **Adaptive Multi-Source Feature Fusion (RQ3):** The GMSFF module employs context-aware gating mechanisms to intelligently weight and integrate features from AGSM, DGE, and MinkowskiNet [10]. The learnable gating predicts per-source weights conditioned on local features, producing a weighted mixture that emphasizes the most discriminative source where it matters and suppresses noise elsewhere. This adaptive fusion replaces static concatenation with scene-dependent weighting that follows viewpoint, density, and sensing changes.

- **Knowledge Transfer for Semantic Gap Bridging (RQ4):** Three specialized knowledge transfer objectives are introduced to establish multi-level semantic connections. Prototype Relation Loss establishes semantic connections between known and novel class prototypes through learnable prototype relationships, aligning relations among class prototypes to transfer structure from known to emerging categories. Contrastive Alignment Loss creates instance-level semantic bridges between categories, enforcing instance-level cohesion versus separation to sharpen cluster boundaries. Semantic Transfer Loss enables distribution-based knowledge propagation through KL divergence alignment, transferring the learned feature distribution structure from known to novel classes to mitigate class imbalance and reduce known–novel drift.

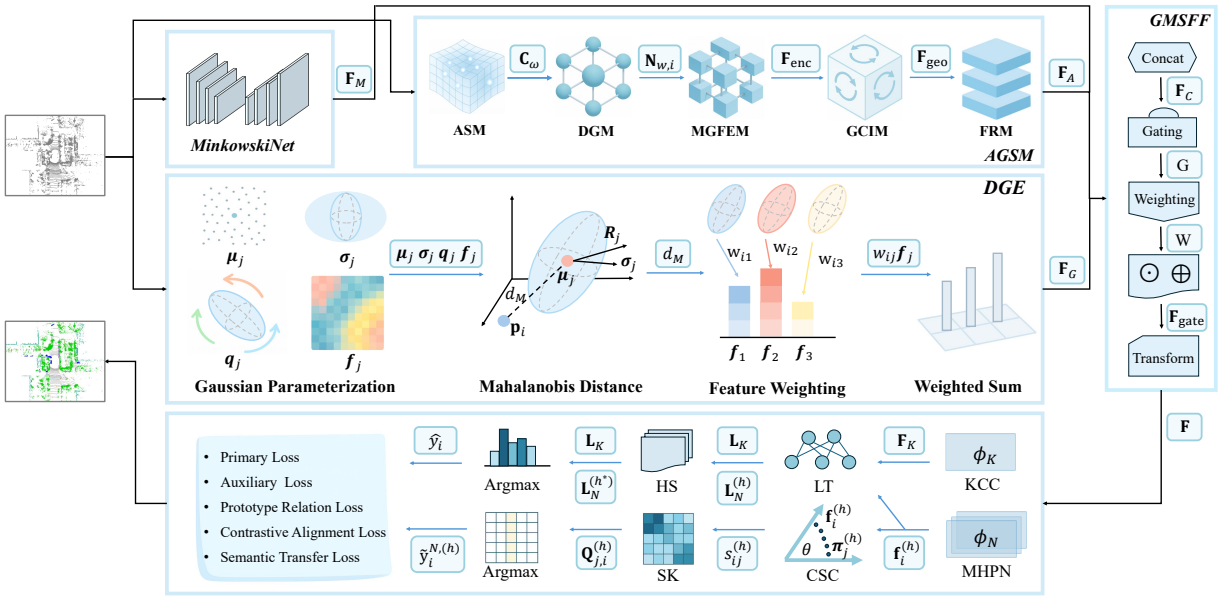

Figure 4.1: Overview of the proposed 3D novel class discovery network AGDNet. The framework extracts complementary features through three branches: MinkowskiNet for sparse convolution, AGSM for adaptive geometric modeling, and DGE for probabilistic Gaussian representation. Features are fused via GMSFF and processed through dual-branch heads for known and novel class segmentation with multi-objective learning.

4.2 Method

4.2.1 Method Overview

The proposed framework addresses the challenge of NCD in 3D point cloud segmentation, where the goal is to segment both known and previously unseen semantic categories. Figure 4.1 illustrates the complete architecture of the AGDNet framework. Given an input point cloud with coordinates $\mathbf{P} \in \mathbb{R}^{N\times 3}$, three parallel branches extract complementary feature representations. First, MinkowskiNet processes the point cloud to generate sparse convolutional features $\mathbf{F}_M$. Second, the Adaptive Geometric Sequence Modeling (AGSM) module directly processes the original point cloud coordinates $\mathbf{P}$ through five sequential components: an Adaptive Sampling Module (ASM) that selects learnable centers $\mathbf{C}_\omega$, a Dynamic Grouping Module (DGM) that forms local neighborhoods $\mathbf{N}_{\omega,i}$ based on local point density, a Multi-aspect Geometric Feature Encoding Module (MGFEM) that pro-

duces group-level features $\mathbf{F}_{\text{enc}}$, a Geometric Context Integration Module (GCIM) that yields refined features $\mathbf{F}_{\text{geo}}$, and a Feature Refinement Module (FRM) that outputs adaptive geometric features $\mathbf{F}_A$. Third, the Dynamic Gaussian Embedding (DGE) module constructs probabilistic spatial embeddings by representing the point cloud as a mixture of Gaussians. Each Gaussian is parameterized by center $\boldsymbol{\mu}_j$, scale $\boldsymbol{\sigma}_j$, orientation $\mathbf{q}_j$, and feature vector $\boldsymbol{f}_j$. The DGE module computes Mahalanobis distances d_M from each point $\mathbf{p}_i$ to each Gaussian, converts these distances into weights w_{ij}, and aggregates them into probabilistic features $\mathbf{F}_G$.

These features are subsequently fused by the Gated Multi-Source Feature Fusion (GMSFF) module through a context-aware gating mechanism. The GMSFF module concatenates $\mathbf{F}_M$, $\mathbf{F}_A$, and $\mathbf{F}_G$, applies a context-aware gating vector $\mathbf{G}$ and source-specific weight vector $\mathbf{W}$ to produce unified point-level features, and transforms the result into the unified per-point feature $\mathbf{F}$. The fused features are forwarded through a dual-branch segmentation head. For known-class prediction, the unified feature $\mathbf{F}$ is transformed by the Known-Class Classifier (KCC) ϕ_K to obtain $\mathbf{F}_K$, followed by a Linear Transformation (LT) to produce logits $\mathbf{L}_K$. For novel-class prediction, $\mathbf{F}$ is passed to a Multi-Head Prototype Network (MHPN) ϕ_N, which outputs H head-specific features $\mathbf{f}_i^{(h)}$ and prototypes $\boldsymbol{\pi}_j^{(h)}$. LT operations are applied to each head's features to generate novel-class logits $\mathbf{L}_N^{(h)}$ for all H heads. At inference, Head Selection (HS) identifies the optimal head h^* with minimal clustering loss. The known-class logits $\mathbf{L}_K$ and novel-class logits $\mathbf{L}_N^{(h^*)}$ from the selected head are then concatenated for unified prediction, with the final class assignment $\hat{y}_i$ obtained via Argmax over the combined logit space. During training, cosine similarities $s_{ij}^{(h)}$ between features and prototypes are computed in a Cosine-Similarity Calculation (CSC) block, then normalized by the Sinkhorn-Knopp algorithm (SK) [11, 71] into soft assignment matrices $Q_{j,i}^{(h)}$ to generate pseudo-labels $\tilde{y}_i^{N,(h)}$ via Argmax. To support semantic generalization, the learning process combines a segmentation loss (composed of Primary and Auxiliary components) with three transfer-driven objectives: Prototype Relation Loss, Contrastive Alignment Loss, and Semantic Transfer Loss. This architecture, together with the proposed multi-objective learning strategy, addresses novel class discovery by bridging the semantic gap between known and novel categories through feature representation, prototype learning, and semantic transfer.

4.2.2 Adaptive geometric sequence modeling

The AGSM module processes point clouds through a sequential pipeline of five functional components: adaptive sampling, dynamic neighborhood grouping, hierarchical feature encoding, geometric context integration, and feature refinement. These components collec-

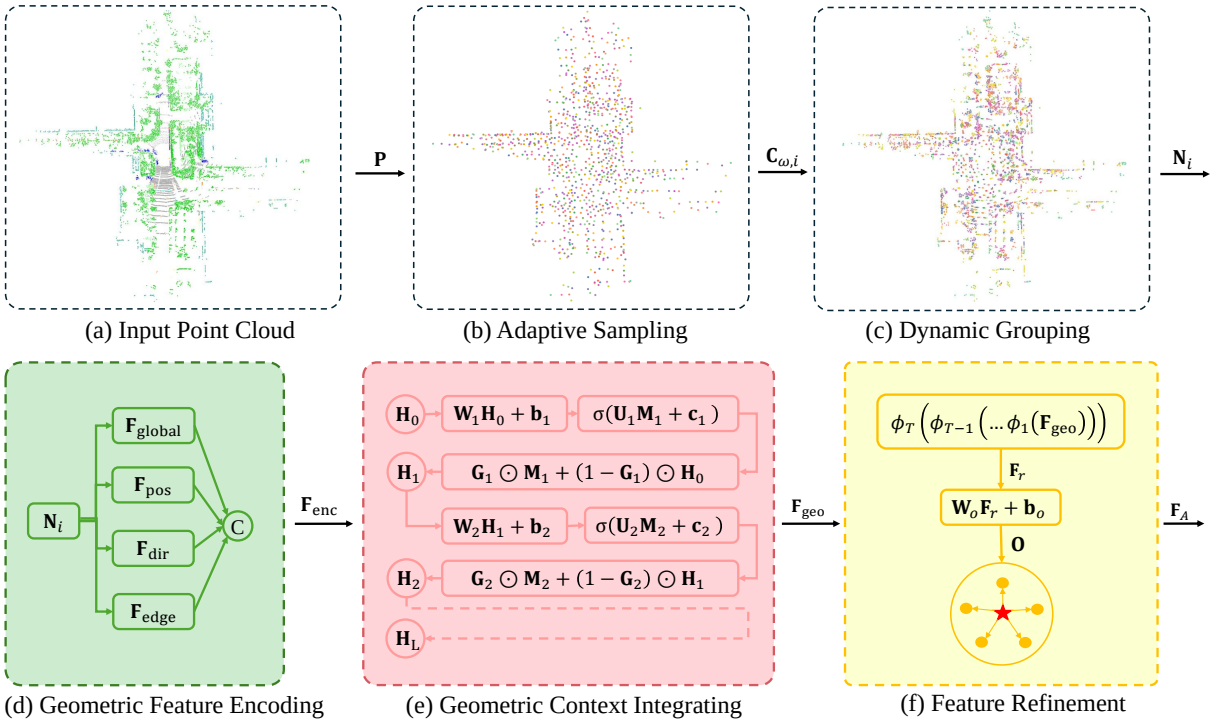

Figure 4.2: Framework of the proposed Adaptive Geometric Sequence Modeling module. $\mathbf{C}_\omega$ represents the selected centers in the point cloud, while $\mathbf{N}_{\omega,i}$ represents the dynamically grouped and normalized points. $\mathbf{F}_{\text{enc}}$ denotes the group-level features, and $\mathbf{F}_{\text{geo}}$ corresponds to the features captured by the geometric context integration module. $\mathbf{F}_A$ represents the features captured by the Adaptive Geometric Sequence Modeling module.

tively model the structural complexity and spatial variability inherent in 3D point clouds, with the complete architecture illustrated in Figure 4.2. From a distributional perspective, AGSM explicitly accounts for spatial heterogeneity through its adaptive sampling and density-aware grouping mechanisms. By dynamically adjusting to local point densities, the module ensures robust feature extraction across regions with uneven spatial distributions. This property is particularly critical for novel class discovery, where the geometric characteristics of previously unseen categories may diverge significantly from those of known classes.

Adaptive Sampling. Traditional uniform sampling strategies in urban point clouds often fail to capture the geometric diversity essential for novel class discovery, as they cannot distinguish between structurally important regions and redundant areas. Urban envi-

ronments exhibit varying geometric significance across different spatial directions, where vertical variations may be critical for distinguishing building structures while horizontal spread defines road networks. Our adaptive sampling addresses this challenge by introducing learnable dimension weighting that automatically prioritizes geometrically discriminative spatial directions, ensuring representative coverage of structural patterns that distinguish novel urban categories from known ones. This innovation enables the sampling process to adapt the distance metric for center selection, ensuring better geometric coverage of structurally important patterns essential for novel class differentiation.

To ensure representative sampling of the point cloud, we employ an Adaptive Farthest Point Sampling (AFPS) strategy. This approach introduces a learnable weighting vector $\boldsymbol{\omega} \in \mathbb{R}^3$ to adaptively adjust the importance of each spatial dimension. The weighted coordinates $\mathbf{P}_\omega$ are computed as:

$$\mathbf{P}_\omega = \mathbf{P} \odot \boldsymbol{\omega}, \tag{4.1}$$

where $\mathbf{P} \in \mathbb{R}^{N\times 3}$ denotes the original coordinates of a point cloud with N points, and $\odot$ represents element-wise multiplication.

The AFPS algorithm selects a set of sampling centers $\mathbf{C}_\omega \in \mathbb{R}^{G\times 3}$ by iteratively choosing the farthest point from the previously selected centers based on the weighted Euclidean distance, where G denotes the number of sampling centers used to form G local groups in the subsequent feature encoding stage:

$$\mathbf{c}_{\omega,i+1} = \arg \max_{\mathbf{p}_\omega \in \mathbf{P}_\omega} \min_{\mathbf{c}_\omega \in \mathbf{C}_\omega} \|\mathbf{p}_\omega - \mathbf{c}_\omega\|^2, \tag{4.2}$$

where $\|\cdot\|^2$ denotes the squared Euclidean distance in the weighted coordinate space.

Dynamic Grouping. Fixed neighborhood sizes cannot adequately capture geometric context across urban environments with heterogeneous point distributions, leading to either insufficient context in sparse regions or redundant information in dense areas. This limitation particularly affects novel class discovery, where appropriate geometric context is crucial for distinguishing new urban structures from existing categories. Our dynamic grouping mechanism solves this problem by adaptively adjusting neighborhood sizes based on local point density, ensuring optimal geometric context capture for diverse urban elements. This density-aware adaptation enables the model to extract meaningful geometric patterns regardless of local point distributions, providing robust feature extraction essential for reliable novel class discovery.

Following the selection of sampling centers, we apply a dynamic grouping strategy that adapts the group size based on local point density. For each center $\mathbf{c}_{\omega,i}$, the local density ρ_i is estimated using the distance to its $k_{\min}$-th nearest neighbor. The group size g_i is then dynamically adjusted according to the local density:

$$g_i = g_{\min} + \left(\frac{\rho_i - \rho_{\min}}{\rho_{\max} - \rho_{\min}} \right) (g_{\max} - g_{\min}), \tag{4.3}$$

where $g_{\min}$ and $g_{\max}$ are the predefined lower and upper bounds of group size, and $\rho_{\min}$ and $\rho_{\max}$ are the minimum and maximum density values across all centers. We use the K-Nearest Neighbors (KNN) algorithm to select g_i neighboring points for each center, and normalize the local coordinates by subtracting the center location:

$$\mathbf{N}_{\omega,i} = \mathbf{p}_{\omega,i} - \mathbf{c}_{\omega,i}, \tag{4.4}$$

where $\mathbf{p}_{\omega,i}$ is a point in the neighborhood of center $\mathbf{c}_{\omega,i}$, and $\mathbf{N}_{\omega,i}$ denotes the normalized neighborhood.

Multi-aspect Geometric Feature Encoding. Novel urban structures often exhibit subtle geometric differences from known categories that cannot be captured through single-aspect feature encoding, requiring comprehensive characterization of spatial relationships at multiple scales and perspectives. The challenge lies in simultaneously capturing point-wise surface properties, positional relationships, directional characteristics, and structural connections that collectively define the geometric signature of urban elements. Our multi-aspect geometric encoding addresses this challenge by implementing parallel encoding pathways that extract complementary geometric representations including local spatial patterns, positional and directional cues, structural relationships, and global contextual information. This comprehensive approach ensures that distinctive geometric signatures of novel urban categories are captured and preserved through the encoding process, providing the rich spatial understanding necessary for effective novel class discovery in complex urban environments.

The dynamically grouped points $\mathbf{N}_{\omega,i}$ from the previous stage undergo feature encoding to extract complementary geometric representations. Our encoding approach captures multiple aspects of the point cloud geometry through a sequential processing pipeline.

To capture local spatial patterns, we apply a point-level encoder to the transposed neighborhood:

$$\mathbf{F}_{\text{point}} = \text{PointEnc}(\mathbf{N}_{\omega,i}^T), \tag{4.5}$$

where PointEnc consists of a convolutional layer followed by batch normalization and ReLU activation, and $\mathbf{F}_{\text{point}}$ denotes the resulting point-wise features. Position and directional encodings are extracted using a shared geometry encoder:

$$\mathbf{F}_{\text{pos}} = \text{GeoEnc}(\mathbf{N}_{\omega,i}^{T}), \quad \mathbf{F}_{\text{dir}} = \text{GeoEnc}\left(\left(\frac{\mathbf{N}_{\omega,i}}{\|\mathbf{N}_{\omega,i}\|}\right)^{T}\right), \tag{4.6}$$

where $\mathbf{F}_{\text{pos}}$ and $\mathbf{F}_{\text{dir}}$ represent position and direction features extracted from absolute and normalized coordinates, respectively.

To capture structural relations, we compute edge features between neighborhood points and representative points from the dynamic grouping stage:

$$\mathbf{E} = \{\text{EdgeEnc}([\mathbf{p}_i, \mathbf{p}_j]^{T}) \mid \mathbf{p}_j \in \mathcal{S}, \forall \mathbf{p}_i \in \mathbf{N}_{\omega,i}\}, \tag{4.7}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ denote neighborhood and representative points, respectively, $\mathcal{S}$ includes the nearest, farthest, and intermediate points, and EdgeEnc is the edge encoding function. Edge features are aggregated via max and average pooling, followed by compression:

$$\mathbf{F}_{\text{edge}} = \text{EdgeComp}([\text{Max}(\mathbf{E}), \text{Avg}(\mathbf{E})]), \tag{4.8}$$

where $\text{Max}(\cdot)$ and $\text{Avg}(\cdot)$ denote max and average pooling, and EdgeComp compresses the concatenated features. Global context is encoded by aggregating the point features:

$$\mathbf{F}_{\text{global}} = \text{Max}(\mathbf{F}_{\text{point}})^{\text{expand}}, \tag{4.9}$$

where ^{expand} denotes broadcasting to match the feature dimensions of other modules.

Finally, all features are integrated through concatenation and transformation:

$$\mathbf{F}_{\text{enc}} = \text{Fusion}([\mathbf{F}_{\text{global}}, \mathbf{F}_{\text{pos}}, \mathbf{F}_{\text{dir}}, \mathbf{F}_{\text{edge}}]), \tag{4.10}$$

where $[\cdot]$ denotes channel-wise concatenation, and Fusion transforms the concatenated features to the desired representation dimension.

Geometric Context Integration. While the multi-aspect geometric encoding effectively captures comprehensive local structural patterns, novel class discovery in urban environments benefits from integrating these detailed representations with broader spatial dependencies and contextual relationships that span across multiple neighborhood scales. The challenge lies in combining the rich local geometric information with long-range spatial

context while preserving the discriminative details essential for distinguishing novel categories from structurally similar known classes. Our geometric context integration module addresses this through hierarchical processing with adaptive gating mechanisms that selectively propagate and refine geometric features across multiple scales. The core innovation lies in the gated residual architecture that enables the network to learn which geometric relationships are most relevant for novel class differentiation, allowing global structural patterns to enhance local feature representations while maintaining the spatial specificity required for accurate segmentation.

The fused features $\mathbf{F}_{\text{enc}}$ are refined through a series of Geometric Context Integration Units (GCIUs), each incorporating adaptive gating, feature mixing, and residual connections. The mixed features at layer l are computed as:

$$\mathbf{M}_l = \mathbf{W}_l \mathbf{H}_{l-1} + \mathbf{b}_l, \tag{4.11}$$

where $l \in \{1, \ldots, L\}$ indexes the GCIU layer, and L is the total number of integration layers. $\mathbf{H}_{l-1}$ is the input from the previous layer, and $\mathbf{W}_l$ and $\mathbf{b}_l$ are learnable parameters. The corresponding gating mask is given by:

$$\mathbf{G}_l = \sigma(\mathbf{U}_l \mathbf{M}_l + \mathbf{c}_l), \tag{4.12}$$

where $\mathbf{U}_l$ and $\mathbf{c}_l$ are learnable parameters, and $\sigma(\cdot)$ is the element-wise sigmoid function. The gated residual output is computed as:

$$\mathbf{H}_l = \mathbf{G}_l \odot \mathbf{M}_l + (1 - \mathbf{G}_l) \odot \mathbf{H}_{l-1}. \tag{4.13}$$

After passing through all L GCIU layers, the final output $\mathbf{H}_L$ is obtained. It is then globally average pooled to obtain a compact representation:

$$\mathbf{F}_{\text{geo}} = \frac{1}{G} \sum_{i=1}^{G} \mathbf{H}_L[i, :], \tag{4.14}$$

where $\mathbf{F}_{geo}$ is the final geometric context feature summarizing all group-level representations.

Feature Refinement and Output Generation. The transition from compact group-level geometric representations to point-wise predictions presents a critical challenge in maintaining spatial consistency and semantic alignment across the entire point cloud, particularly for novel class discovery where prediction accuracy must be preserved at every

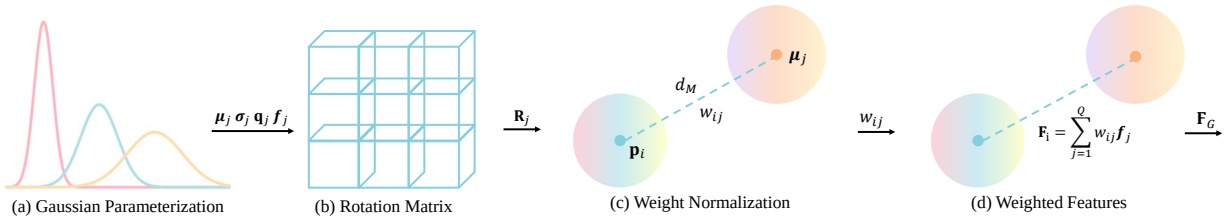

Figure 4.3: Framework of the proposed Dynamic Gaussian Embedding module. Each Gaussian g_j in the Dynamic Gaussian Embedding module is parameterized by its center $\boldsymbol{\mu}_j \in \mathbb{R}^3$, scale $\boldsymbol{\sigma}_j \in \mathbb{R}^3$, rotation $\mathbf{q}_j \in \mathbb{R}^4$ (represented as a unit quaternion), and feature vector $\boldsymbol{f}_j \in \mathbb{R}^D$. The rotation matrix $\mathbf{R}_j$ is computed from the quaternion $\mathbf{q}_j$. w_{ij} represents the contribution of each Gaussian g_j to the point $\mathbf{p}_i$, and $\mathbf{F}_G$ denotes the features captured by the DGE module.

spatial location. Propagating the rich contextual understanding captured at the group level to individual points requires careful preservation of spatial correspondence while ensuring consistent application of novel class discoveries throughout the urban scene. To address this challenge, our feature refinement and output generation module employs a multi-stage processing pipeline that transforms compact geometric context into point-wise features. The key contribution is the adaptive refinement process that preserves the discriminative power of learned geometric representations during the transition to full spatial resolution, enabling accurate segmentation of both known and novel classes throughout the point cloud.

The geometric context feature $\mathbf{F}_{\text{geo}}$ is further processed by a multi-layer perceptron (MLP) composed of T layers, each including normalization and activation:

$$\mathbf{F}_r = \phi_T(\phi_{T-1}(...\phi_1(\mathbf{F}_{\text{geo}}))), \tag{4.15}$$

where ϕ_t denotes the t-th MLP layer and T is the total number of layers. $\mathbf{F}_r$ is the refined feature representation.

The refined feature is then projected to the target output dimension through a linear transformation:

$$\mathbf{O} = \mathbf{W}_o\mathbf{F}_r + \mathbf{b}_o, \tag{4.16}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are the output layer's learnable parameters. To restore point-wise resolution, the output is broadcast to all N points:

$$\mathbf{O}_e = \mathbf{O} \otimes \mathbf{1}_{1\times N}, \tag{4.17}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{1}_{1\times N}$ is a vector of ones.

Finally, the broadcasted output is reshaped to match the batch size B and the number of points N:

$$\mathbf{F}_A = \{\mathbf{O}_e^{(i,j)} \in \mathbb{R}^D : i \in \{1,\ldots,B\}, j \in \{1,\ldots,N\}\}, \tag{4.18}$$

where $\mathbf{F}_A \in \mathbb{R}^{N\times D}$ is the final prediction aligned with the original point cloud structure, and D is the target output dimension.

4.2.3 Dynamic Gaussian Embedding

Point cloud representations face the fundamental challenge of balancing local geometric detail preservation with global spatial context integration, particularly for novel class discovery where effective generalization across varying spatial configurations is essential. Traditional discrete point-based methods struggle to capture the continuous nature of geometric distributions, limiting their ability to interpolate between known training patterns and novel spatial arrangements encountered in complex 3D scenes. The proposed DGE module addresses this challenge by representing point clouds as mixtures of learnable 3D Gaussians that provide continuous probabilistic modeling of spatial structures. The key innovation lies in the differentiable Gaussian parameterization that simultaneously encodes local geometric details through position and orientation parameters while capturing global spatial relationships through learnable feature vectors, enabling robust interpolation across unseen geometric configurations for novel class generalization.

The probabilistic formulation of DGE provides a principled way to model uncertainty and distributional variation between known and novel classes, as illustrated in Figure 4.3. By representing the point cloud as a mixture of learnable Gaussians, DGE captures the continuous structure of geometric distributions, allowing the model to interpolate across unseen spatial configurations. This continuity facilitates better generalization to novel categories that deviate from the training distribution.

Given a point cloud $\mathbf{P}$ consisting of N points $\{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^N$, we define a set of Q Gaussian primitives, indexed by $j = 1, 2, \ldots, Q$. Each Gaussian g_j is parameterized by its center $\boldsymbol{\mu}_j \in \mathbb{R}^3$, scale $\boldsymbol{\sigma}_j \in \mathbb{R}^3$, rotation $\mathbf{q}_j \in \mathbb{R}^4$ (represented as a unit quaternion), and feature vector $\boldsymbol{f}_j \in \mathbb{R}^D$, where D is the feature dimension.

The rotation matrix $\mathbf{R}_j$ is computed from the quaternion $\mathbf{q}_j = (w, x, y, z)$:

$$\mathbf{R}_j = \begin{bmatrix} 1-2(y^2+z^2) & 2(xy-wz) & 2(xz+wy) \\ 2(xy+wz) & 1-2(x^2+z^2) & 2(yz-wx) \\ 2(xz-wy) & 2(yz+wx) & 1-2(x^2+y^2) \end{bmatrix}. \tag{4.19}$$

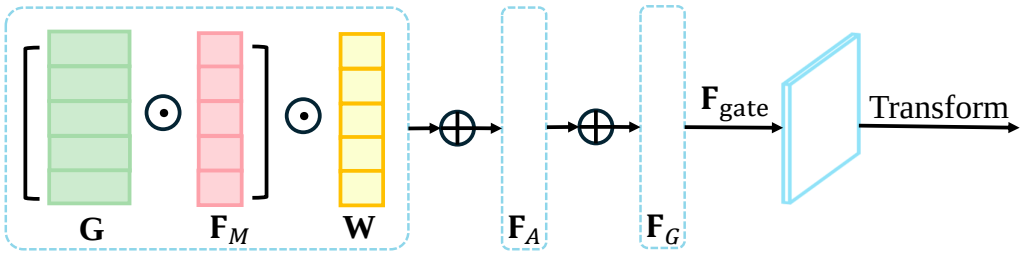

Figure 4.4: Framework of the proposed GMSFF module. The context-aware gating mechanism result is denoted by $\mathbf{G}$, and the source-specific importance weights are represented by $\mathbf{W}$. The operations $\odot$, and $\oplus$ indicate broadcasted element-wise multiplication and element-wise addition, respectively. The fused output $\mathbf{F}_{\text{gate}}$ is obtained by applying gated and weighted combination of the input features.

For each point $\mathbf{p}_i$, the Mahalanobis squared distance $d_M^2(\mathbf{p}_i, g_j)$ to the Gaussian g_j is computed as:

$$d_M^2(\mathbf{p}_i, g_j) = (\mathbf{R}_j(\mathbf{p}_i - \boldsymbol{\mu}_j))^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{R}_j(\mathbf{p}_i - \boldsymbol{\mu}_j)), \tag{4.20}$$

where $\boldsymbol{\Sigma}_j$ denotes the diagonal covariance matrix with entries $\sigma_{j,d}^2$ along the diagonal. The contribution of each Gaussian g_j to the point $\mathbf{p}_i$ is determined by the weight:

$$w_{ij} = \frac{\exp\left(-\frac{1}{2} d_M^2(\mathbf{p}_i, g_j)\right)}{\sum_{k=1}^{Q} \exp\left(-\frac{1}{2} d_M^2(\mathbf{p}_i, g_k)\right)}. \tag{4.21}$$

Finally, the feature representation for each point $\mathbf{p}_i$ is computed as a weighted sum of the Gaussian features:

$$\mathbf{F}_i = \sum_{j=1}^{Q} w_{ij} \boldsymbol{f}_j, \tag{4.22}$$

where $\mathbf{F}_i \in \mathbb{R}^D$ denotes the encoded feature for point $\mathbf{p}_i$, obtained by aggregating Gaussian embeddings. The final point-wise feature representation for the entire point cloud, denoted $\mathbf{F}_G \in \mathbb{R}^{N \times D}$, is obtained by stacking the features $\mathbf{F}_i$ for all points $i = 1, \ldots, N$.

4.2.4 Gated Multi-source Feature Fusion

Integrating diverse feature representations from multiple specialized modules presents significant challenges in maintaining the discriminative power of each source while enabling effective cross-modal information exchange. Simply concatenating features from different modules can lead to feature interference and suboptimal fusion, particularly when individual modules capture complementary but potentially conflicting geometric cues essential for novel class discovery. The GMSFF module addresses this challenge through adaptive weighting mechanisms that intelligently balance contributions from different feature sources based on input context. The approach combines context-aware gating with source-specific importance weighting to achieve optimal integration of complementary representations while preserving the distinctive strengths of each component.

The GMSFF module, illustrated in Figure 4.4, employs a gating mechanism to selectively combine diverse feature representations. Given an input point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$, we obtain three feature representations: $\mathbf{F}_M \in \mathbb{R}^{N\times D}$ from MinkowskiNet, $\mathbf{F}_A \in \mathbb{R}^{N\times D}$ from AGSM, and $\mathbf{F}_G \in \mathbb{R}^{N\times D}$ from DGE. The GMSFF module processes these features as follows: First, the features are concatenated:

$$\mathbf{F}_C = [\mathbf{F}_M; \mathbf{F}_A; \mathbf{F}_G] \in \mathbb{R}^{N\times 3D}. \tag{4.23}$$

Next, a context-aware gating mechanism is applied:

$$\mathbf{G} = \sigma(W_g \mathbf{F}_C + b_g) \in \mathbb{R}^{N\times D}, \tag{4.24}$$

where $W_g \in \mathbb{R}^{D\times 3D}$, $b_g \in \mathbb{R}^D$, and $\sigma(\cdot)$ denotes the sigmoid function. Concurrently, source-specific importance weights are computed:

$$\mathbf{W} = \text{softmax}(W_w \mathbf{F}_C + b_w) \in \mathbb{R}^{N\times 3}, \tag{4.25}$$

where $W_w \in \mathbb{R}^{3\times 3D}$ and $b_w \in \mathbb{R}^3$. The gated and weighted feature fusion is performed as:

$$\begin{aligned} \mathbf{F}_{\text{gate}} = {} & \mathbf{W}[:, 0{:}1] \odot (\mathbf{G} \odot \mathbf{F}_M) \\ & + \mathbf{W}[:, 1{:}2] \odot (\mathbf{G} \odot \mathbf{F}_A) \\ & + \mathbf{W}[:, 2{:}3] \odot (\mathbf{G} \odot \mathbf{F}_G), \end{aligned} \tag{4.26}$$

where $\odot$ denotes broadcasted element-wise multiplication. The gating vector $\mathbf{G} \in \mathbb{R}^{N\times D}$ modulates feature amplitudes based on input context, while the weight matrix $\mathbf{W} \in \mathbb{R}^{N\times 3}$

assigns adaptive importance to each feature source. Finally, a feature transformation is applied:

$$\mathbf{F} = \phi(W_t \mathbf{F}_{\text{gate}} + b_t), \tag{4.27}$$

where $W_t \in \mathbb{R}^{D \times D}$, $b_t \in \mathbb{R}^D$, and $\phi(\cdot)$ is a non-linear activation function.

The resulting $\mathbf{F} \in \mathbb{R}^{N \times D}$ represents the fused features that adaptively combine global context, sequential geometric patterns, and local structural details. The learnable parameters in GMSFF enable dynamic adjustment during training, allowing the fusion strategy to adapt to diverse point cloud structures.

4.2.5 Novel Class Discovery Implementation

To support semantic segmentation of both known and previously unseen categories, the unified feature representation $\mathbf{F} \in \mathbb{R}^{N \times D}$ is processed through two specialized branches: one for known class prediction and one for novel class discovery.

Known Class Prediction. For the known class branch, features $\mathbf{F}$ are projected through a Known-Class Classifier (KCC) ϕ_K to obtain known-specific features $\mathbf{F}_K \in \mathbb{R}^{N \times D_K}$, where D_K is the output dimension. A linear transformation then produces classification logits:

$$\mathbf{L}_K = \mathbf{W}_K \mathbf{F}_K + \mathbf{b}_K, \tag{4.28}$$

where $\mathbf{W}_K \in \mathbb{R}^{K \times D_K}$ and $\mathbf{b}_K \in \mathbb{R}^K$ are learnable weights and biases for K known classes.

Novel Class Discovery. For novel categories, the unified feature $\mathbf{F}$ is passed to a Multi-Head Prototype Network (MHPN) ϕ_N, which produces head-specific features $\mathbf{f}_i^{(h)} \in \mathbb{R}^{D_N}$ for each point i and head $h \in \{1, \ldots, H\}$, where D_N is the dimensionality of novel-class features. Each head maintains N_{novel} learnable prototypes $\boldsymbol{\pi}_j^{(h)} \in \mathbb{R}^{D_N}$.

Cosine similarities between features and prototypes are computed in the Cosine-Similarity Calculation (CSC) block as:

$$s_{ij}^{(h)} = \frac{\langle \mathbf{f}_i^{(h)}, \boldsymbol{\pi}_j^{(h)} \rangle}{\|\mathbf{f}_i^{(h)}\| \, \|\boldsymbol{\pi}_j^{(h)}\|}, \tag{4.29}$$

where $s_{ij}^{(h)}$ is the similarity between point i and prototype j in head h. These similarities are transformed into balanced assignment matrices $\mathbf{Q}^{(h)} \in \mathbb{R}^{N_{\text{novel}} \times N}$ using the Sinkhorn–Knopp algorithm, enforcing the constraints:

$$\sum_{j=1}^{N_{\text{novel}}} Q_{j,i}^{(h)} = 1, \quad \sum_{i=1}^{N} Q_{j,i}^{(h)} = \frac{N}{N_{\text{novel}}}. \tag{4.30}$$

During training, pseudo-labels $\tilde{y}_i^{N,(h)}$ are derived by selecting the class with maximum assignment probability:

$$\tilde{y}_i^{N,(h)} = \arg\max_j Q_{j,i}^{(h)}, \tag{4.31}$$

which serve as supervision targets for the novel-class component of the segmentation loss. To identify the most reliable clustering result at inference, the head h^* with the minimal clustering loss is selected:

$$h^* = \arg\min_h \mathcal{L}_{\text{cluster}}^{(h)}, \tag{4.32}$$

where $\mathcal{L}_{\text{cluster}}^{(h)}$ corresponds to the novel-class component of the primary segmentation loss $\mathcal{L}_{\text{primary}}^{(h)}$ that supervises pseudo-label assignments.

The novel class logits $\mathbf{L}_N^{(h)}$ are computed by applying a linear transformation to the head-specific features:

$$\mathbf{L}_N^{(h)} = \mathbf{W}_N^{(h)} \mathbf{f}^{(h)} + \mathbf{b}_N^{(h)}, \tag{4.33}$$

where $\mathbf{W}_N^{(h)}$ and $\mathbf{b}_N^{(h)}$ are learnable parameters for head h. Known class logits and novel class logits from the selected head are concatenated for unified prediction:

$$\hat{y}_i = \arg\max([\mathbf{L}_{K,i}, \mathbf{L}_{N,i}^{(h^*)}]), \tag{4.34}$$

where $\mathbf{L}_{K,i}$ and $\mathbf{L}_{N,i}^{(h^*)}$ represent the known and novel class logits for point i respectively.

To mitigate the imbalance between known and novel categories, the model processes them in separate branches, allowing dedicated capacity for underrepresented novel structures. Moreover, the Sinkhorn–Knopp algorithm enforces uniform pseudo-label distributions across prototypes, promoting balanced cluster formation and avoiding dominance by majority patterns in the data.

4.2.6 Multi-objective Learning for Novel Class Discovery

Novel class discovery in 3D point cloud segmentation faces fundamental challenges that traditional supervised learning approaches cannot address: how to identify meaningful semantic categories without labeled examples, how to transfer knowledge from well-represented known classes to underrepresented novel categories, and how to establish robust decision boundaries when dealing with severe class imbalance between labeled and unlabeled data. While existing novel class discovery methods have made significant progress, many approaches primarily focus on clustering-based objectives with limited mechanisms for systematic knowledge transfer from known to novel class representations, potentially constraining their effectiveness in discovering semantically meaningful categories in complex urban point cloud environments. The proposed multi-objective learning framework addresses these challenges by implementing specialized knowledge transfer mechanisms that bridge the semantic gap between known and novel classes at multiple levels. This coordinated learning strategy combines prototype-level semantic alignment, instance-level contrastive relationships, and distribution-level knowledge propagation, enabling effective novel class discovery while maintaining robust performance on known categories despite severe data imbalance.

Joint Segmentation Objectives. The foundation of our approach is a unified objective function that integrates supervised learning for known classes with unsupervised discovery for novel classes:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{primary}} + \mathcal{L}_{\text{auxiliary}}, \tag{4.35}$$

$\mathcal{L}_{\text{primary}}$ handles both known and novel class points. For known class points, it employs standard supervised learning with cross-entropy loss based on ground truth labels y_i^K. For novel class points, it utilizes the pseudo-labels $\tilde{y}_i^{N,(h)}$ generated by the Sinkhorn-Knopp algorithm to generate soft cluster assignments based on feature-prototype similarities. This approach enables unsupervised discovery of novel semantic categories by enforcing consistent cluster assignments across different augmented views of the same point cloud.

$\mathcal{L}_{\text{auxiliary}}$ provides additional regularization and also processes both known and novel class points, but applies overclustering to novel classes, partitioning them into finer-grained subcategories. This encourages the discovery of more nuanced semantic boundaries and improves the primary clustering objective. Together, these two losses constitute our base segmentation loss.

Knowledge Transfer Objectives. To bridge the semantic gap between known and novel classes, we introduce three specialized knowledge transfer objectives:

Prototype Relation Loss: This objective establishes semantic connections between the prototype representations of known and novel classes:

$$\mathcal{L}_{\text{PR}} = \alpha(t) \cdot \mathcal{S}(\mathbf{\Pi}_K, \mathbf{\Pi}_N) - (1 - \alpha(t)) \cdot (\mathcal{C} + \mathcal{D}), \tag{4.36}$$

where $\mathbf{\Pi}_K \in \mathbb{R}^{K \times D_K}$ and $\mathbf{\Pi}_N = \{\mathbf{\Pi}_N^{(h)}\}_{h=1}^{H}$ represent the prototype matrices for known and novel classes respectively, $\mathcal{S}$ measures their semantic alignment, while $\mathcal{C}$ and $\mathcal{D}$ quantify intra-class cohesion and inter-class diversity. The dynamic weight $\alpha(t) = \sigma(t/\tau_\alpha)$ progressively shifts emphasis from diversity to semantic alignment throughout training, where σ denotes the sigmoid function, t represents the training step, and τ_α is a scaling parameter.

Contrastive Alignment Loss: This objective facilitates instance-level semantic relationships through contrastive learning:

$$\mathcal{L}_{\text{CA}} = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \log \frac{\exp(\langle \mathbf{z}_i, \mathbf{z}_i^+ \rangle / \tau_{CA})}{\exp(\langle \mathbf{z}_i, \mathbf{z}_i^+ \rangle / \tau_{CA}) + \exp(\langle \mathbf{z}_i, \mathbf{z}_i^- \rangle / \tau_{CA})}, \tag{4.37}$$

where $\mathcal{U}$ represents the set of points from novel classes, $\mathbf{z}_i$ denotes the normalized feature representation for novel class point i, $\mathbf{z}_i^+$ is the feature of the most similar point from known classes, $\mathbf{z}_i^-$ represents a more distant known-class feature, and τ_{CA} is the temperature parameter. This mechanism creates semantic bridges between known and novel categories while preserving their distinctive characteristics.

Semantic Transfer Loss: This objective propagates class-level knowledge through distribution alignment:

$$\mathcal{L}_{\text{ST}} = D_{\text{KL}} \left(\mathbf{P}_N(\mathbf{z}) \parallel \sum_{k \in \mathcal{K}} w_k \cdot \mathbf{P}_k(\mathbf{z}) \right), \tag{4.38}$$

where $\mathbf{P}_N(\mathbf{z})$ is the predicted distribution over novel classes for feature $\mathbf{z}$, $\mathbf{P}_k(\mathbf{z})$ is the distribution for known class k, $\mathcal{K} = \{1, ..., K\}$ represents the set of known classes, weights w_k are determined by feature similarities, and D_{KL} denotes the Kullback-Leibler divergence. This alignment guides novel class formation based on structured knowledge from known classes.

The complete learning objective integrates these components with progressively increasing weights:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{PR}}(t) \cdot \mathcal{L}_{\text{PR}} + \lambda_{\text{CA}}(t) \cdot \mathcal{L}_{\text{CA}} + \lambda_{\text{ST}}(t) \cdot \mathcal{L}_{\text{ST}}, \tag{4.39}$$

where the coefficients $\lambda_{\text{PR}}(t)$, $\lambda_{\text{CA}}(t)$, and $\lambda_{\text{ST}}(t)$ increase progressively during training to gradually introduce specialized constraints for novel class discovery.

Class Imbalance Mitigation through Knowledge Transfer. The three proposed knowledge transfer objectives contribute to mitigating class imbalance challenges in novel class discovery tasks. The Prototype Relation Loss establishes semantic connections between known and novel classes, allowing novel classes without labeled training data to benefit from the supervision signals available for known classes. Through the semantic alignment measurement $\mathcal{S}(\mathbf{\Pi}_K, \mathbf{\Pi}_N)$, the framework enables novel classes to utilize knowledge encoded in well-represented known class prototypes, thereby compensating for the absence of direct supervision.

The Contrastive Alignment Loss facilitates balanced representation learning by providing each novel class point with selected positive and negative examples from the known class space. This mechanism helps ensure that novel classes receive more consistent learning signals even when they have limited representation in the unlabeled data. By establishing feature boundaries through instance-level contrastive learning, the approach helps prevent well-represented known classes from dominating the learning process for less frequent novel classes.

The Semantic Transfer Loss addresses distributional imbalance by aligning novel class distributions with weighted combinations of known class distributions. The adaptive weighting scheme w_k, determined by feature similarities, directs novel classes to receive knowledge from the most relevant known classes. This approach helps maintain learning stability when novel class instances constitute a small proportion of the unlabeled point cloud data. The knowledge propagation mechanism enables effective learning for novel classes with limited data representation while preserving inter-class discriminability.

The progressive weight scheduling $\lambda(t)$ for all three objectives provides gradual integration of these constraints. This approach helps prevent early training instability that may arise from class imbalance while facilitating effective knowledge transfer in later training stages. These coordinated mechanisms collectively enable robust novel class discovery even under data imbalance conditions.

4.3 Experiments

4.3.1 Datasets

We evaluate the proposed framework on three diverse urban point cloud datasets to demonstrate its effectiveness across different environments and conditions.

Table 4.1: Detailed network architecture of the proposed framework. All dimensions refer to feature channels. "–" denotes operations without feature dimensionality.

Module	Layer	Input Dim	Output Dim	Kernel / Operation	#Params
MinkowskiNet	Initial Conv	1	32	Conv3d(1→32, k=5×5×5)	4,000
	Encoder Stage 1 (2 blocks)	32	64	Conv3d(32→64, k=3×3×3)	37,056
	Encoder Stage 2 (3 blocks)	64	128	Conv3d(64→128, k=3×3×3)	222,848
	Encoder Stage 3 (4 blocks)	128	256	Conv3d(128→256, k=3×3×3)	1,188,096
	Encoder Stage 4 (6 blocks)	256	256	Conv3d(256→256, k=3×3×3)	3,559,680
	Decoder Stage 1 (2 blocks)	512	256	Conv3d(512→256, k=3×3×3)	2,365,952
	Decoder Stage 2 (2 blocks)	384	128	Conv3d(384→128, k=3×3×3)	885,120
	Decoder Stage 3 (2 blocks)	192	96	Conv3d(192→96, k=3×3×3)	332,160
	Decoder Stage 4 (2 blocks)	128	96	Conv3d(128→96, k=3×3×3)	222,528
AGSM	Adaptive Sampling Weights	–	–	Learnable parameters	3
	Weighted FPS	–	–	Farthest Point Sampling	0
	Density-aware KNN Grouping	–	–	Dynamic grouping $k \in [16, 32]$	0
	Point Encoder	3	64	Conv1d(3→64)+BN+ReLU	384
	Geo Encoder	3	64	Conv1d(3→64)+BN+LeakyReLU	384
	Edge Encoder	6	64	Conv1d(6→128)+BN+ReLU; Conv1d(128→64)	9,408
	Edge Compressor	128	64	Conv1d(128→64)+BN+ReLU	16,576
	Fusion	256	384	Conv1d(256→384)+BN+ReLU	99,456
	GCIM (3 layers)	384	384	3×[Linear(384→384)+Gate]	887,040
	FRM MLP Layer 1	384	256	Linear(384→256)+LayerNorm+ReLU	99,072
	FRM MLP Layer 2	256	96	Linear(256→96)+LayerNorm+ReLU	24,864
	Output Projection	96	96	Linear(96→96)	9,312
DGE	Gaussian Parameters	–	–	1000×(3+3+4+96)	106,000
	Mahalanobis + Softmax	–	96	Weighted Gaussian aggregation	0
GMSFF	Gate Network	288	96	Linear(288→96)+Sigmoid	27,744
	Weight Network	288	3	Linear(288→3)+Softmax	867
	Transform 1	288	96	Linear(288→96)+ReLU	27,744
	Transform 2	96	96	Linear(96→96)+ReLU	9,312
Segmentation	Base Prototypes Head	96	K	Conv1d(96→K)	96×K
	Novel Prototypes Head	96	N_{novel}	Conv1d(96→N_{novel})	96×N_{novel}

The Toronto-3D dataset [78] is an urban point cloud dataset acquired by a Teledyne Optech Maverick mobile LiDAR system in Toronto, Canada. Covering approximately 1 km of roadway, it comprises 78.3 million points across 8 semantic classes. The dataset provides detailed point attributes, presenting challenges such as variations in point cloud density and the inclusion of distant objects. This dataset enables evaluation of the model's ability to handle variations in urban scene complexity and to accurately segment large-scale urban areas, contributing to effective urban planning and infrastructure development.

The SemanticSTF dataset [94] is a pioneering benchmark for 3D semantic segmentation under diverse weather conditions, comprising 2,076 LiDAR scans across snow, rain, and fog environments. With 21 annotated semantic categories, 19 of which are used for evaluation, the dataset presents unique challenges related to weather-specific distortions. The SemanticSTF dataset facilitates evaluating the robustness of segmentation models in adverse weather scenarios, which is critical for environmental monitoring and maintaining infrastructure integrity in regions affected by extreme weather.

The SemanticPOSS dataset [66] was collected in the intricate campus environment of Peking University, comprising 2,988 LiDAR scans with over 216 million points. It provides detailed point-wise labels and instance-level annotations for both dynamic and static objects, including an average of 8.29 people instances and 15.02 cars per frame, offering a significantly richer dynamic object representation than existing datasets. SemanticPOSS is particularly suited for urban scene analysis, enabling the evaluation of models aimed at urban infrastructure monitoring, pedestrian flow analysis, and public space management.

4.3.2 Implementation Details

Table 4.1 summarizes the architectural configuration of all components in the proposed framework, including input/output dimensions, kernel sizes, and parameter counts for each layer. The MinkowskiNet consists of four encoder stages and four decoder stages. The encoder progressively increases the feature dimension from 32 to 256 using $3 \times 3 \times 3$ sparse convolutions, while the decoder reduces it back to 96. The AGSM module includes several 1D convolutional blocks for point-wise, geometric, and edge encodings. The fusion stage integrates 256 input channels and outputs 384 channels using a Conv1d operation. Three gated context integration layers preserve the 384-dimensional representation, followed by two fully connected layers that reduce the dimensionality to 96. The DGE module employs 1000 learnable Gaussians, each defined by a 3D position, scale, rotation (4D quaternion), and a 96-dimensional feature vector. The Gaussian parameters are directly optimized during training, contributing a total of 106,000 parameters. The GMSFF module receives three 96-dimensional feature sources, concatenated into a 288-dimensional tensor. It applies a gating network (288→96), a weight network (288→3), and two linear transformation layers to produce a 96-dimensional fused representation. The segmentation head contains two parallel Conv1d layers: one for K known classes and another for N_{novel} novel classes per prototype head, both mapping from 96-dimensional features.

4.3.3 Training Settings

For fair comparison, all baseline methods in our experiments were implemented and evaluated using the same data preprocessing steps, data splitting steps, training parameters, and evaluation protocols as in [71, 18]. The proposed model was trained using the following configuration across all datasets: each point cloud file used only (x, y, z) coordinates, with a voxel size of 0.05 and downsampling to 60,000 points for training. We used SGD optimizer with learning rate 1×10^{-2} for newly initialized modules, momentum 0.9, and weight decay 1×10^{-4}. The learning rate schedule applied linear warm-up for 10 epochs

Table 4.2: Novel class discovery results on 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.

		Per-class IoU								mIoU		
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
	CHNCD	**87.10**	10.39	94.32	55.80	78.21	70.84	**67.44**	29.92	**71.45**	58.52	61.75
	Ours	86.55	**12.57**	**94.85**	**55.99**	**80.70**	**80.79**	64.85	**34.80**	71.27	**61.43**	**63.89**
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
	CHNCD	91.49	14.82	**55.49**	61.75	74.51	76.50	6.82	29.47	31.16	58.09	51.36
	Ours	**93.57**	**17.13**	50.00	**64.79**	**84.57**	**76.87**	**20.16**	**34.86**	**35.08**	**61.97**	**55.24**
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
	CHNCD	91.78	10.53	94.31	61.34	70.61	55.17	70.65	39.02	32.85	71.29	61.68
	Ours	**94.31**	**14.17**	**95.40**	**64.38**	**85.01**	**72.69**	**77.59**	**44.03**	**43.43**	**76.79**	**68.45**
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
	CHNCD	91.09	14.11	94.44	62.46	49.53	75.82	68.66	**14.52**	**32.03**	67.76	58.83
	Ours	**93.66**	**16.29**	**95.14**	**63.03**	**49.76**	**81.53**	**79.49**	13.49	31.63	**71.52**	**61.55**

followed by cosine annealing to 1×10^{-5}. Training was conducted with batch size 4 and 8 worker threads for 500 epochs on Toronto-3D and 100 epochs on SemanticSTF. DropPath regularization (initial ratio 0.1) was applied with progressive increase by network depth. The multi-objective framework combined Prototype Relation Loss (maximum weight 0.3), Contrastive Alignment Loss (0.2), and Semantic Transfer Loss (0.1) with linearly increasing weights during training.

4.3.4 Evaluation and Visualization Results on Toronto-3D

Table 4.2 presents results on the Toronto-3D dataset, acquired by a mobile LiDAR system in urban Toronto environments. The proposed method demonstrates consistent performance gains across all metrics, with average improvements in overall mIoU of 6.59% over NOPS [71] and 3.88% over CHNCD [18]. Novel class segmentation increases by 6.47% compared to NOPS and 3.48% over CHNCD, while base class performance advances by 6.64% and 4.01% relative to NOPS and CHNCD respectively.

The capacity to segment thin vertical structures emerges as a distinctive strength throughout multiple splits. Pole segmentation in the Toronto-3D-2^0 split reaches 80.79%, representing a 16.04% improvement over NOPS (64.75%) and 9.95% over CHNCD (70.84%). This remarkable enhancement for slender objects with sparse point representation illustrates how adaptive sampling strategies preserve critical geometric details in regions where traditional approaches often fail. Similar patterns appear in utility line segmentation,

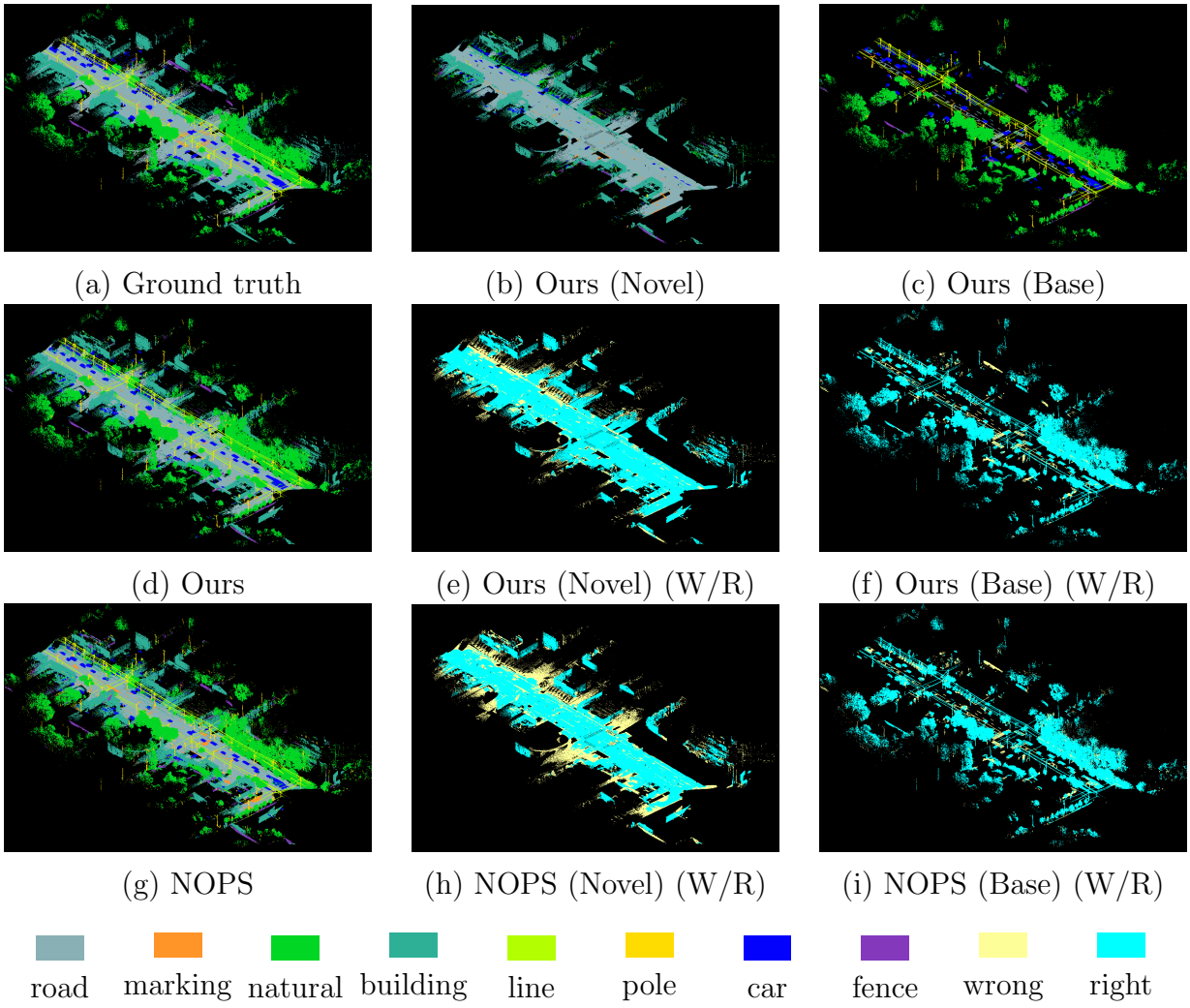

Figure 4.5: Segmentation results on Toronto-3D-2^0 split. (a) Ground truth, (b) our novel class predictions (road, building), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

which reaches 80.70% compared to 71.70% for NOPS, contributing to a base class mIoU of 61.43% that substantially outperforms NOPS (54.00%) and CHNCD (58.52%).

The Toronto-3D-2^1 split demonstrates how semantic relationships influence cross-category enhancement. Despite the natural class showing moderate improvement at 50.00% (compared to NOPS at 46.96% but below CHNCD at 55.49%), the car category demonstrates

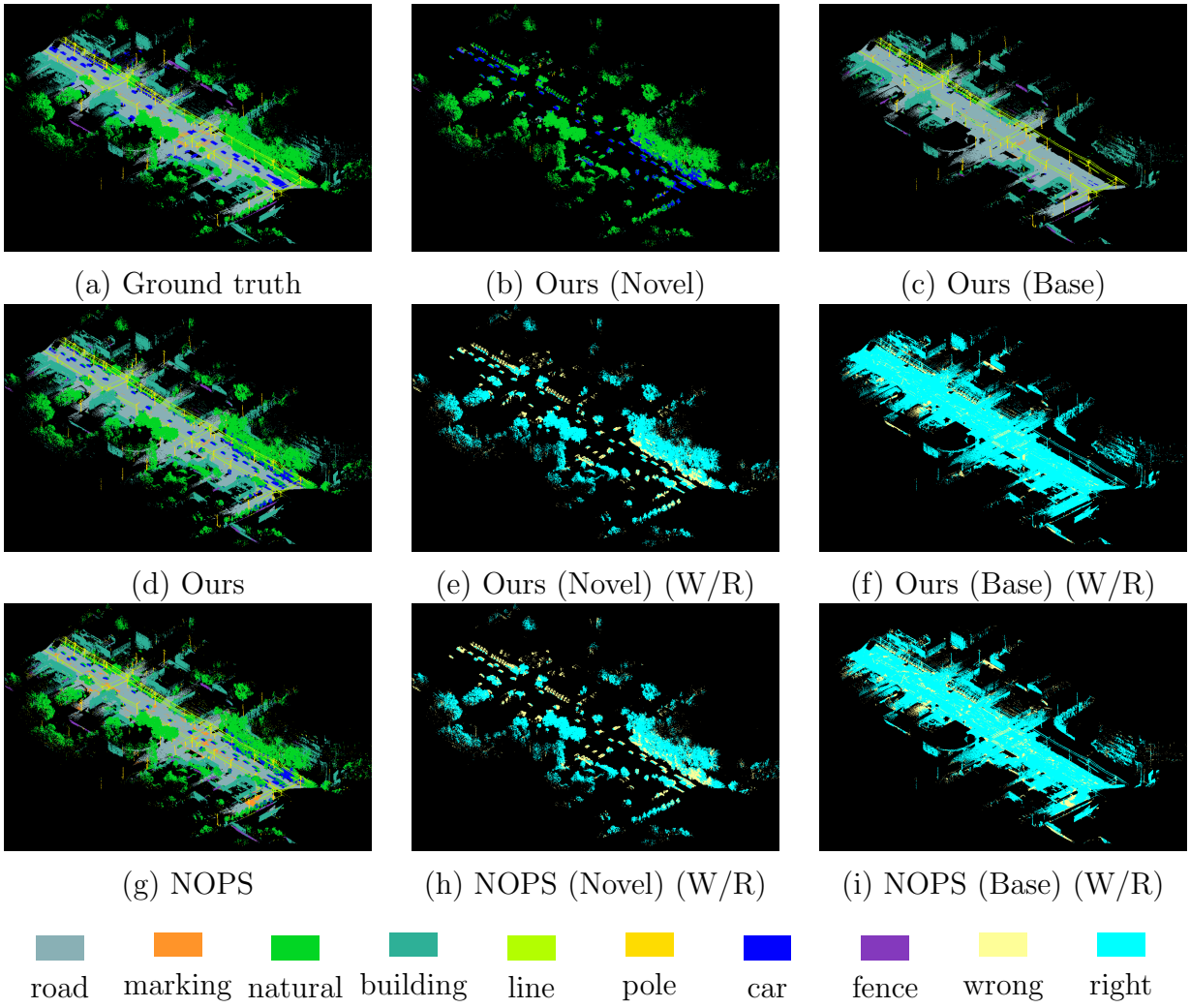

Figure 4.6: Segmentation results on Toronto-3D-2[1] split. (a) Ground truth, (b) our novel class predictions (natural, car), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

significant advancement, reaching 20.16% versus 17.17% for NOPS. This significant improvement for a geometrically complex object category illustrates how probabilistic representation effectively captures structural variations in vehicles with diverse orientations. Base class performance benefits similarly from these knowledge transfer mechanisms, with utility line segmentation increasing from 74.67% (NOPS) to 84.57%. These complemen-

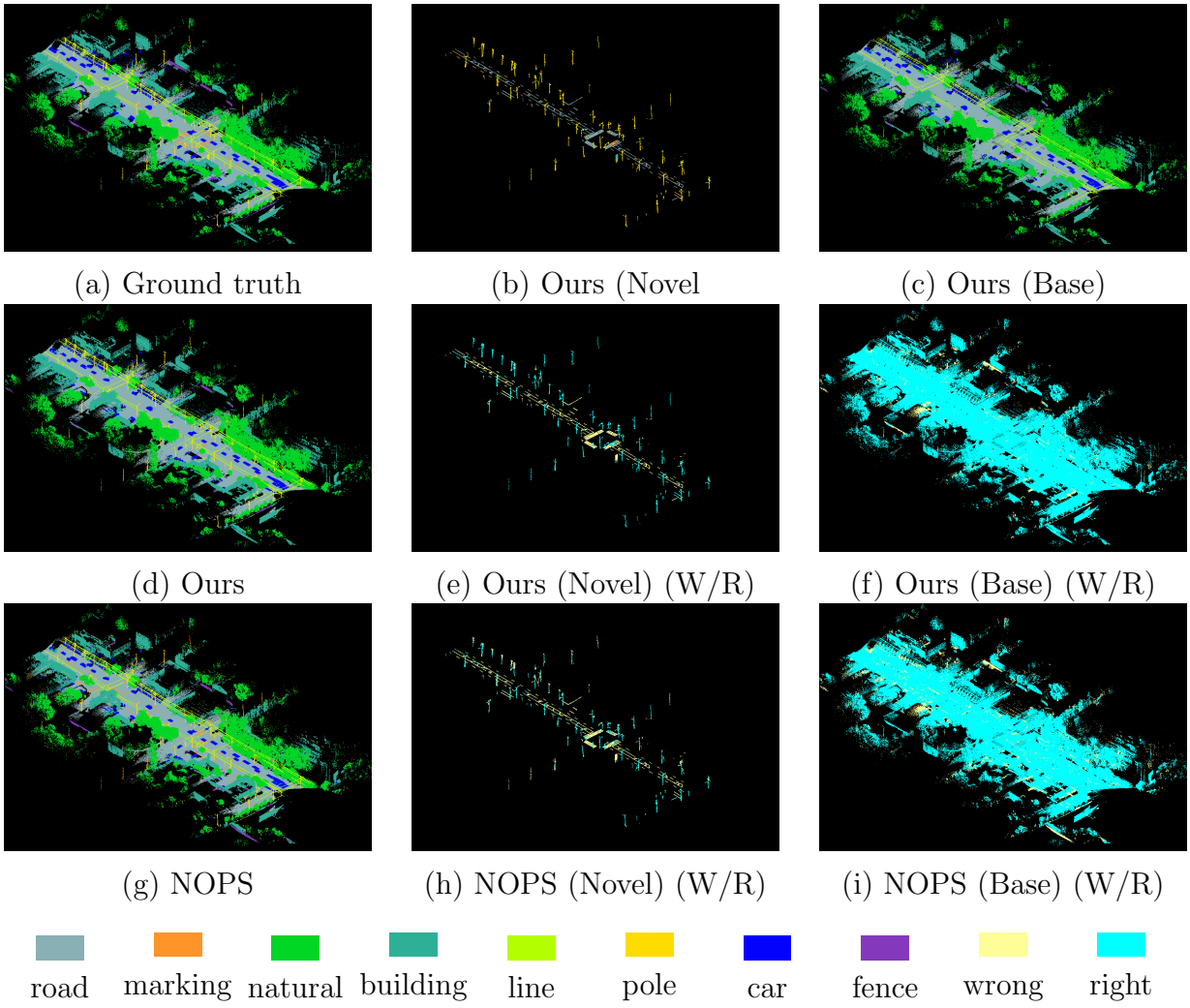

Figure 4.7: Segmentation results on Toronto-3D-2^2 split. (a) Ground truth, (b) our novel class predictions (road marking, pole), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

tary enhancements yield a novel class mIoU of 35.08% and base class mIoU of 61.97%, producing an overall mIoU of 55.24% that surpasses previous approaches by 5.85% and 3.88% respectively.

Knowledge transfer proves particularly effective when applied to structurally challenging categories in the Toronto-3D-2^2 split. Road marking segmentation reaches 14.17%,

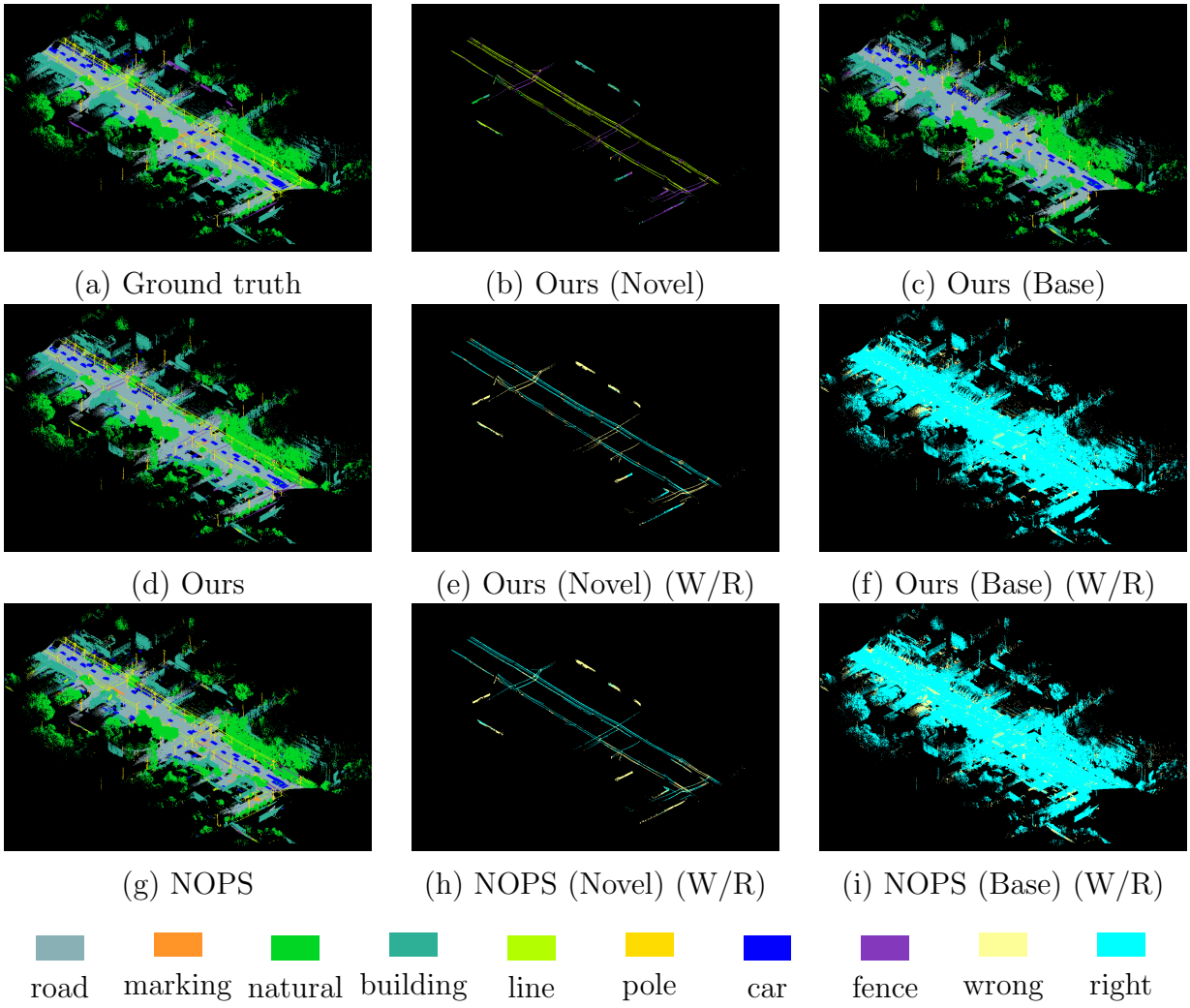

Figure 4.8: Segmentation results on Toronto-3D-2^3 split. (a) Ground truth, (b) our novel class predictions (utility line, fence), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

more than tripling NOPS (4.16%) and outperforming CHNCD (10.53%) by 3.64%. Concurrently, pole segmentation as a novel class achieves 72.69%, substantially surpassing both NOPS (56.50%) and CHNCD (55.17%). These combined improvements yield a novel class mIoU of 43.43%, representing significant gains of 13.10% over NOPS and 10.58% over CHNCD. The ability to effectively segment such challenging categories demonstrates how

prototype-based learning establishes robust semantic bridges between known and novel classes that share geometric characteristics. Base class performance follows this pattern with an impressive mIoU of 76.79%, contributing to the highest overall performance across all splits at 68.45%.

Network structures and boundary delineation challenges emerge in the Toronto-3D-2^3 split, where utility line and fence appear as novel classes. The modest improvement in utility line segmentation (49.76% versus 48.61% for NOPS) illustrates the inherent difficulty of modeling complex network-like structures with limited point density. Fence segmentation shows similar patterns at 13.49%, falling between NOPS (8.08%) and CHNCD (14.52%). Despite these limitations for certain novel classes, base class performance demonstrates robust improvement at 71.52%, significantly outperforming both NOPS (66.71%) and CHNCD (67.76%). The most substantial gains appear in car segmentation (79.49% versus 65.92% for NOPS) and pole segmentation (81.53% versus 74.16%), illustrating how adaptive feature fusion enhances performance across both dynamic and static urban elements. These balanced improvements produce an overall mIoU of 61.55%, advancing 4.43% beyond NOPS.

Figure 4.5 demonstrates the framework's novel class discovery performance on the Toronto-3D-2^0 split, where road and building serve as target novel categories. The novel class segmentation in Figure 4.5(b) shows accurate road surface segmentation with continuous lane boundaries and precise geometric transitions at intersection areas, while building structures exhibit consistent vertical wall segmentation and clear architectural boundary definition. The complete prediction result in Figure 4.5(d) maintains spatial consistency between novel and base class predictions, preserving geometric relationships across different urban infrastructure elements. The base class visualization in Figure 4.5(c) reveals enhanced performance for vertical structures including poles and utility lines, with improved connectivity and reduced geometric discontinuities.

The error analysis in Figures 4.5(e) and 4.5(f) indicates concentrated accuracy improvements at urban infrastructure boundaries and geometric transition zones. Comparative evaluation against NOPS in Figures 4.5(g), 4.5(h), and 4.5(i) shows reduced error patterns in complex geometric regions, particularly at road-building interfaces and utility network connections. The framework demonstrates consistent performance across varying structural complexity, from large-scale planar surfaces to thin linear elements, while maintaining accurate boundary delineation and topological preservation throughout the urban scene.

Figure 4.6 reveals the framework's capability to identify natural and car as novel classes in the Toronto-3D-2^1 split. The natural discovery in Figure 4.6(b) shows segmentation of scattered tree areas and vegetation patches distributed throughout the scene, with clear

segmentation of green spaces amid the urban infrastructure. Car segmentation captures individual cars positioned along roadways and in designated areas, maintaining geometric boundary definition despite the sparse point cloud representation of these objects. The integrated visualization in Figure 4.6(d) demonstrates consistent semantic classification across the entire scene, with proper distinction between the discovered novel categories and existing infrastructure elements.

Wrong/Right analysis in Figures 4.6(e) and 4.6(f) indicates improved accuracy in novel class discovery compared to baseline performance. The base class visualization in Figure 4.6(c) shows maintained performance across road surfaces, building structures, and linear infrastructure elements including utility lines and poles. NOPS results in Figures 4.6(g), 4.6(h), and 4.6(i) show increased error patterns in vegetation boundary definition and reduced accuracy in vehicle segmentation, with more fragmented segmentation in both novel categories. The framework achieves superior performance in maintaining spatial coherence for vegetation areas and preserving vehicle geometric integrity compared to the baseline approach.

Figure 4.7 presents novel class discovery results for the Toronto-3D-2^2 split, where road marking and pole constitute the target novel categories. The road marking segmentation in Figure 4.7(b) shows segmentation of linear pavement markings with improved connectivity compared to baseline methods, though the thin geometric characteristics present inherent challenges due to sparse point representation. Pole discovery exhibits accurate segmentation of individual vertical structures distributed throughout the scene, with preserved cylindrical boundaries and clear geometric separation from adjacent infrastructure elements. The base class performance in Figure 4.7(c) demonstrates maintained segmentation quality across road surfaces, building structures, and vegetation areas, ensuring that novel class discovery does not compromise established category recognition.

The Wrong/Right evaluation in Figures 4.7(e) and 4.7(f) shows concentrated cyan regions indicating accurate predictions, with notable improvements in pole segmentation and enhanced road marking segmentation. NOPS baseline performance in Figure 4.7(g) exhibits more fragmented results, while the error analysis in Figures 4.7(h) and 4.7(i) reveals increased yellow error patterns compared to our method. The framework demonstrates better preservation of linear structure continuity for road markings and maintains more consistent pole boundary definition throughout the scene.

The Toronto-3D-2^3 split results in Figure 4.8 target utility line and fence as novel categories, representing some of the most challenging structures for discovery due to their sparse point representation and complex geometric arrangements. Utility line segmentation in Figure 4.8(b) shows segmentation of linear infrastructure connections with performance similar to baseline methods, reflecting the inherent difficulty of these thin network

Table 4.3: Novel class discovery results on 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.

		Per-class IoU																			mIoU		
Split	Model	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	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
	CHNCD	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
	Ours	**67.66**	**31.28**	14.40	37.09	32.35	**53.20**	**70.94**	**28.85**	56.34	**25.97**	9.46	**25.33**	**37.37**	**46.55**	**44.15**	**35.20**	**29.90**	**44.08**	**45.19**	**35.44**	**39.86**	**38.70**
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
	CHNCD	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
	Ours	**60.19**	**33.87**	**14.95**	**50.04**	**34.35**	**57.31**	**73.30**	**22.14**	**80.57**	5.47	**50.44**	**11.48**	**75.69**	22.53	63.18	**18.94**	**60.12**	**47.28**	**48.13**	**23.72**	**50.81**	**43.68**
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
	CHNCD	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
	Ours	**77.14**	31.82	15.81	**22.93**	0.27	10.87	**76.94**	**25.22**	**82.51**	22.98	**54.28**	**35.05**	**76.76**	**51.74**	65.25	**37.21**	**61.96**	20.17	12.47	13.34	**51.05**	**41.13**
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
	CHNCD	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
	Ours	**77.61**	8.79	**13.18**	**43.54**	**40.31**	**60.37**	12.50	**15.75**	**82.12**	24.53	**55.08**	**32.88**	**76.48**	**51.46**	64.98	**39.09**	**60.96**	**49.02**	**48.13**	12.56	**53.77**	**45.09**

structures. Fence segmentation demonstrates modest improvements in boundary element capture, though both methods face substantial challenges with these linear boundary structures.

Base class performance in Figure 4.8(c) reveals substantial improvements across multiple categories, with car segmentation achieving enhanced geometric boundary preservation and pole segmentation showing significant advances in vertical structure segmentation. Road surfaces maintain excellent spatial coverage with improved connectivity, while road marking segmentation demonstrates better linear feature preservation. The Wrong/Right analysis in Figures 4.8(e) and 4.8(f) shows concentrated cyan regions for base classes, indicating strong performance maintenance. NOPS results in Figures 4.8(g), 4.8(h), and 4.8(i) exhibit more fragmented patterns across base categories, with increased yellow error distributions particularly visible in car and pole segmentation, demonstrating the proposed framework's superior performance in preserving known category accuracy while attempting novel class discovery.

4.3.5 Evaluation and Visualization Results on SemanticSTF

Table 4.3 presents comparative results on the SemanticSTF dataset, which evaluates segmentation under challenging weather conditions including snow, rain, and fog. The proposed method delivers consistent performance gains with an average improvement in overall mIoU of 7.32% over NOPS and 4.80% over CHNCD. For novel classes, the average improvement reaches 4.61% over NOPS and 3.12% over CHNCD, while for base classes, it achieves 8.21% over NOPS and 5.30% over CHNCD.

The ability to segment vegetation shows the most significant improvement in the SemanticSTF-5^0 split, achieving 44.15% IoU compared to 21.80% for NOPS and 23.73% for CHNCD. This substantial 22.35% gain for vegetation, a novel class in this split, illustrates how the AGSM module effectively handles complex natural structures even when precipitation affects point cloud density. Building segmentation also benefits significantly (37.37% versus 23.57% for NOPS), contributing to a novel class mIoU of 35.44% which surpasses NOPS by 9.27% and CHNCD by 5.83%. Dynamic objects in the base classes show equally impressive gains, with bicyclist reaching 70.94% (versus 57.13% for NOPS) and person achieving 53.20% (versus 45.33% for NOPS). These balanced improvements across both static and dynamic elements yield an overall mIoU of 38.70%, representing gains of 9.19% over NOPS and 6.12% over CHNCD.

Vehicle segmentation under adverse weather represents a particularly challenging task that the proposed framework handles remarkably well. In the SemanticSTF-5^1 split, car segmentation (a novel class) reaches 60.19%, doubling the performance of NOPS (30.41%) and CHNCD (29.89%). This notable improvement stems from the knowledge transfer capabilities of the framework, which establish semantic connections between geometrically similar classes despite weather-induced distortions. The challenging trunk category also sees substantial gains (18.94% versus 4.99% for NOPS), highlighting the method's effectiveness on thin structures that typically suffer from point sparsity during precipitation. Base class performance shows similar enhancements, with truck segmentation improving by 18.30% over NOPS. These combined improvements yield an overall mIoU of 43.68%, an 8.36% gain over NOPS and 5.69% over CHNCD.

Weather conditions particularly affect the segmentation of highly variable object categories, as revealed in the Seman-ticSTF-5^2 split. While truck segmentation shows a respectable improvement (22.93% versus 15.87% for NOPS), limitations appear with the challenging other-vehicle class (0.27% versus 0.48% for NOPS). This pattern suggests that the knowledge transfer objectives work most effectively when clear geometric correspondences exist between known and novel categories. The effectiveness of the approach becomes more evident in base class performance, achieving a remarkable 51.05% mIoU, outperforming NOPS by 7.79% and CHNCD by 4.12%. The bicyclist category shows an impressive 35.02% absolute improvement over NOPS, demonstrating the DGE module's capability to maintain distinctive shape representation even under degraded visibility conditions. The overall mIoU reaches 41.13%, representing gains of 5.57% over NOPS and 3.08% over CHNCD.

The SemanticSTF-4^3 split presents a case where NOPS completely fails to segment a particular category. For motorcycle, one of the novel classes in this split, the proposed method achieves 13.18% IoU compared to 0.00% for NOPS and 5.59% for CHNCD. This

ability to segment narrow two-wheeled vehicles under challenging weather conditions highlights the strength of the prototype-based approach in capturing essential geometric characteristics despite limited point density. The balanced performance extends to base classes as well, with significant improvements across structurally diverse elements including person (60.37% versus 48.46% for NOPS), pole (49.02% versus 40.86% for NOPS), and traffic-sign (48.13% versus 39.05% for NOPS). This comprehensive enhancement yields a base class mIoU of 53.77%, representing gains of 7.07% over NOPS and 5.59% over CHNCD, and an overall mIoU of 45.09%.

The SemanticSTF-5^0 split visualization in Figure 4.9 targets road, sidewalk, vegetation, building, and terrain as novel categories under adverse weather conditions, where the majority of the scene consists of novel class elements. Novel class segmentation in Figure 4.9(b) demonstrates successful segmentation across multiple visible categories, with particularly strong performance in road surface coverage and vegetation boundary preservation despite weather-induced point cloud degradation. Terrain areas show consistent spatial organization. Base class segmentation in Figure 4.9(c) shows trunk structures maintain accurate segmentation throughout the scene and fence elements exhibit clear linear boundary definition across the roadway environment. The extensive spatial coverage of discovered novel classes validates the proposed method's capability to handle scenarios where most scene elements lack supervised training data. Error evaluation through Wrong/Right encoding in Figures 4.9(e) and 4.9(f) reveals concentrated cyan regions across vegetation and road areas, confirming superior novel class discovery. NOPS baseline results in Figures 4.9(g), 4.9(h), and 4.9(i) exhibit increased fragmentation with more scattered yellow error patterns throughout the scene, demonstrating our approach's enhanced performance in challenging weather conditions where novel categories dominate the environment.

Figure 4.10 presents qualitative results on the SemanticSTF-5^1 split, where the novel categories include car, parking, ground, fence, and trunk. In Figure 4.10(b), the proposed method correctly segments most car instances aligned along the main road, with clear object boundaries and consistent shapes. The trunk category is also effectively captured, showing continuous vertical structures with minimal fragmentation. Figure 4.10(e) confirms these observations through concentrated cyan regions in both car and trunk areas, suggesting accurate segmentation. The composite prediction in Figure 4.10(d) produces a consistent reconstruction of the urban layout, aligning closely with the spatial structure observed in the ground truth from Figure 4.10(a). For base classes, Figure 4.10(c) demonstrates reliable segmentation of building and sidewalk classes, with well-preserved linear edges and coherent area coverage, especially near the lower and right-hand side of the scene. As shown in Figure 4.10(f), these base categories correspond to large regions of correct predictions.

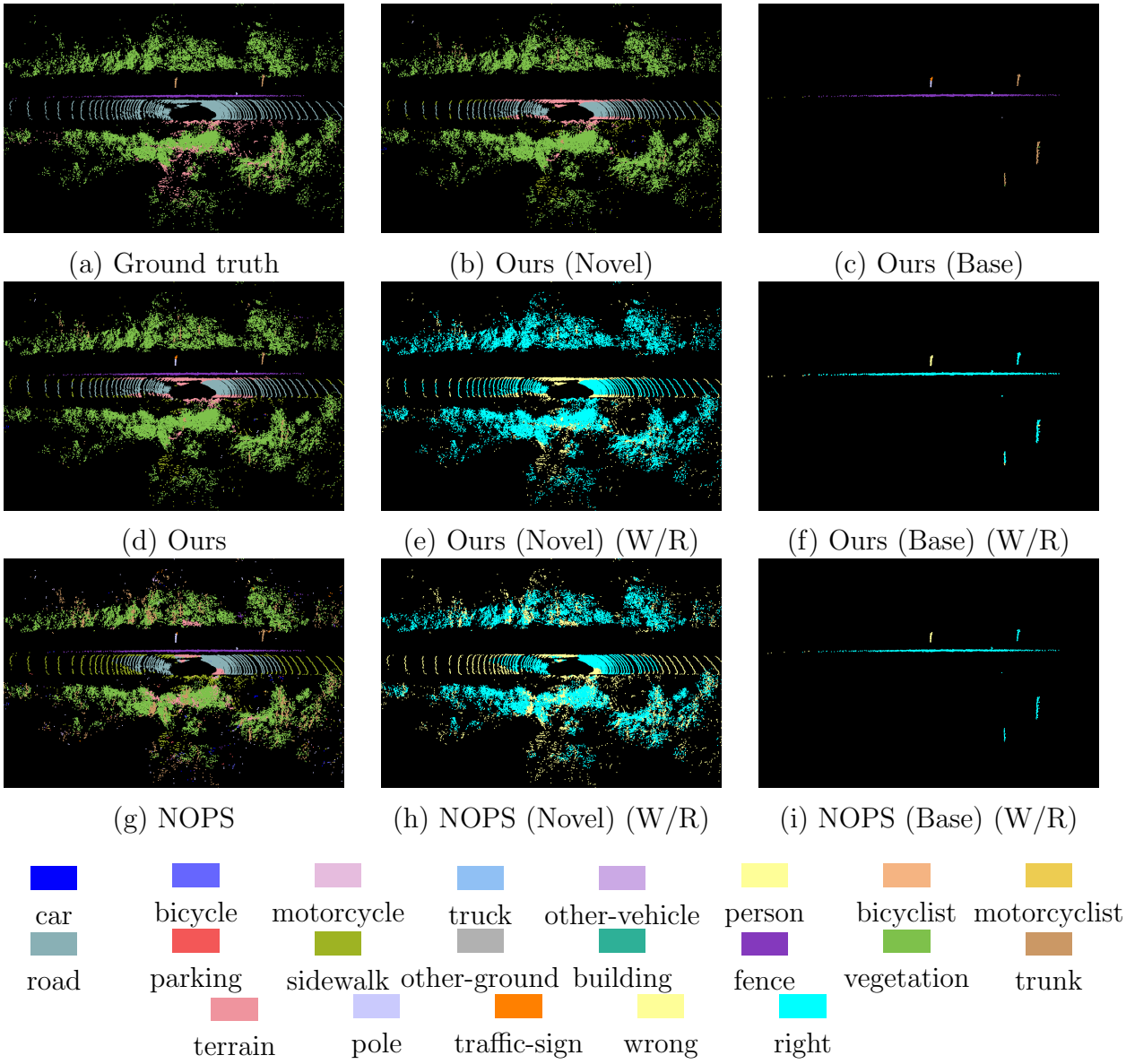

Figure 4.9: Segmentation results on SemanticSTF-5^0 split. (a) Ground truth, (b) our novel class predictions (road, sidewalk, building, vegetation, and terrain), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

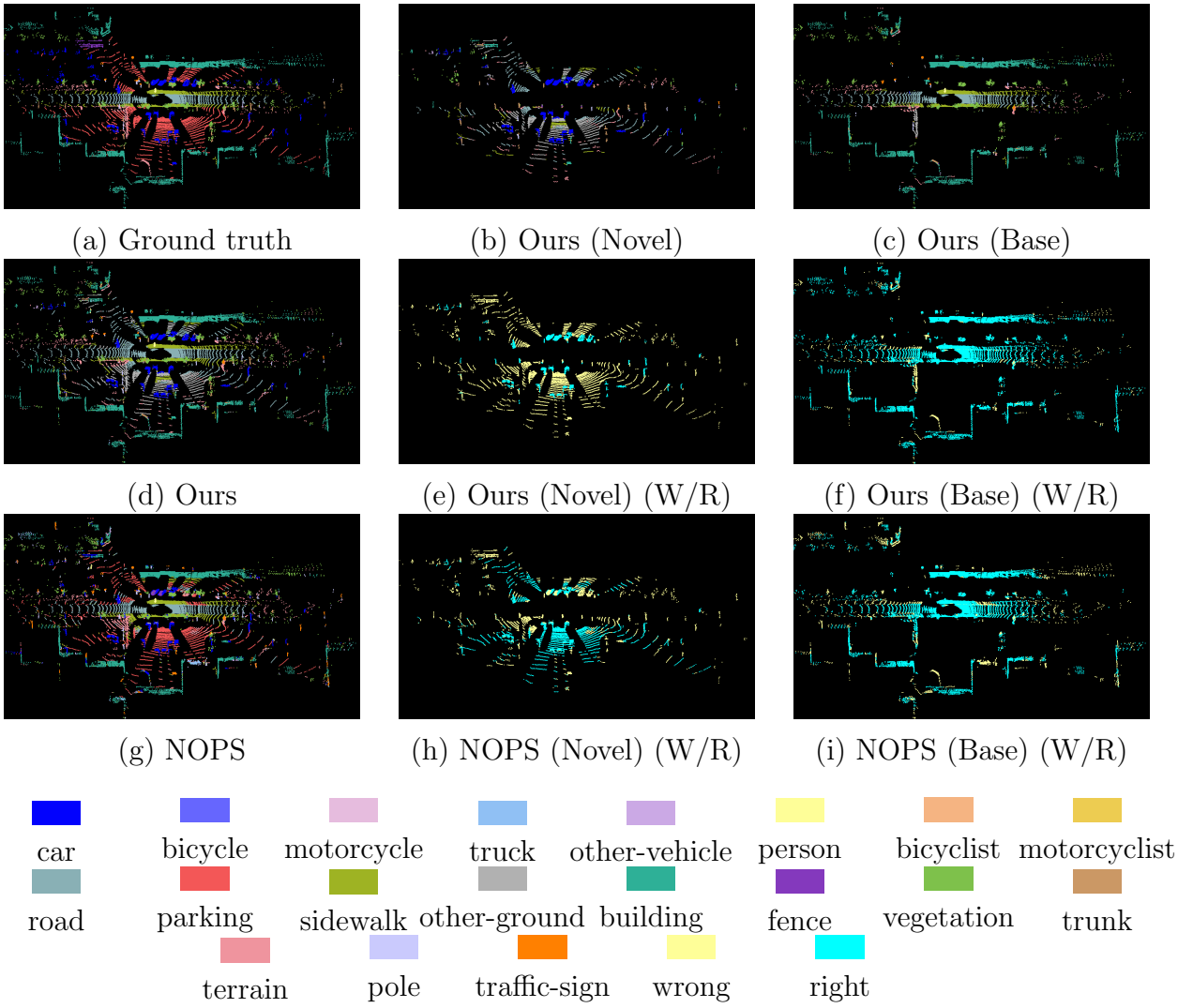

Figure 4.10: Segmentation results on SemanticSTF-5[1] split. (a) Ground truth, (b) our novel class predictions (car, parking, ground, fence, and trunk), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

Table 4.4: Novel class discovery results on 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.

		Per-class IoU													mIoU		
Split	**Model**	**bike**	**building**	**car**	**cone**	**fence**	**ground**	**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
	CHNCD	**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
	ours	37.86	45.58	**7.19**	**37.32**	**37.90**	**80.08**	**58.15**	56.40	**30.90**	**62.24**	24.75	3.08	19.67	**47.31**	**34.65**	**38.55**
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
	CHNCD	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
	ours	20.93	**77.69**	**41.19**	**35.68**	**28.57**	77.92	**57.76**	74.62	**33.24**	**57.63**	**25.86**	3.95	16.42	**35.75**	**44.42**	**42.42**
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
	CHNCD	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
	ours	**45.73**	**80.53**	**46.56**	**31.66**	**45.48**	**78.02**	**63.50**	**76.62**	**33.65**	**55.22**	1.34	0.75	7.46	**14.15**	**52.41**	**43.58**
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
	CHNCD	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
	ours	**44.16**	**81.49**	**46.21**	3.54	**46.68**	76.89	**60.32**	**75.98**	**33.89**	**40.25**	28.09	1.08	15.18	14.96	**50.89**	**42.60**

4.3.6 Evaluation and Visualization Results on SemanticPOSS

Table 4.4 demonstrates results on the SemanticPOSS dataset, which captures the dynamic environment of a university campus with complex interaction between pedestrians, vehicles, and static infrastructure. Across all splits, the proposed method consistently outperforms comparative approaches, yielding average improvements of 7.27% in overall mIoU over NOPS and 4.62% over CHNCD. Novel class segmentation advances by 6.64% compared to NOPS, while base class performance increases by 7.37% and 4.80% relative to NOPS and CHNCD respectively.

The SemanticPOSS-4^0 split reveals how different geometric structures respond to the framework's approach. Ground segmentation excels at 80.08%, outperforming NOPS (69.14%) and CHNCD (58.33%), demonstrating effectiveness with large planar surfaces. Car segmentation shows substantial improvement from 1.24% to 7.19%, while plant segmentation reaches 56.40% versus 42.07% for NOPS. These consistent gains across diverse object types yield a novel class mIoU of 47.31% and overall mIoU of 38.55%, advancing 9.17% beyond NOPS.

The SemanticPOSS-3^1 split demonstrates balanced improvements across both novel and base classes. For novel classes, fence segmentation shows significant enhancement from 3.94% to 28.57%, while people segmentation reaches 57.76% versus 56.78% for NOPS. Base class performance benefits significantly from knowledge transfer, with car segmentation improving from 28.70% to 41.19%.

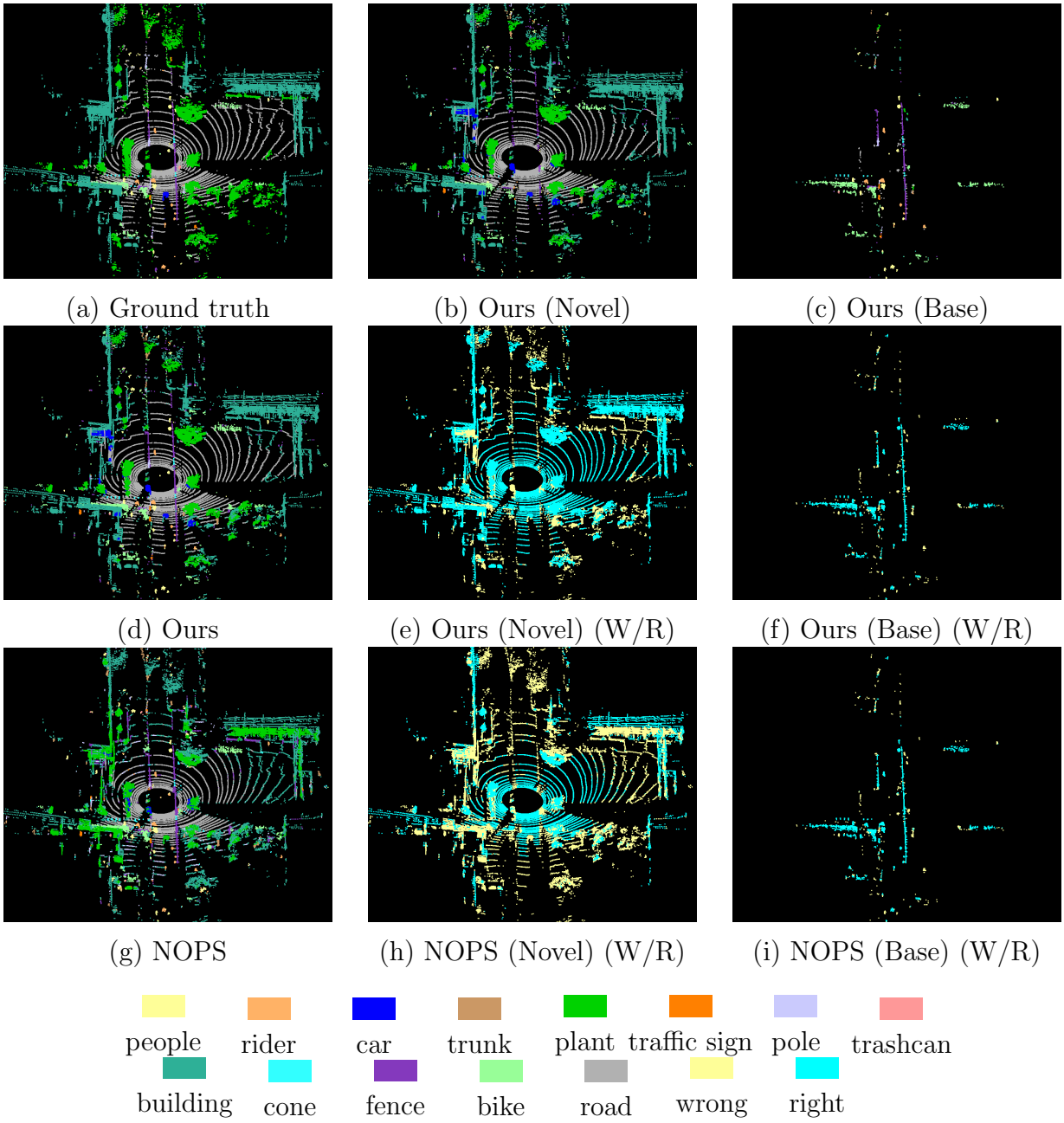

Figure 4.11: Segmentation results on SemanticPOSS-4^0 split. (a) Ground truth, (b) our novel class predictions (building, car, ground, and plant), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

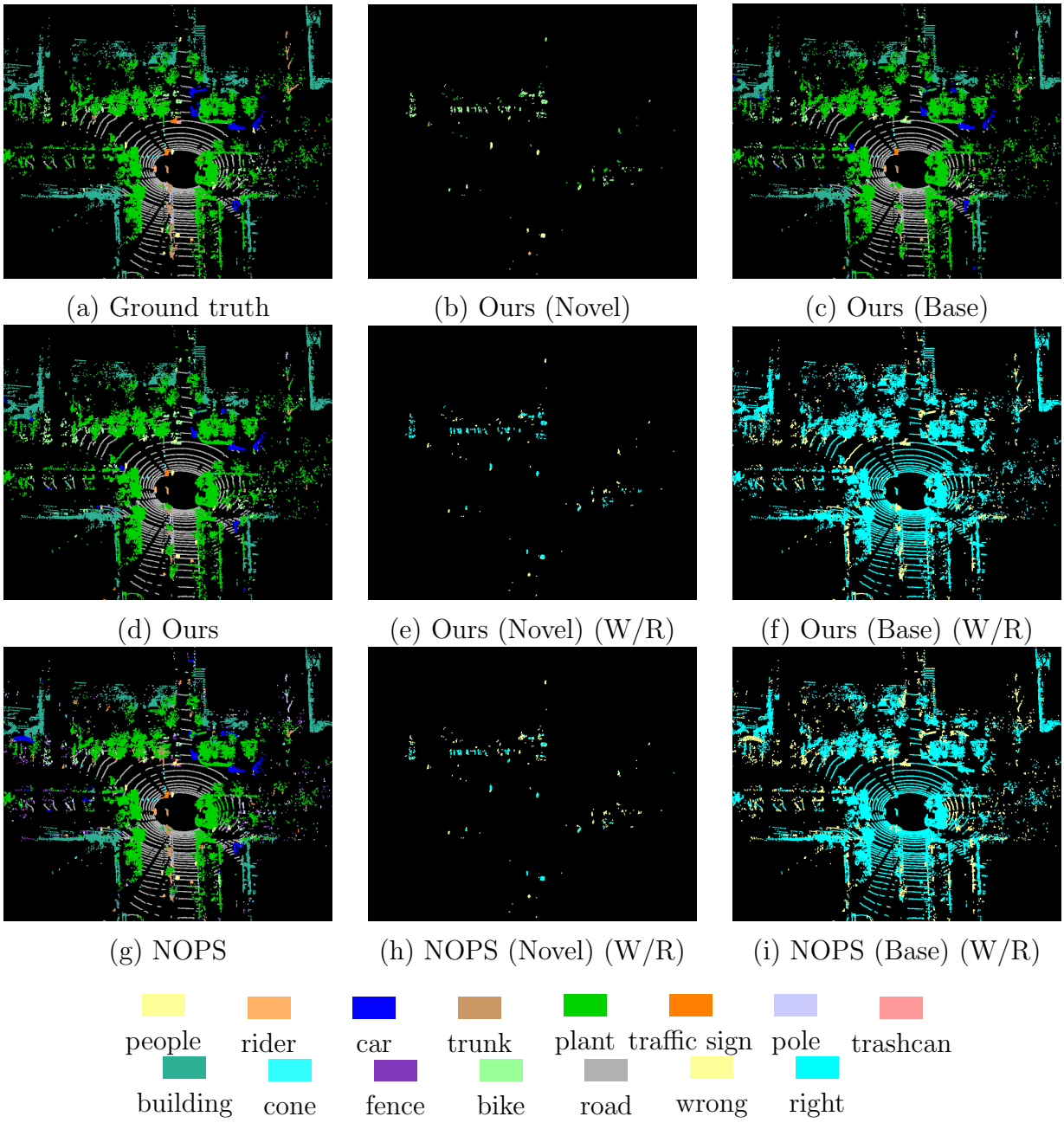

Figure 4.12: Segmentation results on SemanticPOSS-3[1] split. (a) Ground truth, (b) our novel class predictions (bike, fence, and people), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

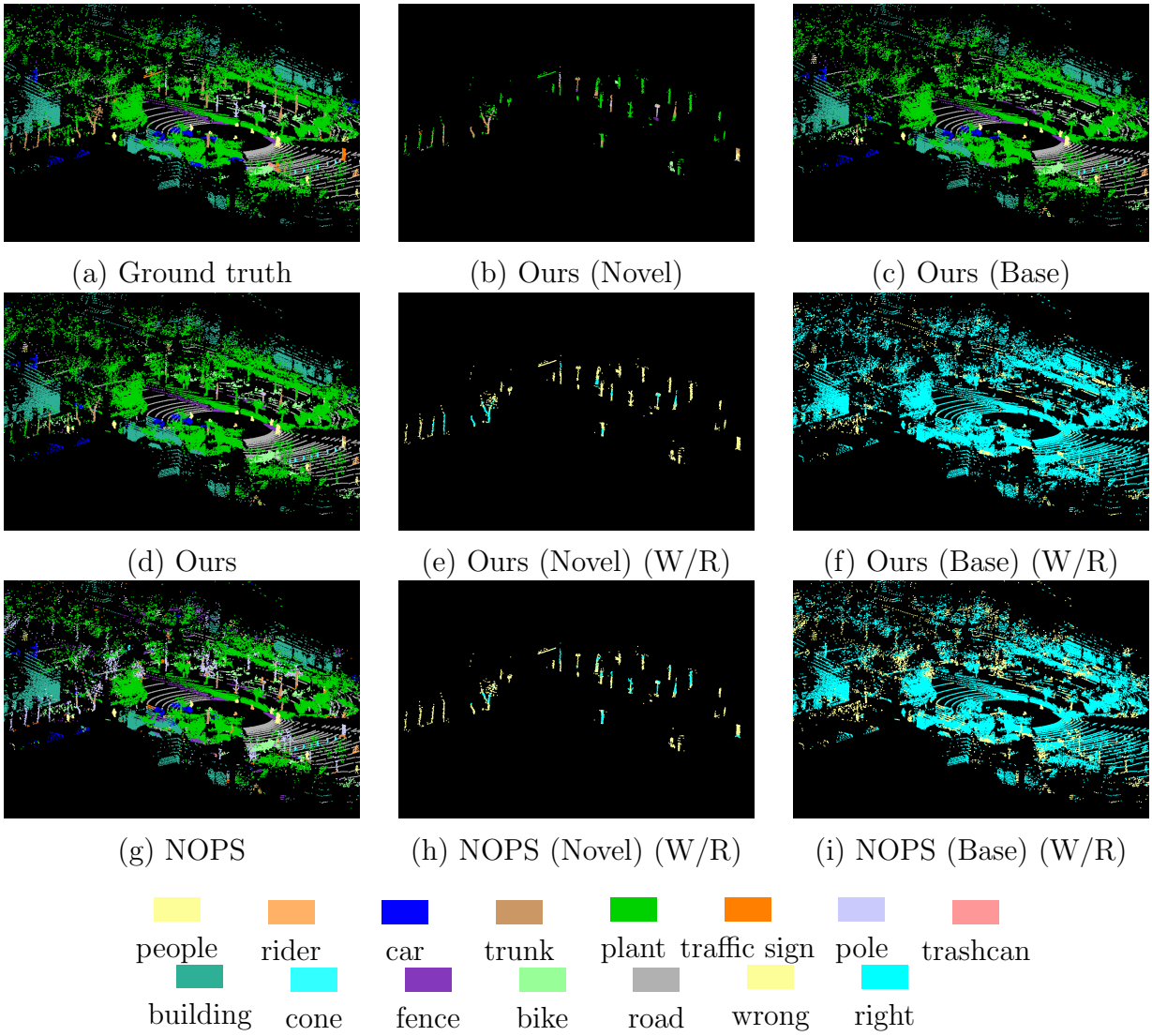

Figure 4.13: Segmentation results on SemanticPOSS-3^2 split. (a) Ground truth, (b) our novel class predictions (pole, traffic sign, and trunk), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

The SemanticPOSS-3^2 split targets pole, traffic sign, and trunk as novel classes, presenting varied segmentation challenges. Pole segmentation achieves remarkable improvement,

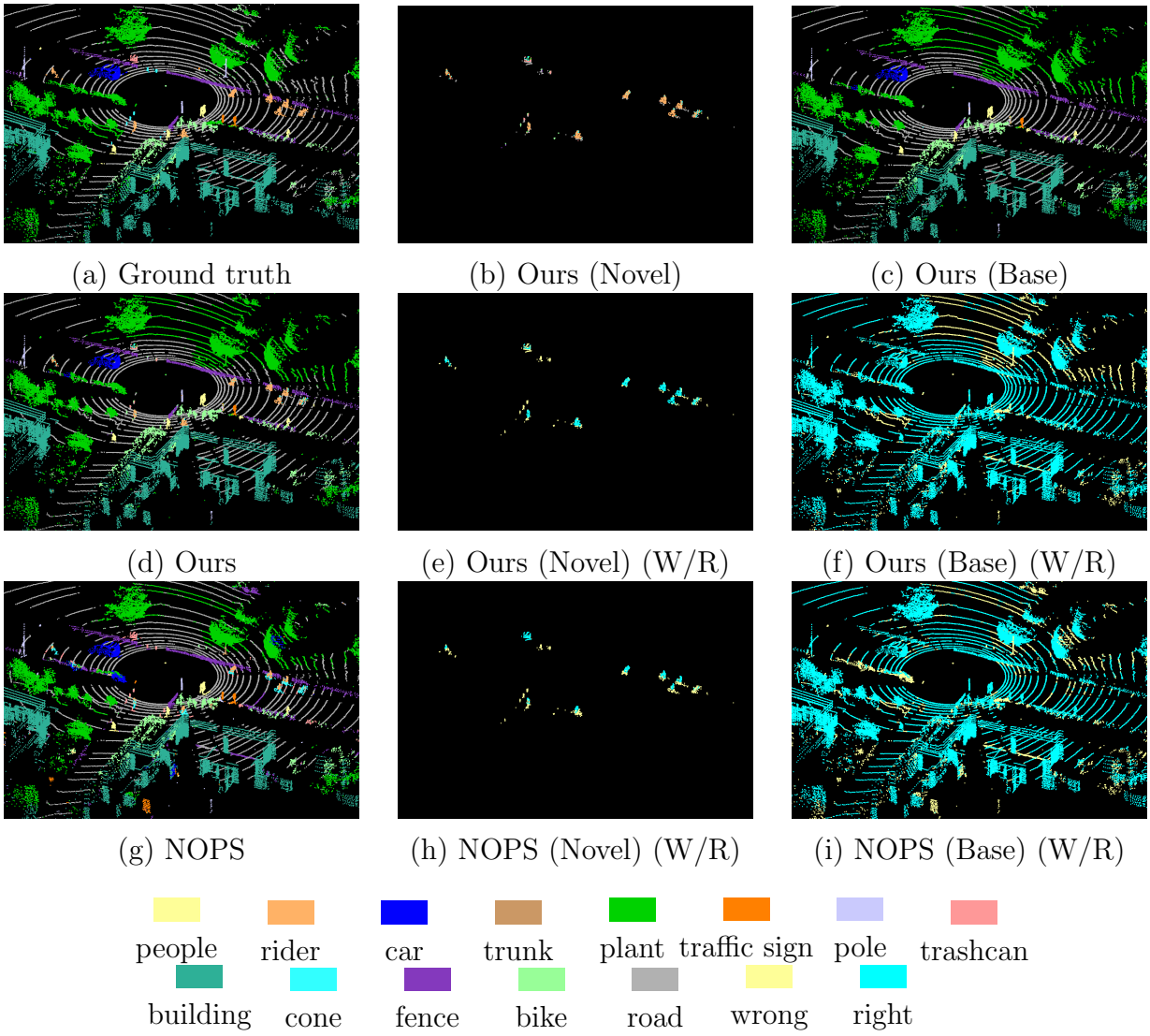

Figure 4.14: Segmentation results on SemanticPOSS-3^3 split. (a) Ground truth, (b) our novel class predictions (cone, rider, and trashcan), (c) our base class predictions, (d) our complete segmentation, (e-f) our error analysis for novel/base classes (cyan: correct, yellow: error), (g) NOPS complete segmentation, (h-i) NOPS error analysis for novel/base classes.

reaching 33.65% compared to 11.51% for NOPS and 9.61% for CHNCD, demonstrating the framework's effectiveness on vertical structures. Traffic sign segmentation shows modest gains from 0.54% to 1.34%, while trunk segmentation faces challenges, declining from

14.79% to 7.46%. Base class performance compensates with significant improvements, particularly in car segmentation (46.56% versus 29.74% for NOPS) and people segmentation (63.50% versus 52.09%). These mixed novel class results yield a novel mIoU of 14.15%, while strong base class performance (52.41%) produces an overall mIoU of 43.58%, advancing 7.54% beyond NOPS.

SemanticPOSS-3^3 highlights the challenge of segmenting small, geometrically varied objects. The proposed method successfully detects cones at 3.54% where NOPS completely fails (0.00%), though this remains below CHNCD's 13.66%. Rider segmentation shows more substantial improvement at 40.25% versus 31.87% for NOPS, benefiting from established knowledge of person form dynamics. Trashcan segmentation remains difficult at 1.08% versus 0.95% for NOPS, illustrating limitations in transferring knowledge to categories with limited semantic correspondence to known classes. Base class performance compensates significantly, reaching 50.89% with exceptional building segmentation (81.49% versus 70.36% for NOPS) and car segmentation (46.21% versus 30.91%). These balanced improvements produce an overall mIoU of 42.60%, representing a 6.28% gain over NOPS.

Figure 4.11 illustrates novel class discovery performance on SemanticPOSS-4^0 split within a complex campus environment, where building, car, ground, and plant serve as novel categories. The ground segmentation in Figure 4.11(b) achieves extensive spatial coverage with consistent surface segmentation throughout the scene, representing the strongest novel class performance. Plant segmentation shows coherent vegetation clustering with clear spatial boundaries, while building structures maintain accurate facade segmentation despite architectural complexity. Car segmentation demonstrates improved boundary definition compared to baseline methods, though challenges remain due to the sparse point representation of individual vehicles in the campus setting.

Base class performance in Figure 4.11(c) shows substantial improvements across multiple categories, with people segmentation achieving enhanced boundary preservation and cone segmentation demonstrating significantly better spatial accuracy. Pole structures maintain consistent vertical segmentation throughout the scene. The Wrong/Right analysis in Figures 4.11(e) and 4.11(f) reveals concentrated cyan regions for ground and plant categories, confirming superior novel class discovery. NOPS performance in Figures 4.11(g), 4.11(h), and 4.11(i) exhibits increased fragmentation with more scattered yellow error patterns, particularly visible in ground surface continuity and building facade preservation. The proposed method achieves improved performance in large-scale category discovery while maintaining enhanced accuracy across base classes in the dynamic campus environment.

Figure 4.12 presents segmentation outcomes on the SemanticPOSS-3^1 split, focusing

on the novel classes bike, fence, and people within a campus environment characterized by dense and heterogeneous object distributions. As illustrated in Figure 4.12(b), the model effectively identifies human figures dispersed across the scene, with boundaries that generally align with the true object extents. Bicycle instances are also clearly delineated, including those positioned adjacent to buildings or along pathways. In the full-scene prediction shown in Figure 4.12(d), the integration of novel and base class predictions preserves semantic consistency across diverse object types, capturing the spatial layout of campus facilities without evident boundary degradation.

In terms of base class segmentation, Figure 4.12(c) highlights improved performance in delineating car and rider categories, where object shapes and edges are more coherent compared to previous methods. Vertical elements such as pole structures are retained with minimal disruption across varying point densities, and cone instances are successfully identified in locations where object scale and background interference may pose challenges. The prediction correctness maps in Figures 4.12(e) and 4.12(f) indicate large cyan areas across both novel and base classes, suggesting consistent agreement with the ground truth. Compared with the NOPS results in Figures 4.12(g)–4.12(i), where error regions appear more dispersed and fragmented, our method exhibits better alignment with scene geometry and maintains more stable class boundaries throughout the segmented campus area.

Figure 4.13(b) demonstrates complex misclassification patterns where trunk instances are predominantly misclassified as plant categories, reflecting the geometric similarity between tree trunks and plant elements that share vertical cylindrical characteristics and natural textures. Pole segmentation exhibits confusion with portions being incorrectly assigned to trunk categories due to shared cylindrical vertical characteristics, while traffic sign segmentation shows difficulties with instances being misclassified as both people and trunk categories. This indicates the challenge of distinguishing small mounted objects from human figures and vertical structures based on sparse geometric features. Base class performance maintains robust quality in Figure 4.13(c), while Figure 4.13(f) shows extensive cyan regions confirming stable performance for well-represented categories. NOPS encounters similar confusion patterns among these vertical infrastructure categories, as demonstrated in Figures 4.13(g), 4.13(h), and 4.13(i), where pole, trunk, and traffic sign categories are frequently misclassified with each other. The fundamental challenge stems from distinguishing categories that share structural primitives but serve different functional roles in urban environments. These outcomes illustrate the specific limitations of geometric-based approaches when handling structurally similar but semantically distinct urban elements.

Figure 4.14 presents segmentation outcomes on the SemanticPOSS-3^3 split, where the novel classes (cone, rider, and trashcan) correspond to the categories with the fewest point

Table 4.5: Ablation study on Toronto-3D-2^0. Pink highlighted values indicate novel classes (road and building). Each row corresponds to a model variant: w/o AGSM indicates the model without the AGSM module; w/o DGE removes the DGE module; w/o GMSFF disables GMSFF by replacing gated fusion with simple concatenation; Full model includes all modules.

	Per-class IoU								mIoU		
Model	**road**	**road marking**	**natural**	**building**	**utility line**	**pole**	**car**	**fence**	**Novel**	**Base**	**All**
w/o AGSM	83.09	11.48	94.40	46.86	**82.81**	74.69	66.04	24.14	64.98	58.93	60.44
w/o DGE	85.95	15.64	93.19	46.85	79.59	76.41	66.37	25.27	66.40	59.41	61.16
w/o GMSFF (concat)	78.22	**16.15**	94.71	48.42	81.98	76.98	**69.35**	**36.51**	63.32	**62.61**	62.79
Full model	**86.55**	12.57	**94.85**	**55.99**	80.70	**80.79**	64.85	34.80	**71.27**	61.43	**63.89**

instances in the dataset. This split specifically targets the most challenging scenario where novel classes suffer from severe data scarcity, limited size, sparse point representation, and high geometric variability. In Figure 4.14(b), the novel class segmentation reveals substantial cross-category misclassification patterns where cone instances are frequently misclassified as bike and trashcan categories, likely due to geometric similarity in cylindrical structures and insufficient point samples to establish reliable discriminative features. Rider segmentation demonstrates boundary definition challenges where rider boundaries are incorrectly assigned to cone categories, reflecting the difficulty of distinguishing human silhouettes from nearby small objects. Trashcan segmentation exhibits complex error patterns with portions of trashcan instances being misclassified as cone categories. This cross-category confusion among the most point-sparse categories indicates that the framework struggles to establish reliable geometric discriminants when limited point density, severe data imbalance, and small object size collectively constrain feature learning. The Wrong/Right analysis in Figure 4.14(e) confirms these misclassification patterns, showing concentrated yellow error regions particularly evident in the sparse distribution of cone and trashcan instances throughout the scene.

Comparative analysis with NOPS reveals that both methods face similar fundamental challenges when dealing with the most data-scarce categories, though our approach demonstrates reduced error severity. In Figures 4.14(g), 4.14(h), and 4.14(i), NOPS exhibits more severe misclassification patterns, particularly for rider segmentation where substantially larger boundary regions are incorrectly assigned to cone categories compared to our method shown in Figure 4.14(b). Base class performance in Figure 4.14(c) demonstrates significantly better segmentation quality with accurate building and people segmentation. The Wrong/Right analysis in Figure 4.14(f) shows extensive cyan regions, confirming maintained performance for point-abundant categories. These results highlight the fundamental challenge of disambiguating small, point-sparse objects with limited geometric distinctive-

Table 4.6: Ablation study of AGSM module components on Toronto-3D-2[1] split (%). Pink highlighted values indicate novel classes (natural and car). Each row demonstrates the performance impact of removing or simplifying a specific AGSM component: (1) w/o Adaptive Sampling removes learnable dimension weights and uses standard FPS; (2) w/o Dynamic Grouping uses fixed group sizes instead of density-aware adaptation; (3) w/o Multi-aspect Encoding simplifies to basic convolution without multi-scale feature fusion; (4) w/o Context Integration replaces GCIU units with simple linear transformations.

	Per-class IoU								mIoU		
AGSM Configuration	**road**	**road marking**	**natural**	**building**	**utility line**	**pole**	**car**	**fence**	**Novel**	**Base**	**All**
w/o Adaptive Sampling	93.77	16.65	52.27	63.32	83.51	77.81	8.12	27.65	30.20	60.45	52.89
w/o Dynamic Grouping	93.64	16.90	**54.64**	62.44	81.98	74.92	5.66	33.81	30.15	60.62	53.00
w/o Multi-aspect Encoding	**93.90**	**19.82**	50.01	63.42	82.32	77.39	6.24	**39.17**	28.13	**62.67**	54.03
w/o Context Integration	93.60	17.51	47.10	61.93	81.14	**79.32**	15.65	31.13	31.38	60.77	53.42
Full AGSM	93.57	17.13	50.00	**64.79**	**84.57**	76.87	**20.16**	34.86	**35.08**	61.97	**55.24**

ness in current geometric-based novel class discovery approaches.

4.3.7 Ablation Study on Framework Modules

To evaluate the impact of each individual module, we conduct an ablation study on the Toronto-3D-2[0] split. Table 4.5 summarizes the results. Compared to the full model, removing AGSM leads to a noticeable drop in the Road and Building categories, which are part of the novel set, indicating that geometric sequence modeling improves spatial reasoning for unfamiliar structures. Excluding DGE causes a moderate decline in novel class performance, confirming the benefit of probabilistic feature encoding. Disabling GMSFF results in reduced performance across both base and novel classes, demonstrating the advantage of gated feature integration over simple concatenation. These findings suggest that the three modules contribute complementary capabilities: AGSM captures structured spatial patterns, DGE enhances feature expressiveness, and GMSFF ensures effective multi-source fusion. Their combined effect enables more accurate segmentation of both seen and unseen categories in complex urban scenes.

4.3.8 Ablation Study on AGSM

To address the critical need for understanding individual module contributions in novel class discovery frameworks, we conduct comprehensive ablation studies on the AGSM module components using the Toronto-3D-2[1] split. Table 4.6 presents the systematic evaluation

of each component's contribution to the overall framework performance, with particular emphasis on novel class discovery capabilities.

The ablation results reveal distinct performance patterns that illuminate the hierarchical importance of AGSM components for novel class discovery. The removal of Adaptive Sampling demonstrates significant impact on novel class performance, with novel class mIoU dropping from 35.08% to 30.20%, a decrease of 4.88%. This degradation indicates that learnable dimension weighting in the adaptive FPS algorithm is crucial for representative center selection across diverse urban geometries. Without adaptive sampling weights, the framework struggles to prioritize geometrically relevant spatial dimensions, leading to suboptimal sampling centers that fail to capture essential structural information for novel class differentiation.

Dynamic Grouping removal exhibits substantial impact on novel class discovery, with novel class mIoU declining to 30.15%, representing a 4.93% decrease. The performance degradation demonstrates that density-aware grouping mechanisms are fundamental for handling the heterogeneous point distributions characteristic of urban environments. Fixed grouping strategies fail to adapt to varying local densities, resulting in information loss in sparse regions and redundancy in dense areas, ultimately compromising the framework's ability to extract discriminative features for novel categories.

The Multi-aspect Encoding ablation reveals the most severe performance degradation, with novel class mIoU dropping to 28.13%, a substantial decrease of 6.95%. This significant decline underscores the importance of comprehensive geometric feature extraction that captures point-level patterns, positional relationships, directional cues, edge structures, and global context information. The simplified convolution-based encoding fails to capture the rich geometric relationships essential for distinguishing novel classes from structurally similar known categories, particularly in complex urban scenarios where subtle geometric differences determine semantic boundaries.

Context Integration removal shows moderate impact, with novel class mIoU decreasing to 31.38%, a decline of 3.70%. This demonstrates that while GCIU units with adaptive gating and residual connections enhance feature quality, the framework maintains reasonable performance through the foundational geometric encoding capabilities. The smaller impact suggests that hierarchical context integration primarily refines features rather than providing fundamental discriminative information, though its contribution remains valuable for optimal novel class discovery performance.

Notably, base class performance exhibits remarkable stability across all ablations (60.45%-62.67% vs 61.97% for full model), with variations within 2.22%. This stability indicates that AGSM enhancements specifically target novel class discovery challenges rather than

Table 4.7: Ablation study of DGE module components on SemanticPOSS-3^2 split (%). Pink highlighted values indicate novel classes (pole, traffic sign, and trunk). Each row demonstrates the performance impact of modifying a specific DGE component: (1) w/o Rotation removes quaternion-based rotation parameters, using only position and scale; (2) Isotropic Gaussian simplifies anisotropic Gaussians to isotropic ones with uniform scaling; (3) w/o Adaptive Weighting replaces Mahalanobis distance-based adaptive weights with fixed uniform weights; (4) MLP-based Features generates features from Gaussian positions via MLP instead of learning independent feature parameters. Performance drops relative to the full DGE model quantify each component's contribution to novel class discovery and overall segmentation accuracy.

	Per-class IoU													**mIoU**		
DGE Configuration	**bike**	**building**	**car**	**cone**	**fence**	**ground**	**people**	**plant**	**pole**	**rider**	**traffic sign**	**trashcan**	**trunk**	**Novel**	**Base**	**All**
w/o Rotation	42.69	78.47	44.36	26.27	44.97	76.96	59.41	75.62	13.60	54.98	**17.27**	2.60	5.85	12.24	50.63	41.77
Isotropic Gaussian	43.82	77.96	40.91	22.89	45.06	76.43	60.20	75.97	16.59	54.94	16.45	**3.59**	**11.01**	**14.68**	50.18	41.99
w/o Adaptive Weighting	43.02	78.10	41.49	22.07	44.21	76.41	60.96	75.11	18.33	55.12	14.89	2.85	7.44	13.55	49.93	41.54
MLP-based Features	42.13	76.94	44.72	27.39	43.94	76.44	59.86	73.52	21.63	55.12	12.80	2.42	6.35	13.59	50.25	41.79
Full DGE	**45.73**	**80.53**	**46.56**	**31.66**	**45.48**	**78.02**	**63.50**	**76.62**	**33.65**	**55.22**	1.34	0.75	7.46	14.15	**52.41**	**43.58**

general segmentation capabilities. The targeted improvement validates our design philosophy of addressing the semantic gap between known and novel categories through enhanced geometric understanding. The overall mIoU degradation ranges from 1.21% to 2.35% across ablations, demonstrating that each component contributes meaningfully to the framework's core objective.

These results collectively validate the synergistic design of AGSM components, where adaptive sampling ensures representative geometric coverage, dynamic grouping handles spatial heterogeneity, multi-aspect encoding captures comprehensive geometric relationships, and context integration refines feature representations. The coordinated operation of these components enables effective novel class discovery in complex urban point cloud environments.

4.3.9 Ablation Study on DGE

To evaluate the contribution of each component within the DGE module to novel class discovery performance, we conduct comprehensive ablation experiments on the SemanticPOSS-3^2 split. Table 4.7 presents the performance impact of modifying or removing specific DGE components, revealing critical insights into the geometric modeling requirements for effec-

Table 4.8: Ablation study of knowledge transfer objectives on SemanticPOSS-3[3] split (%). Pink highlighted values indicate novel classes (cone, rider, and trashcan). Each row demonstrates the performance impact of removing a specific loss component from the complete framework. $\mathcal{L}_{\text{PR}}$, $\mathcal{L}_{\text{CA}}$, and $\mathcal{L}_{\text{ST}}$ denote Prototype Relation Loss, Contrastive Alignment Loss, and Semantic Transfer Loss, respectively. The checkmark (✓) indicates the loss is included, while the cross (×) indicates it is excluded.

			Per-class IoU													**mIoU**		
$\mathcal{L}_{\text{PR}}$	$\mathcal{L}_{\text{CA}}$	$\mathcal{L}_{\text{ST}}$	bike	building	car	cone	fence	ground	people	plant	pole	rider	traffic sign	trashcan	trunk	**Novel**	**Base**	**All**
×	✓	✓	43.10	76.62	37.26	0.00	42.66	**77.71**	**60.85**	**76.59**	33.46	**50.13**	**29.01**	0.12	**19.69**	**16.75**	49.70	42.09
✓	×	✓	42.92	78.51	37.92	0.27	41.88	76.82	57.39	75.10	**39.46**	41.22	25.04	0.69	19.26	14.06	49.43	41.27
✓	✓	×	40.36	78.19	42.10	0.01	40.68	76.19	48.65	74.40	38.35	26.96	24.57	**1.51**	17.25	9.49	48.07	39.17
✓	✓	✓	**44.16**	**81.49**	**46.21**	**3.54**	**46.68**	76.89	60.32	75.98	33.89	40.25	28.09	1.08	15.18	14.96	**50.89**	**42.60**

tive novel class discovery in urban point cloud environments.

The removal of quaternion-based rotation parameters demonstrates significant impact on novel class discovery, with novel class mIoU dropping from 14.15% to 12.24%, representing a 1.91% decrease. The impact varies notablely across different novel categories, revealing the heterogeneous geometric complexity requirements for urban structure discovery.

Converting anisotropic Gaussians to isotropic ones yields substantial decline from 33.65% to 16.59% for pole category, demonstrating that anisotropic modeling provides significant benefits for elongated structures, where the 17.06% performance drop indicates that directional scaling contributes meaningfully to capturing the geometric anisotropy inherent in vertical urban infrastructure.

Removing adaptive weighting mechanisms results in a novel class mIoU of 13.55%, representing a 0.60% decrease that demonstrates the moderate but consistent importance of spatial-aware feature aggregation based on Mahalanobis distances. The pole category maintains performance at 18.33% compared to 33.65% in the full model, showing a 15.32% decline that indicates spatial proximity-based weighting contributes meaningfully to vertical structure modeling, though the fundamental representational capacity depends more heavily on rotation and scale parameters.

Replacing learnable feature parameters with MLP-generated features yields a novel class mIoU of 13.59%, representing a 0.56% decrease that demonstrates the value of independent feature learning while revealing the relative importance of geometric versus semantic modeling components within the DGE framework. The trunk category shows

modest degradation to 6.35% from 7.46%, indicating that natural structures benefit from the flexibility of independent feature learning, where learnable parameters can adapt to the high geometric variability characteristic of tree structures more effectively than position-based generation schemes.

4.3.10 Ablation Study on Knowledge Transfer Objectives

To evaluate the contribution of the proposed knowledge transfer mechanisms to novel class discovery in 3D point clouds, comprehensive ablation experiments were conducted on the SemanticPOSS-3[3] split. Table 4.8 presents the quantitative impact of removing each loss component individually, revealing the intricate interplay between Prototype Relation Loss ($\mathcal{L}_{\text{PR}}$), Contrastive Alignment Loss ($\mathcal{L}_{\text{CA}}$), and Semantic Transfer Loss ($\mathcal{L}_{\text{ST}}$) in addressing the fundamental challenges of semantic gap bridging and distributional shift mitigation.

The removal of Prototype Relation Loss reveals its essential function in maintaining semantic coherence through prototype relationships. Without this loss, cone segmentation fails completely (0.00% IoU), indicating that the semantic alignment term $\mathcal{S}(\mathbf{\Pi}_K, \mathbf{\Pi}_N)$ is crucial for establishing meaningful connections between known urban structures and geometrically distinct novel objects. The absence of cohesion and diversity regularization ($\mathcal{C}$ and $\mathcal{D}$) leads to unstable prototype formation, where the model cannot maintain consistent representations for novel categories that lack extensive training data.

Contrastive Alignment Loss contributes through instance-level feature relationship preservation. As defined by the contrastive formulation with positive ($\mathbf{z}_i^+$) and negative ($\mathbf{z}_i^-$) pairs from known classes, this objective ensures that novel class instances maintain semantically meaningful relationships with known class features. The performance degradation from 14.96% to 14.06% novel mIoU demonstrates that instance-level contrastive learning prevents feature space degeneration. The near-zero cone segmentation (0.27% IoU) indicates that without proper instance-level anchoring to known classes, novel class features drift into poorly structured regions of the embedding space. The temperature-controlled softmax normalization (τ_{CA}) in the contrastive loss provides calibrated similarity measurements that are essential for maintaining discriminative boundaries.

The absence of Semantic Transfer Loss results in the most substantial performance reduction, with novel class mIoU decreasing to 9.49%. The significant impact demonstrates that aligning novel class distributions with weighted combinations of known class distributions is fundamental for addressing data scarcity. The adaptive weighting scheme w_k, determined by feature similarities, enables selective knowledge transfer from the most relevant known classes. Without this distributional guidance, novel classes with limited

samples cannot leverage the rich statistical properties learned from well-represented known categories, resulting in poor generalization.

The differential impact patterns across the three objectives reveal their complementary roles in addressing distinct aspects of the novel class discovery challenge. Prototype Relation Loss operates at the category level, establishing relationships that guide the discovery of structurally novel urban elements. Contrastive Alignment Loss functions at the instance level, maintaining local feature coherence and preventing representation collapse. Semantic Transfer Loss bridges these scales through distribution-level alignment, enabling comprehensive knowledge propagation that addresses both semantic gap and class imbalance challenges. The interdependence becomes evident: removing any single component leads to specific failure modes that the remaining components cannot compensate for, validating the multi-objective design.

4.3.11 Feature Discriminability Analysis

To validate the discriminative power of our learned feature representations, especially for novel class discovery, we employ t-SNE dimensionality reduction to visualize feature clustering quality in 2D space. This analysis is conducted on a file from the SemanticPOSS-4^0 split to provide intuitive evidence of our method's ability to cluster semantically similar features more effectively than baseline approaches.

Figure 4.15 presents a comprehensive comparison of predicted class distributions across different semantic groups in the t-SNE feature space. The visualization reveals that our method's predictions consistently align better with the ground truth clustering patterns. Most notably, for novel classes (bottom row), NOPS exhibits significant misclassification patterns with extensive orange regions that deviate from the ground truth distribution, while our method maintains clustering consistency much closer to the ground truth layout.

These complementary visualizations on SemanticPOSS-4^0 provide direct evidence that our approach learns more discriminative feature representations, enabling both better semantic clustering and more accurate point cloud semantic segmentation for novel class discovery tasks.

4.3.12 Computational Efficiency Analysis

To evaluate the practical deployment viability of our framework across diverse urban environments, we conducted comprehensive computational efficiency analyses during the inference phase on three representative datasets using identical hardware configurations and

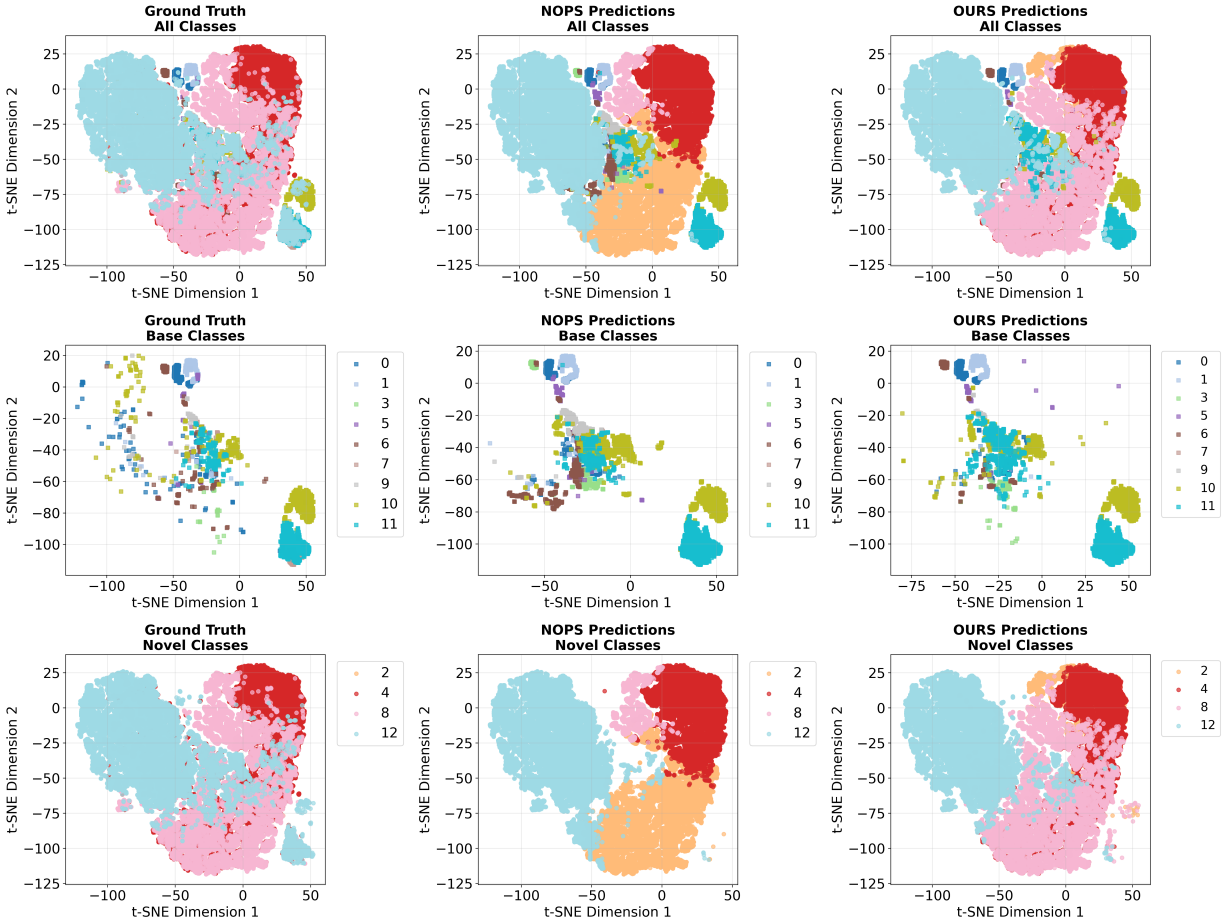

Figure 4.15: Feature discriminability comparison using t-SNE visualization on SemanticPOSS-4^0 split. The 3×3 grid compares ground truth labels (left column), NOPS predictions (center column), and our method's predictions (right column) for all classes (top row), base classes (middle row), and novel classes (bottom row). Circle and square markers represent novel and base classes, respectively.

experimental settings. All inference experiments were performed on an NVIDIA A100-SXM4-40GB GPU with consistent batch sizes, data preprocessing procedures, and evaluation protocols to ensure fair comparison between methods. Table 4.9 presents the inference computational resource requirements for our method and both baseline approaches NOPS and CHNCD across SemanticPOSS, SemanticSTF, and Toronto-3D datasets. The results demonstrate that our framework consistently requires additional computational resources compared to both baseline methods during inference, with memory overhead ranging from

Table 4.9: Comprehensive computational efficiency comparison across three datasets.

Dataset	Method	Initial Memory (GB)	Peak Memory (GB)	Memory Increase (GB)	Test Time (s)
SemanticPOSS-4^0	NOPS	0.14	1.39	1.25	100.96
	CHNCD	0.14	1.40	1.26	177.23
	Ours	0.31	3.09	2.78	151.00
SemanticSTF-5^0	NOPS	0.14	2.42	2.28	35.96
	CHNCD	0.14	2.43	2.29	77.82
	Ours	0.15	2.60	2.45	61.30
Toronto-3D-2^0	NOPS	0.14	7.73	7.59	18.72
	CHNCD	0.14	7.75	7.61	61.26
	Ours	0.16	8.44	8.28	30.90

1.07× to 2.22× compared to NOPS and 1.07× to 2.21× compared to CHNCD, and processing times that fall between the two baselines in most cases, being faster than CHNCD but slower than NOPS. Specifically, our method shows processing time ratios of 1.5× to 1.7× compared to NOPS and 0.50× to 0.85× compared to CHNCD across different urban scenarios. The variations in computational requirements across datasets reflect the inherent complexity differences between urban environments.

The inference computational overhead is justified by the consistent performance improvements achieved across all evaluated datasets under identical experimental conditions. Compared with both NOPS and CHNCD baselines, our framework achieves notable improvements in both novel class discovery and overall segmentation performance across all three datasets. The peak inference memory requirements remain well within the capacity of modern GPU infrastructure commonly deployed in urban management systems, ranging from a minimum of 2.60GB on SemanticSTF to a maximum of 8.44GB on Toronto-3D, representing only 6.5% to 21.1% of the available A100 memory capacity. While our method requires more memory than both baselines, the memory increase is generally modest, ranging from 0.17GB to 1.70GB additional compared to the baseline methods. The inference processing time overhead, while noticeable compared to NOPS, remains acceptable for most practical urban applications that typically operate in batch processing modes where accuracy and reliability take precedence over real-time constraints. Furthermore, our method achieves faster processing times than CHNCD across all datasets, demonstrating computational efficiency advantages over this more recent baseline while maintaining enhanced segmentation performance.

4.4 Discussion

4.4.1 Framework Performance Analysis

The proposed framework addresses novel class discovery in 3D point cloud semantic segmentation through three coordinated components. The AGSM module implements density-aware adaptation with learnable dimensional weighting in the sampling process. Ablation studies show a 4.88% performance decrease when the Adaptive Sampling component is removed, indicating its necessity for handling heterogeneous point distributions in urban environments. The adaptive sampling strategy adjusts to varying point densities across different urban structures, enabling more consistent feature extraction compared to fixed sampling approaches.

The DGE module provides continuous probabilistic representation of point clouds through learnable 3D Gaussians parameterized by position, scale, orientation, and features. The quaternion-based rotation parameters contribute to improved vertical structure segmentation, with pole segmentation showing 17.06% improvement over isotropic alternatives in the ablation study. This geometric modeling proves useful for urban infrastructure where elongated structures exhibit directional characteristics. The Mahalanobis distance-based weighting enables spatial-aware feature aggregation while maintaining computational tractability.

Our knowledge transfer strategy operates through three loss components at different semantic scales. Prototype Relation Loss establishes connections between known and novel class prototypes through learnable relationships. Contrastive Alignment Loss creates instance-level correspondences that help maintain feature quality for categories with limited examples. Semantic Transfer Loss implements distribution-level knowledge propagation via KL divergence alignment. Ablation results demonstrate that removing Semantic Transfer Loss causes the largest performance drop (5.47%), indicating its importance for transferring statistical knowledge from well-represented to scarce categories.

Experimental evaluation across three urban datasets shows consistent improvements compared to state-of-the-art methods. On Toronto-3D, overall mIoU increases by 6.59% over NOPS and 3.88% over CHNCD, with our method achieving robust performance across diverse urban structures including roads, buildings, and infrastructure elements. SemanticSTF results demonstrate 7.32% overall mIoU improvement over NOPS and 4.80% over CHNCD under adverse weather conditions including snow, rain, and fog, with vegetation segmentation showing particularly strong performance (44.15% vs 21.80% for NOPS, representing a 22.35% absolute improvement). SemanticPOSS evaluation yields 7.27% overall mIoU improvement over NOPS and 4.62% over CHNCD in dynamic campus environments.

Base class performance remains stable with improvements ranging from 6.64% to 8.21% over NOPS across the three datasets, while novel class discovery gains range from 4.61% to 6.64% over NOPS and 3.12% to 4.24% over CHNCD across the three datasets.

4.4.2 Practical Deployment Considerations

The framework addresses practical deployment considerations in urban management systems. The ability to segment novel infrastructure elements without extensive manual annotation provides value for infrastructure monitoring. Performance improvements on utility lines and pole segmentation translate to applications in maintenance planning where automated segmentation of new installation types can reduce inspection requirements.

Computational requirements represent a practical constraint. Memory usage ranges from 2.60GB to 8.44GB across datasets, representing 1.07× to 2.22× overhead compared to NOPS and 1.07× to 2.21× compared to CHNCD. Processing time increases by 1.5× to 1.7× compared to NOPS, while remaining faster than CHNCD, which is acceptable for batch processing scenarios typical in urban management applications where accuracy takes precedence over real-time constraints.

4.5 Chapter Summary

This chapter introduces AGDNet, which deepens representation and transfer through adaptive geometric sequence modeling, dynamic Gaussian embeddings, and gated multi-source fusion. AGSM provides adaptive sampling and dynamic grouping for heterogeneous spatial distributions, DGE implements continuous probabilistic modeling through learnable 3D Gaussians for local-global context capture, and GMSFF integrates these representations through context-aware weighting that adapts to scene characteristics.

The knowledge transfer approach operates at three semantic levels through Prototype Relation Loss for category-level connections, Contrastive Alignment Loss for instance-level correspondences, and Semantic Transfer Loss for distribution-level knowledge propagation. This strategy addresses semantic gap challenges while managing class imbalance issues in novel class discovery tasks.

Evaluation across Toronto-3D, SemanticSTF, and SemanticPOSS datasets demonstrates consistent improvements over state-of-the-art methods. The framework shows substantial gains in both overall performance and novel class discovery capabilities while maintaining

stable base class performance, validating its practical viability for deployment in urban systems.

This work advances representation and transfer through adaptive geometric modeling and multi-level knowledge transfer mechanisms. The following chapter bridges research frameworks with operational deployment by situating discovery within land cover mapping applications.

Chapter 5

3DLCDM: Hybrid Supervision for Land Cover Discovery Mapping[3]

5.1 Introduction

Chapter 3 established geometry-aware organization through cluster-based sampling and proximity hashing, improving boundary fidelity and neighborhood stability across diverse point cloud datasets. Chapter 4 deepened representation through adaptive geometric sequence modeling and gated multi-source fusion, enabling principled knowledge transfer from known to novel categories. Building on these advances, this chapter addresses a distinct challenge: how can discovery mechanisms be integrated with supervised segmentation to enable continuous land cover mapping in evolving urban environments characterized by extreme class imbalance and operational scale constraints?

To address this challenge, we introduce Land Cover Discovery Mapping (LCDM), which integrates NCD [41, 81, 31, 106, 20] with high-precision 3D land cover mapping to enable dynamic adaptation to emerging urban features. Unlike conventional methods [68, 109, 110, 104, 99] that maintain static labels, LCDM dynamically segments and incorporates novel urban features as they emerge, yielding continuously updated semantic representations that adapt to the evolving urban fabric and provide more comprehensive characterization of urban morphology compared to static, fixed-label methods. This approach holds significant promise for environmental monitoring [64] and sustainable urban management [43].

[3]The content of the chapter has been published in Remote Sensing of Environment with the paper entitled "3DLCDM: Hybrid supervision for land cover discovery mapping of emerging urban structures in 3D remote sensing".

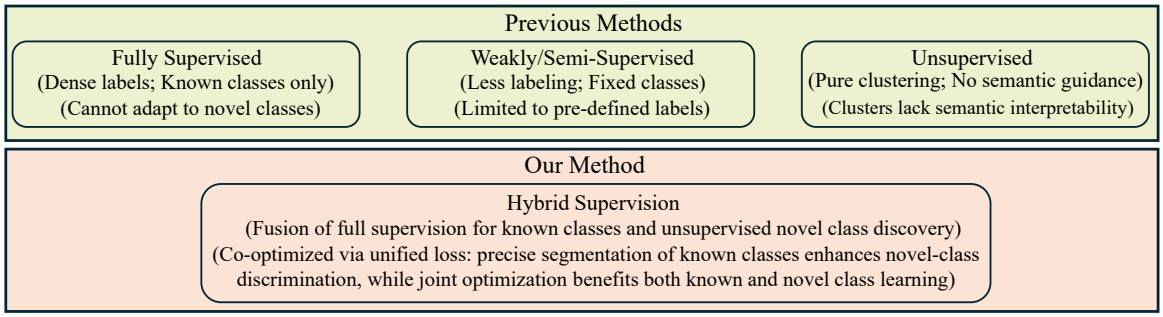

Figure 5.1: Comparison of traditional semantic segmentation methods and our unified end-to-end framework for land cover discovery mapping. The proposed approach integrates full supervision for known classes with unsupervised signals for novel class discovery, addressing the limitations of traditional methods.

While the promise of LCDM is evident, the task of segmenting urban 3D point clouds to support such mapping faces several substantial challenges. First, urban scenes are highly heterogeneous and continuously evolving [77, 39]. Although standard classes such as buildings, roads, and vegetation are typically well annotated, many environmentally critical features remain unlabelled [36, 39]. For example, emerging elements such as adaptive green infrastructure, smart city installations, temporary urban fixtures, and interactive urban elements frequently fall outside predefined label sets [33, 78, 82, 48, 65]. Their sporadic appearance and evolving nature demand land cover mapping systems with the capability to discover and segment novel classes as they emerge.

Second, urban point clouds are characterized by severe class imbalance [42, 52]. Dominant categories such as ground surfaces and large buildings may account for millions of points, while less frequent but environmentally significant features, for example chimneys, power lines, or minor street fixtures, are represented by only a sparse number of points [78, 82, 65]. This imbalance complicates model training, as the overwhelming number of points from dominant classes can obscure the learning signals required to accurately segment rarer elements [27, 7, 53].

Furthermore, urban data exhibit immense scale and structural complexity that is evident in variable point densities, heavy occlusions, and overlapping structures, all of which further exacerbate segmentation challenges [101]. Extracting fine-grained geometric details is essential for robust segmentation, particularly when the goal is to monitor dynamic urban evolution [56, 51].

Together, addressing this multifaceted problem is crucial for developing a mapping

framework that continuously adapts to the evolving urban fabric, thereby establishing a foundation for comprehensive environmental monitoring and guiding sustainable urban planning and management.

As described above, the task of segmenting urban 3D point clouds to support LCDM faces several substantial challenges, ranging from high scene heterogeneity with emerging, unannotated features, severe class imbalance, to the immense scale and structural complexity. Therefore, we have analyzed existing methods in remote sensing and found that they can be broadly classified into three categories: fully supervised methods [12, 100], weakly/semi-supervised methods [102, 13] and unsupervised methods [57, 83], each with its own shortcomings:

- Fully supervised methods: These approaches, which depend on dense labeling, achieve high accuracy on well-established classes but are inherently limited to a fixed set of semantic categories. They cannot automatically adapt to newly emerging urban features without retraining.
- Weakly/Semi-supervised methods: Although these methods reduce the annotation burden, they still operate within a predefined label space, thereby failing to capture novel, unannotated structures that are critical for understanding urban evolution.
- Unsupervised methods: Pure clustering techniques do not require manual labels; however, they lack semantic guidance and typically produce clusters based on geometric similarities rather than meaningful land cover categories, limiting their ability to discover interpretable novel features relevant to environmental monitoring.

Motivated by these limitations, we propose a novel end-to-end hybrid supervision framework tailored for LCDM. Figure 5.1 compares conventional semantic segmentation techniques with our unified framework. While standard methods rely on static semantic labels or unsupervised clustering without semantic guidance, our hybrid approach combines explicit supervision for known classes with unsupervised signals to discover novel urban features, enabling precise segmentation of established urban structures while simultaneously discovering emerging ones.

To systematically address the challenges of deploying NCD for operational land cover mapping, the following research questions (RQ) are formulated:

RQ1: How can supervised segmentation for known classes be coupled with unsupervised discovery mechanisms to enable continuous identification of emerging urban features without exhaustive re-annotation?

RQ2: How can pseudo-label generation be stabilized to ensure reliable novel class discovery under the noisy clustering conditions typical of unlabeled urban point clouds?

RQ3: How can severe class imbalance between dominant urban categories and rare but environmentally significant features be mitigated to achieve balanced learning?

RQ4: How can large-scale urban point clouds be processed efficiently while preserving geometric fidelity necessary for discovering fine-grained novel structures?

We propose 3DLCDM, a hybrid supervision framework that realizes LCDM through coupled pathways: precise segmentation of established classes via supervised learning, and unsupervised discovery of emerging features via dual prototype learning with temporal normalization and frequency-aware weighting. This integration enables operational land cover mapping at scale in large-scale remote sensing applications. The framework addresses each research question as follows:

- **Hybrid Supervision for Continuous Discovery (RQ1):** The 3DLCDM framework realizes LCDM through a fully integrated hybrid supervision method that bridges supervised segmentation of known classes with unsupervised discovery of emerging urban features. Unlike traditional methods that rely on fixed semantic labels, our framework continuously updates semantic representations by simultaneously segmenting established urban features and automatically discovering emerging classes. This capability is essential for accurately capturing the evolving urban fabric and provides a more comprehensive characterization of urban morphology.

- **Stabilized Pseudo-Label Generation (RQ2):** To effectively segment novel urban features through dedicated novel class discovery mechanisms, our method employs two complementary unsupervised heads. A primary branch maintains a fixed number of prototypes corresponding to the expected novel classes, while an over-segmentation branch initially leverages a larger number of prototypes that are progressively reduced through a scheduled process. Both heads compute raw similarity scores between features and prototypes, which are then transformed into balanced probability distributions through temporal Sinkhorn-Knopp normalization with dynamic temperature scheduling, ultimately generating discrete pseudo-labels that enhance the discovery of novel urban structures relevant to LCDM.

- **Adaptive Imbalance Mitigation (RQ3):** Recognizing the pronounced class imbalance inherent in urban point clouds, we propose a frequency-based weighting strategy that dynamically combines per-iteration batch weights with a long-term global weight vector through momentum updates. This adaptive mechanism ensures that rare yet semantically important classes receive sufficient emphasis during training,

leading to balanced gradient updates, improved segmentation performance, and more accurate land cover maps that support sustainable urban management.

- **Efficient Large-Scale Processing (RQ4):** The framework employs a shared sparse-convolutional MinkUNet backbone that maintains computational tractability for city-scale data while preserving geometric detail. Block partitioning strategies adapted to dataset characteristics (spatial grid for naturally tiled DALES, FPS+KNN for heterogeneous H3D) enable scalable processing without sacrificing the fine-grained geometric information necessary for discovering emerging urban structures.

5.2 Study Area and Dataset Processing

5.2.1 Study Area

The evaluation of our 3DLCDM framework is conducted in two distinct study areas, each capturing different aspects of built environment morphology and environmental conditions. The first region is located in Surrey, British Columbia, Canada, representing a diverse North American urban landscape characterized by a mixture of urban areas with multi-story buildings, suburban residential neighborhoods, commercial districts, and rural environments. The second region, located in Hessigheim, Baden-Württemberg, Germany, offers a uniquely different perspective due to its multi-temporal and ultra-high-resolution characteristics. Figure 5.2 presents an overview of the two study areas, highlighting their geographic locations and key dataset characteristics of the acquired point cloud data.

DALES is a large-scale aerial LiDAR dataset designed to facilitate remote sensing research in semantic segmentation and environmental analysis. It consists of over half a billion hand-labeled points captured from an airborne LiDAR system operating at approximately 1300 meters altitude, covering 10 km^2 of diverse terrain. Each scene spans 0.5 km^2 and contains key environmental elements such as ground surfaces, vegetation, and built infrastructure. The data encompasses both suburban and commercial regions, reflecting realistic aerial scanning conditions in urban and peri-urban settings. DALES offers eight semantic classes—ground, vegetation, building, car, truck, pole, power line, and fence. Each point's label was meticulously assigned following a thorough annotation protocol that leverages digital elevation models and expert verification. The dataset primarily supports evaluation of spatial variations in urban structure on a broader scale, enabling rigorous assessment of the LCDM's scalability, robustness, and generalizability across complex urban environments.

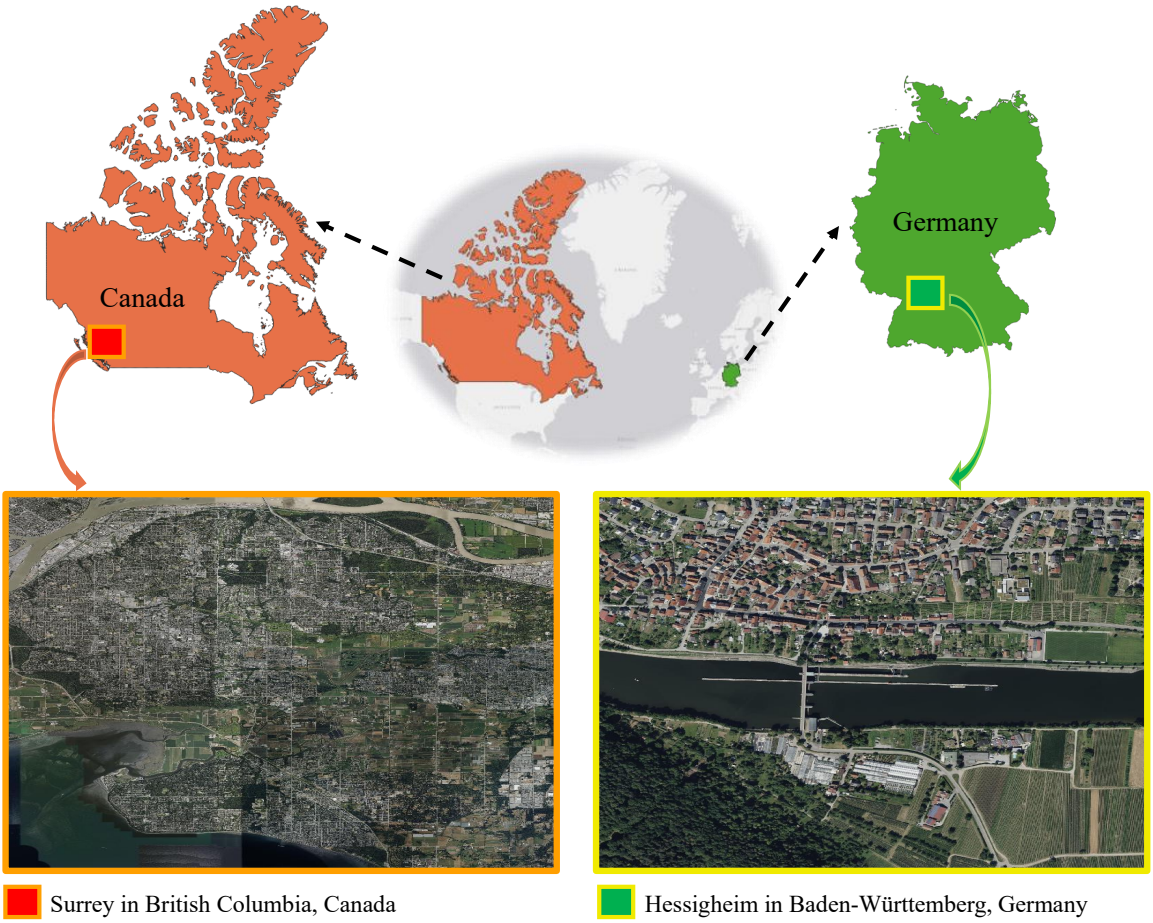

Figure 5.2: Geographic distribution and dataset characteristics of the DALES and H3D datasets used in this study. The DALES dataset covers 10 km^2 of urban area in Surrey, British Columbia, Canada, while the H3D dataset covers built environment in Hessigheim, Baden-Württemberg, Germany. Dataset statistics including point density, acquisition methods, and class information are provided to illustrate the complementary nature of these datasets for land cover discovery mapping evaluation.

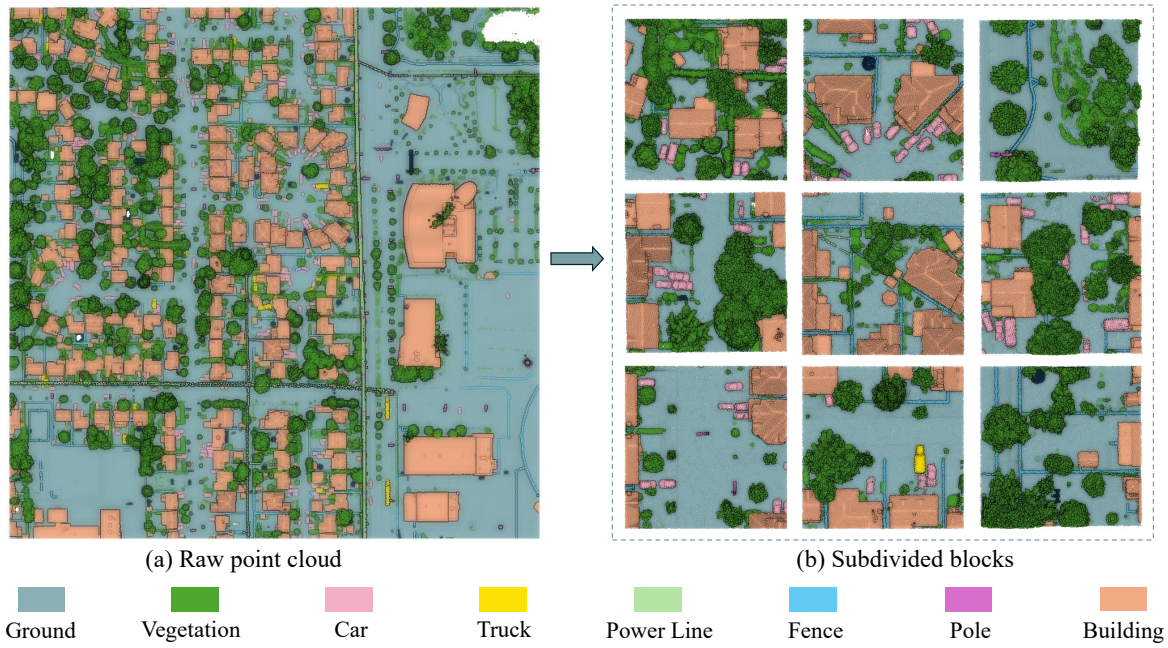

Figure 5.3: Illustration of the block partitioning results for the DALES dataset. (a) The raw point cloud. (b) The resulting subdivided blocks generated by spatial grid partitioning, each enforcing both a minimum point count and label diversity.

H3D is an ultra-high-resolution dataset featuring multi-temporal acquisitions across four epochs in Hessigheim, Baden-Württemberg, Germany. The primary three epochs (March 2018, November 2018, and March 2019) were collected using an unmanned aerial vehicle equipped with both a RIEGL VUX-1LR laser scanner and oblique Sony Alpha 6000 cameras, with the aerial platform flying at a nominal altitude of around 50 m and yielding a point density of approximately 800 points per square meter after strip overlap. An additional LiDAR-only epoch from March 2016 was acquired using manned aircraft with typical characteristics of national mapping agency LiDAR data. Each surveyed epoch covers various built environment features, allowing thorough investigations of low vegetation, impervious surface, vehicle, urban furniture, roof, façade, shrub, tree, soil/gravel, vertical surface, and chimney. The dataset serves as a complementary evaluation platform that focuses on fine-grained feature discrimination and precise semantic segmentation, enabling thorough assessment of the framework's capability in discovering novel structural features critical for comprehensive land cover mapping.

5.2.2 Data Preprocessing and Block Partitioning for DALES

The DALES dataset provides large-scale aerial LiDAR point clouds acquired over diverse urban, suburban, rural, and commercial areas. Owing to the large spatial coverage and substantial total number of points, directly processing the entire dataset at once can be both time-consuming and memory-intensive. In order to overcome these challenges and enable more efficient workflows in remote sensing, we adopt a block partitioning strategy, followed by a merging step for blocks that fail to meet specific criteria. This methodology ensures that each subdivided data file remains sufficiently large and diverse to be representative of its localized environment, as shown in Figure 5.3. Importantly, this partitioning and merging process preserves the complete dataset integrity, with the total number of points in the processed blocks remaining identical to the original dataset, ensuring no data loss during preprocessing.

(1) Spatial Partitioning. We begin by computing the bounding box of the entire tile. This bounding box is then discretized into a grid of rectangular cells of fixed size, for example, $50\,\mathrm{m} \times 50\,\mathrm{m}$. Each cell defines a potential block, and all points whose coordinates fall within the spatial extent of that cell are extracted into a corresponding point subset. This approach respects the natural spatial layout of the DALES data and preserves local context. Crucially, the choice of block size is important: smaller cells make each block easier to process, but risk insufficient representation of complex or large environmental structures such as expansive building rooftops or dense tree canopies. Larger cells preserve continuity but may still be too large for feasible memory and runtime constraints. Through empirical testing, a block size around 50 m provides a balance between manageability and coherent scene coverage for the DALES dataset.

(2) Block Evaluation and Merging. Some blocks will inevitably be underpopulated (*i.e.*, containing fewer than 80,000 points) or may contain only a single object category (for example, exclusively ground or vegetation). Such blocks have limited utility for downstream classification tasks, as they lack the diversity needed to train or test deep learning models effectively. To remedy this, we evaluate each block using two criteria:

1. Minimum point count: Blocks must contain at least 80,000 points.
2. Label diversity: Blocks should comprise more than one semantic label.

If a block fails either criterion, it is merged with one of its spatial neighbors, thereby increasing point count and potentially adding more class labels. By repeating this merging process iteratively, we ensure that all final blocks meet minimum data requirements, fostering a richer training context while retaining meaningful local structure. Note that the

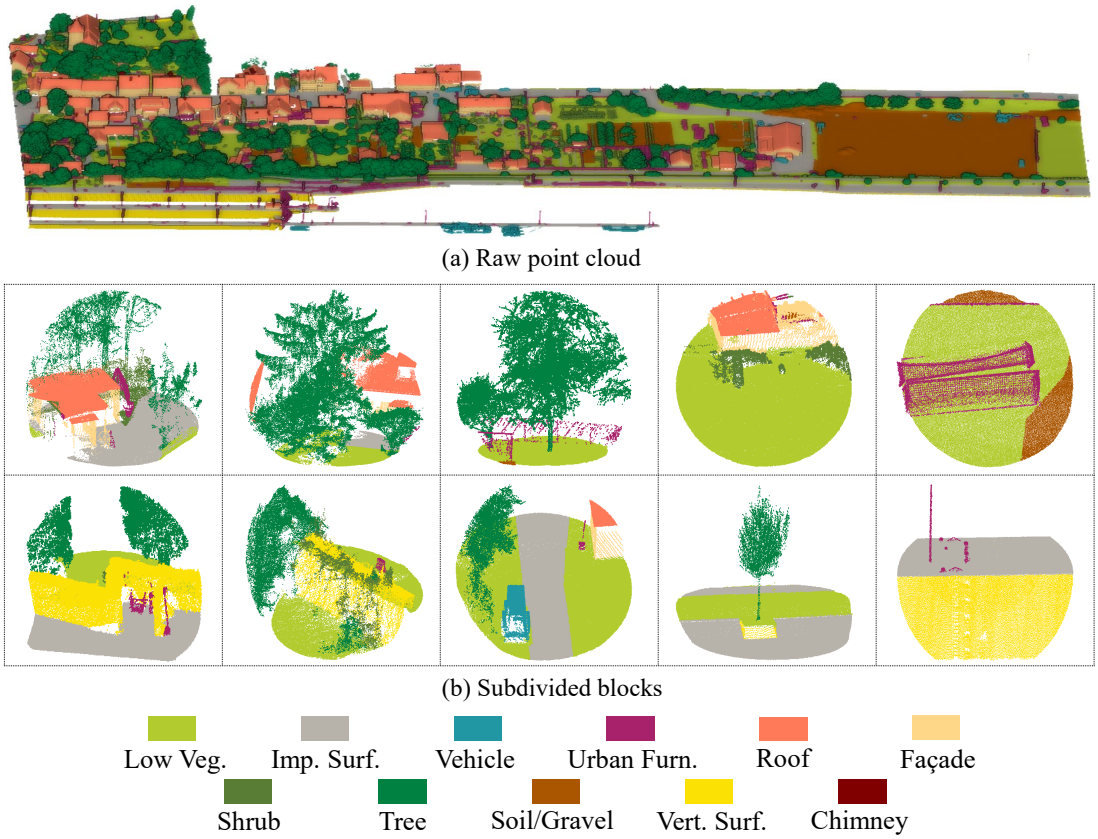

Figure 5.4: Illustration of the block partitioning results for the H3D dataset. (a) The raw point cloud. (b) The resulting subdivided blocks generated by applying FPS and KNN queries, each focusing on a smaller portion of the scene.

majority of blocks maintain their original 50m $\times$ 50m squared boundaries as shown in Figure 5.3, while only a small percentage of blocks, primarily those at dataset boundaries or in sparse regions, require merging and thus have irregular boundaries.

In summary, partitioning the DALES point clouds into smaller, well-defined blocks reduces memory requirements and computational load while preserving the local spatial and semantic diversity of the environment. This process ensures that each block contains a representative sample of its local land cover, which is essential for accurately capturing heterogeneous urban features in dynamic land cover mapping. Furthermore, the block-based approach facilitates parallel processing and timely updates of land cover maps, thereby supporting robust environmental monitoring and adaptive urban management. Our partitioning-and-merging pipeline transforms massive, unwieldy point clouds

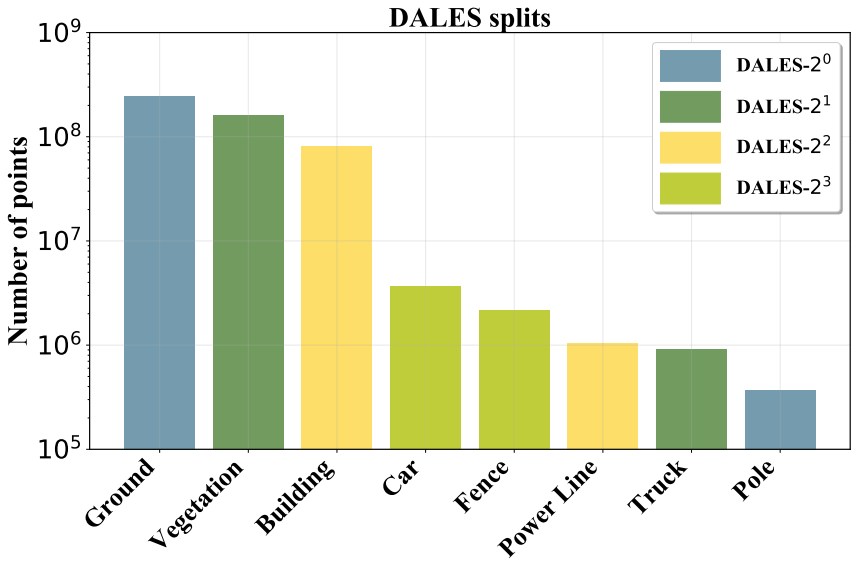

Figure 5.5: Point count distribution for each class in the processed DALES dataset. Classes are sorted by point count. Colors indicate split assignments: DALES-2^0 (blue-gray), DALES-2^1 (olive green), DALES-2^2 (yellow), and DALES-2^3 (yellow-green) for novel class discovery experiments.

into a set of manageable blocks that retain sufficient point density and semantic diversity, streamlining computational workflows while preserving critical environmental features for robust 3D scene understanding and advanced deep learning research in remote sensing.

The spatial grid partitioning strategy is specifically designed to leverage DALES' inherent tile-based organization, where the dataset is naturally structured as regular 0.5 km^2 tiles. By employing 50m $\times$ 50m grid cells that align with this pre-existing spatial structure, we maximize the preservation of original geometric relationships and spatial continuity across large-scale urban features. This approach maintains the integrity of extended urban structures such as building complexes, transportation corridors, and vegetation patches that span multiple spatial scales, ensuring that the natural spatial dependencies critical for comprehensive land cover discovery mapping are preserved throughout the preprocessing pipeline.

5.2.3 Data Preprocessing and Block Partitioning for H3D

In order to efficiently handle the extensive point cloud data provided by the H3D dataset for environmental analyses, we employ a two-stage partitioning strategy that combines

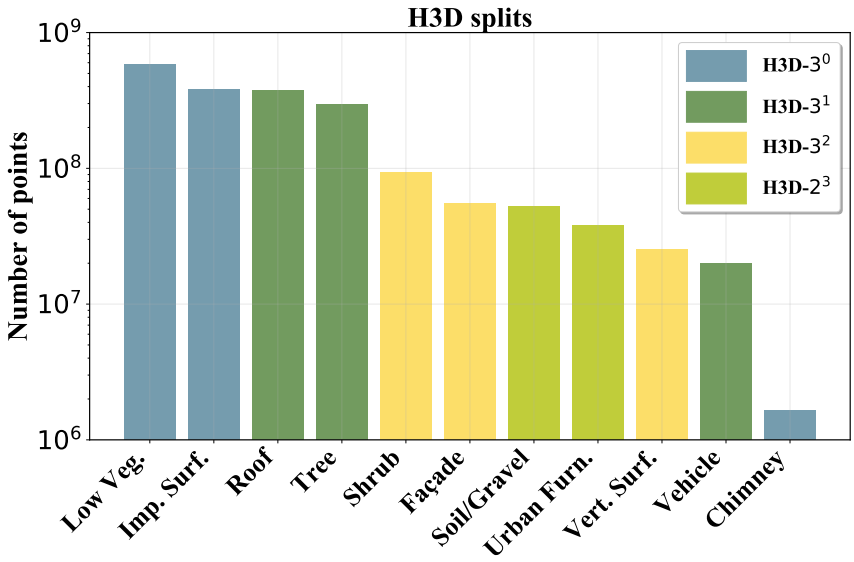

Figure 5.6: Point count distribution for each class in the processed H3D dataset. Classes are arranged by point count (∼5× original size). Colors represent split configurations: H3D-3^0 (blue-gray), H3D-3^1 (olive green), H3D-3^2 (yellow), and H3D-2^3 (yellow-green) for novel class discovery experiments.

Farthest Point Sampling (FPS) [69] and k-Nearest Neighbors (KNN) queries [87]. As illustrated in Figure 5.4, the raw H3D point cloud is subdivided into multiple blocks for fine-grained analysis.

First, we read the original PLY files to extract 3D coordinates (x, y, z), color values $(\text{red}, \text{green}, \text{blue})$, and labels. These labels are crucial for distinguishing various environmental features, such as different vegetation types or man-made structures. As an initial step, we subtract the coordinate offset $[513700, 5426400, 0]$ from each point's coordinates to bring the dataset into a more manageable range.

Next, we perform FPS to identify center points distributed throughout the scene. To ensure comprehensive coverage and robust block generation, we select 5 times more center points than the minimum required number. FPS selects points that maximize their mutual distances, ensuring that the sampled centers are representative of the entire spatial domain. This coverage is critical for environmental remote sensing tasks, as it avoids clustering all sample blocks in dense urban areas while neglecting sparsely populated regions. By selecting multiple FPS centers, we prepare the dataset for subsequent neighborhood searches without losing important landscape features.

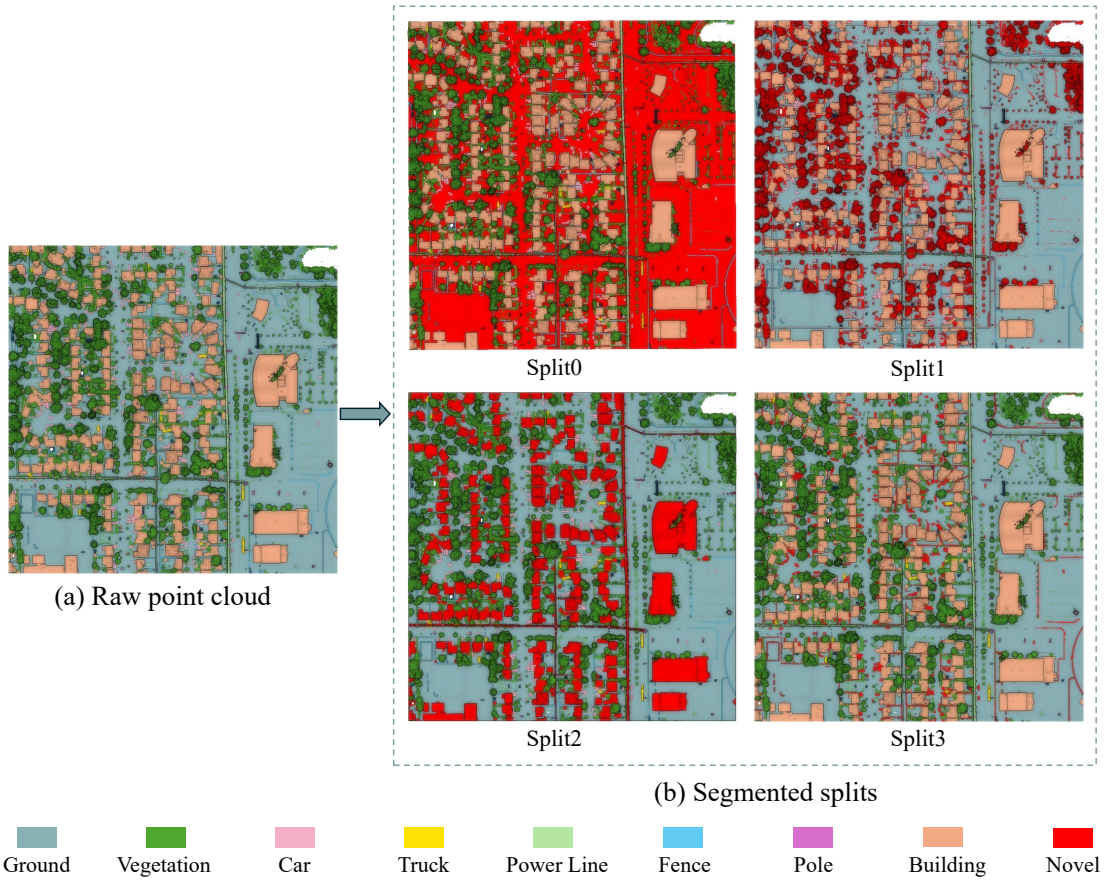

Figure 5.7: Class splitting in the DALES dataset. (a) Raw point cloud, (b) segmented splits showing different novel class discovery scenarios. Split0 (DALES-2^0): Ground and Pole as novel classes; Split1 (DALES-2^1): Vegetation and Truck as novel classes; Split2 (DALES-2^2): Building and Power Line as novel classes; Split3 (DALES-2^3): Car and Fence as novel classes.

Following this, each FPS center is used in a KNN search to retrieve a fixed number of nearest points, forming individual blocks. Specifically, we use KNN to gather 100,000 nearest neighbors around each FPS center, generating blocks of approximately 100,000 points each. The KNN approach naturally adapts to varying densities in the dataset: regions with dense LiDAR returns (e.g., dense foliage or complex urban structures) are contained in relatively small spatial extents to meet the target point count, while sparser areas are covered by larger spatial blocks but still maintain a similar number of points. This uniformity in block size is advantageous for parallel processing and reduces the memory footprint

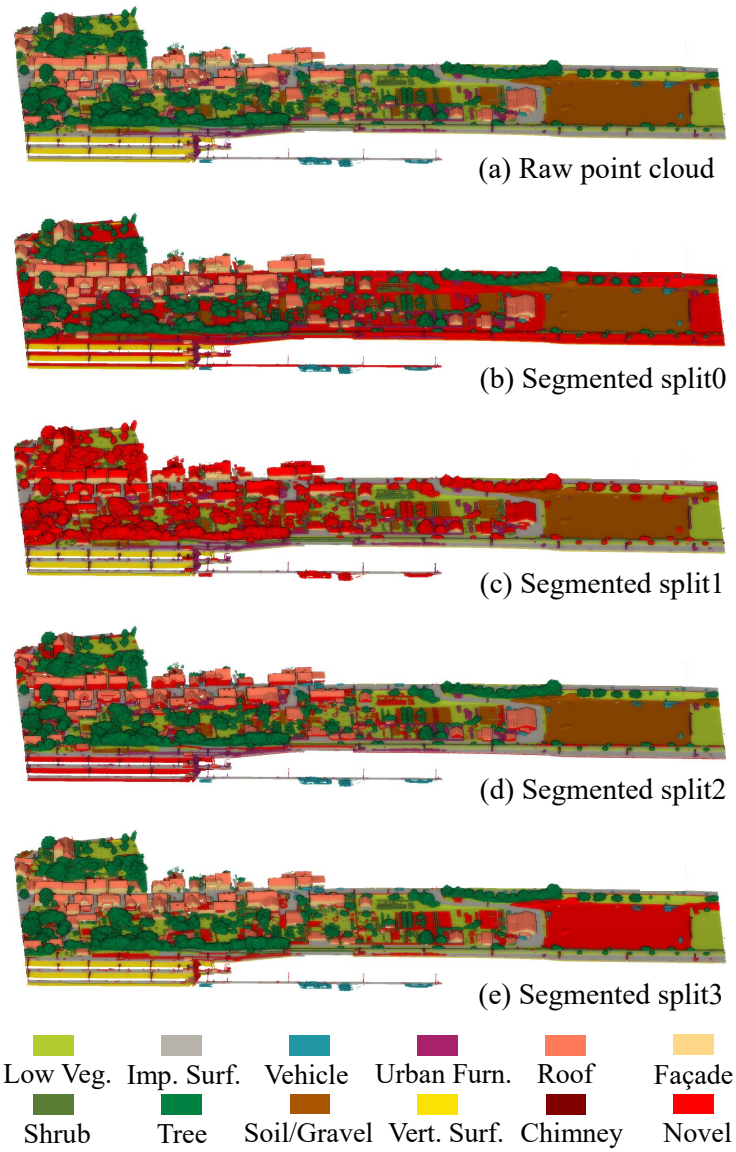

Figure 5.8: Class splitting in the H3D dataset. (a) Raw point cloud, (b) segmented splits showing different novel class discovery scenarios. Split0 (H3D-3^0): Low Vegetation, Impervious Surface, and Chimney as novel classes; Split1 (H3D-3^1): Vehicle, Tree, and Roof as novel classes; Split2 (H3D-3^2): Shrub, Vertical Surface, and Façade as novel classes; Split3 (H3D-2^3): Urban Furniture and Soil as novel classes.

during classification or segmentation tasks, making it feasible to deploy computationally intensive deep learning models for environmental analysis.

Moreover, special handling is implemented to ensure that blocks with only one class label are merged with multi-label blocks, promoting more informative local contexts in each block. After the initial FPS-based center selection and subsequent KNN partitioning, the algorithm checks for any points that remain unassigned to a block (unassigned points). If such points exist, an additional FPS sampling is performed on these remaining points to select new center points, followed by another round of nearest-neighbor allocation. Finally, if some points still remain unassigned after this secondary allocation, these residual points are either appended to existing blocks (when they are few in number or homogeneous in class) or organized into new separate blocks. These steps guarantee that the entire dataset is covered without creating excessively fragmented files.

Overall, this FPS and KNN strategy yields multiple smaller PLY files, each containing approximately the same number of points and a balanced distribution of urban classes. This partitioning facilitates scalable and memory-efficient analysis, enabling the deployment of advanced deep learning models for dynamic land cover mapping. By dividing H3D into manageable yet representative sub-blocks, our approach preserves fine-scale urban details and supports the continuous discovery of emerging urban structures. Notably, this preprocessing pipeline ensures complete coverage of all points in the original dataset while resulting in approximately 5 times as many data points as the original dataset without any data loss.

The FPS+KNN partitioning strategy is tailored to address H3D's higher class complexity and the inherent spatial heterogeneity of fine-grained urban features. Given the dataset's rich semantic composition with eleven distinct classes and the pronounced spatial imbalance characteristic of detailed urban scenes, where minority classes such as chimneys and urban furniture exhibit sparse and irregular distributions, the adaptive partitioning approach ensures that each block maintains representative coverage of diverse local geometric characteristics. The FPS center selection mechanism guarantees spatial coverage while the subsequent KNN gathering preserves the local geometric context essential for discriminating between semantically similar but geometrically distinct fine-scale urban elements, thereby maintaining the geometric integrity required for effective novel class discovery in complex urban environments.

5.2.4 Base-novel Class Splitting

To evaluate the efficacy of novel class discovery in dynamic urban environments, we implement a splitting strategy that designates specific semantic categories as novel during

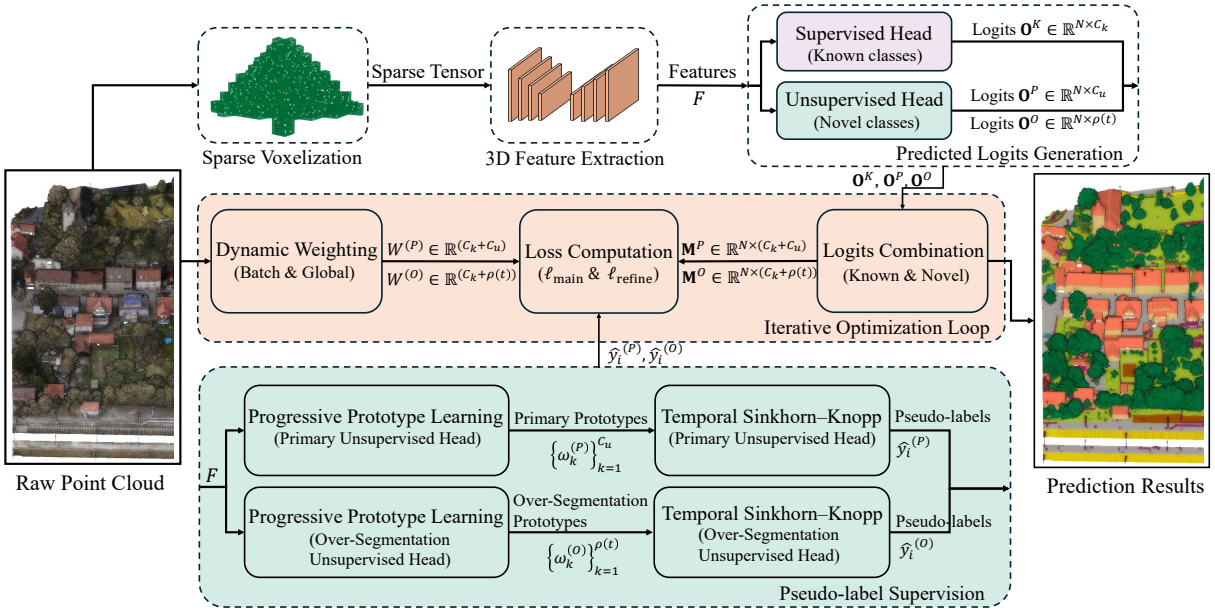

Figure 5.9: Network architecture of the proposed 3DLCDM framework for land cover discovery mapping. The system processes raw point clouds through sparse voxelization and extracts deep features F using MinkUNet34C. The extracted features are processed through two complementary pathways: (i) a supervised head generating known-class logits $O^K \in \mathbb{R}^{N\times C_k}$ for established urban categories, and (ii) a dual unsupervised head comprising a primary branch producing logits $O^P \in \mathbb{R}^{N\times C_u}$ with fixed prototypes $\{\omega_k^{(P)}\}_{k=1}^{C_u}$, and an over-segmentation branch generating logits $O^O \in \mathbb{R}^{N\times\rho(t)}$ with dynamic prototypes $\{\omega_k^{(O)}\}_{k=1}^{\rho(t)}$. Both unsupervised branches undergo temporal Sinkhorn-Knopp normalization to generate discrete pseudo-labels $\hat{y}_i^{(P)}$ and $\hat{y}_i^{(O)}$ for novel class discovery. The supervised and unsupervised logits are combined into unified score matrices $M^P \in \mathbb{R}^{N\times(C_k+C_u)}$ and $M^O \in \mathbb{R}^{N\times(C_k+\rho(t))}$. A dynamic weighting strategy computes adaptive weights $W^{(P)} \in \mathbb{R}^{C_k+C_u}$ and $W^{(O)} \in \mathbb{R}^{C_k+\rho(t)}$ based on batch and global frequency analysis to address class imbalance. The framework optimizes combined loss components ℓ_{main} and ℓ_{refine} through an iterative optimization loop, enabling accurate segmentation of known classes while discovering emerging urban features for continuous land cover map updates. N represents the number of points, C_k denotes the number of known classes, C_u represents the number of novel classes, and $\rho(t)$ indicates the dynamic prototype count.

both training and testing. The true ground truth labels of novel classes are only utilized to quantify discovery performance, enabling direct comparison between predicted novel class assignments and actual semantic categories. This experimental design enables quantitative assessment of the framework's ability to identify novel urban features, thereby simulating the continuous emergence of new land cover types characteristic of evolving urban landscapes.

The splitting strategy is designed based on class frequency distributions to ensure balanced experimental evaluation, as shown in Figure 5.5 for DALES and Figure 5.6 for H3D. Each dataset is divided into four experimental configurations with different subsets of classes designated as novel. The split assignment follows a largest-smallest alternation pattern: classes are ranked by their point counts, then distributed across splits to ensure each configuration contains both dominant and minority categories. This approach prevents any single split from being dominated exclusively by high-frequency or low-frequency classes, thereby maintaining experimental validity across varying class distributions. Following the unified experimental setting of the DALES dataset, the label value (0) representing the invalid class is excluded. Points with this label are entirely excluded from both training and evaluation processes.

The DALES dataset implements DALES-2^0, DALES-2^1, DALES-2^2, and DALES-2^3 configurations, while the H3D dataset employs H3D-3^0, H3D-3^1, H3D-3^2, and H3D-2^3 splits. For each experimental configuration, the designated novel classes have their original semantic labels withheld during training, with these classes treated as unlabeled data for novel class discovery, while the remaining categories retain their original labels as known classes with full supervisory signals. Figures 5.7 and 5.8 demonstrate the visual effect of these split configurations, where the red regions represent the novel classes within each scene for different experimental splits.

5.3 Method

5.3.1 Network Architecture

Figure 5.9 presents the overall pipeline of the proposed 3DLCDM framework. Raw point cloud data are first transformed into sparse voxel tensors using the MinkowskiEngine [10], which reduces computational overhead while preserving fine spatial details. The voxelized data are then processed by MinkUNet34C that extracts multi-scale deep features. These features are subsequently fed into two complementary pathways: a supervised head produces logits for known urban classes, and a dual unsupervised head employs prototype-

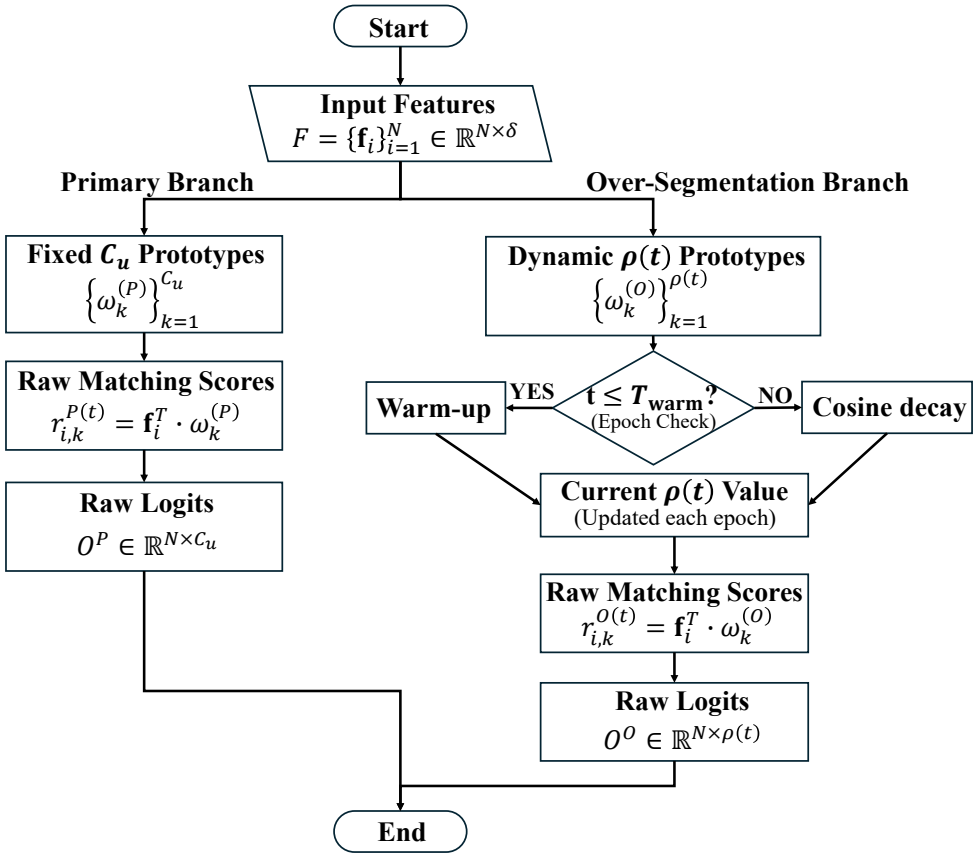

Figure 5.10: Dual-branch prototype learning architecture with progressive prototype scheduling. The framework processes unified features $F = \{\mathbf{f}_i\}_{i=1}^{N}$ through two parallel prototype learning branches: (1) A primary branch maintaining fixed C_u prototypes throughout training for stable clustering, and (2) An over-segmentation branch with adaptive prototype count $\rho(t)$. Both branches generate raw logits that are processed differently during training.

based learning to discover novel classes. In the dual unsupervised head, the primary branch uses a fixed number of prototypes while the over-segmentation branch initially employs a larger set of prototypes that are gradually reduced during training through progressive scheduling. The similarity scores between features and prototypes are normalized via the Sinkhorn-Knopp algorithm [76, 11] with temporal temperature scheduling to yield balanced discrete pseudo-labels, and a dynamic weighting strategy combines per-batch and global class frequency statistics to address class imbalance during loss computation. The supervised and unsupervised logits are combined into unified prediction matrices, enabling

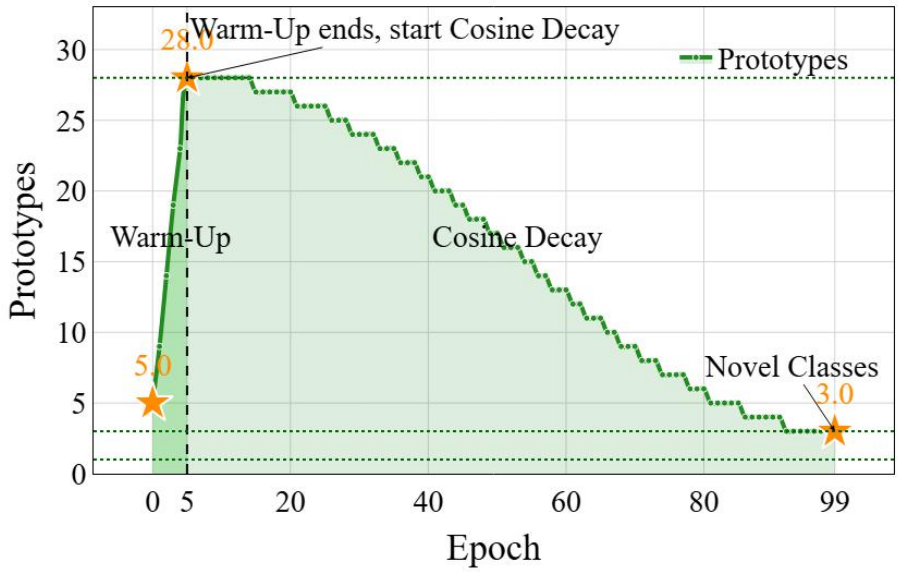

Figure 5.11: Illustration of how the prototype dimension $\rho(t)$ evolves over 100 epochs. A warm-up phase ramps the dimension from a small initial value (e.g., 5) to a high capacity (e.g., 28) by epoch 5, after which it holds steady briefly. A cosine decay then gradually lowers the dimension to the final number of novel classes (e.g., 3) by epoch 99.

accurate segmentation of established classes while continuously discovering emerging urban features, thereby supporting robust land cover discovery mapping. We define known classes as land cover categories with complete supervision during training, and novel classes as land cover categories that require discovery through unsupervised mechanisms, either in experimental settings where labels are withheld from training but retained for evaluation, or in real-world scenarios where new urban features emerge without prior annotations.

5.3.2 Progressive Dual-branch Prototype Learning

Effective novel class discovery requires robust clustering mechanisms that can capture fine-grained distinctions among unlabeled urban features while forming stable, interpretable clusters. Early training phases benefit from fine-grained over-segmentation to explore potential class boundaries, while later phases require consolidation to form coherent semantic groups. We address this requirement through a dual-branch prototype learning mechanism with progressive capacity scheduling. Our architecture processes the unified feature rep-

resentations $F = \{\mathbf{f}_i\}_{i=1}^{N}$ extracted by MinkUNet34C through two specialized branches employing distinct prototype strategies. Figure 5.10 illustrates the complete workflow.

The dual-branch unlabeled module comprises two separate sets of learnable prototype parameters. For the primary unsupervised branch, we maintain a fixed set of prototypes $\{\omega_k^{(P)}\}_{k=1}^{C_u}$, and for the over-segmentation branch, we use a dynamic set of prototypes $\{\omega_k^{(O)}\}_{k=1}^{\rho(t)}$. Here, C_u denotes the fixed number of prototypes in the primary branch corresponding to the target number of novel classes, $\rho(t)$ represents the progressive partition dimension indicating a time-varying integer for the current number of prototypes in the over-segmentation branch at epoch t, and $\omega_k^{(P)}, \omega_k^{(O)} \in \mathbb{R}^{\delta}$ are the learnable prototype parameters for the corresponding branches.

The primary unsupervised branch maintains a fixed number C_u of prototypes throughout training, providing consistent clustering capacity and ensuring stable convergence to reliable cluster centers. For each point p_i with feature $\mathbf{f}_i$, this branch computes raw matching scores $\gamma_{i,k}^{P(t)}$ via dot products $\mathbf{f}_i^{\top}\omega_k^{(P)}$, where $\gamma_{i,k}^{P(t)}$ represents the raw logit at epoch t for assigning point p_i to the k-th prototype in the primary branch. The aggregated raw logits from the primary branch are denoted as $O^P \in \mathbb{R}^{N\times C_u}$, where N represents the total number of points.

The over-segmentation branch employs a dynamic number of prototypes governed by the progressive partition dimension $\rho(t)$. This branch enables adaptive capacity that matches the discovery needs at different training stages. For each point p_i with feature $\mathbf{f}_i$, this branch computes raw matching scores $\gamma_{i,k}^{O(t)}$ via dot products $\mathbf{f}_i^{\top}\omega_k^{(O)}$, where $\gamma_{i,k}^{O(t)}$ represents the raw logit at epoch t for assigning point p_i to the k-th prototype in the over-segmentation branch. The aggregated raw logits from the over-segmentation branch are denoted as $O^O \in \mathbb{R}^{N\times\rho(t)}$.

The progressive partition dimension $\rho(t)$ follows a carefully designed two-phase evolution as illustrated in Figure 5.11. During the warm-up phase for epochs $t \in \{0, 1, \ldots, T_{\text{warm}} - 1\}$, $\rho(t)$ increases gradually:

$$\rho(t) = \text{round}\left(\rho_{\max} \cdot \frac{t+1}{T_{\text{warm}}}\right), \tag{5.1}$$

where $\rho_{\max}$ represents the maximum number of prototypes, T_{warm} denotes the number of warm-up epochs, and the rounding operation ensures integer prototype counts.

Subsequently, a cosine decay schedule is immediately applied, commencing at epoch T_{warm} and continuing until the final epoch:

$$\rho(t) = \text{round}\Big(C_u + \tfrac{1}{2}(\rho_{\max} - C_u)\times\big[\,1 + \cos\big(\pi\,\tfrac{t-T_{\text{warm}}}{T_{\text{total}}-T_{\text{warm}}}\big)\big]\Big), \tag{5.2}$$

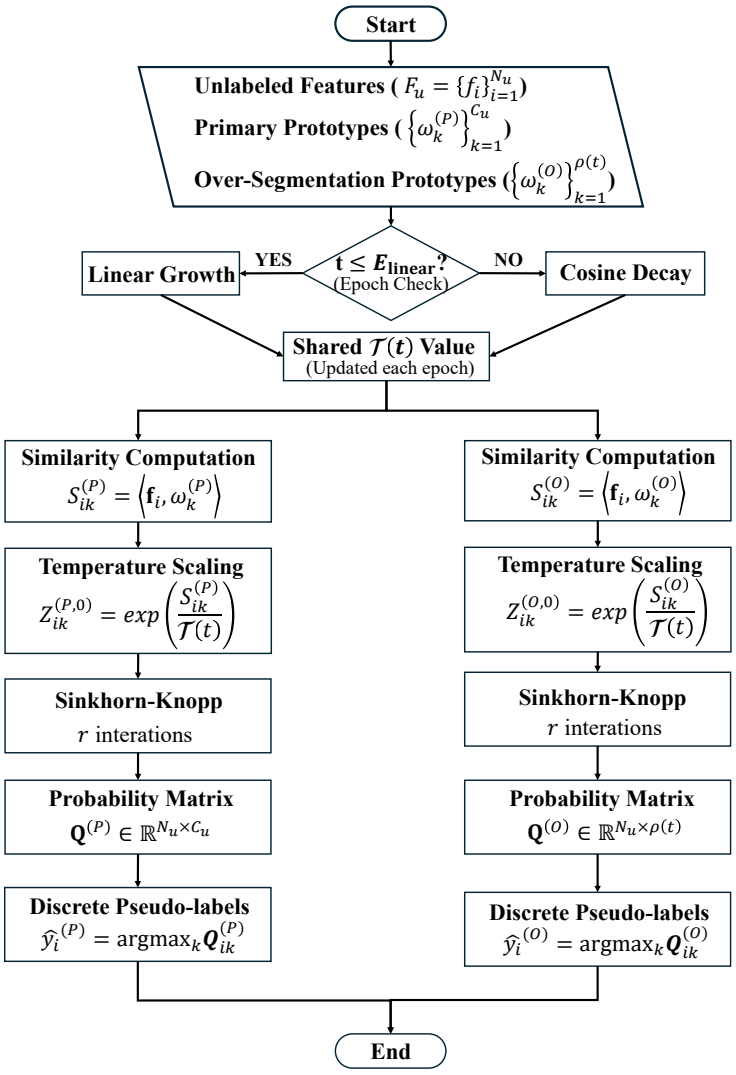

Figure 5.12: Temporal temperature scheduling mechanism for robust Sinkhorn-Knopp pseudo-label generation. The framework employs epoch-dependent temperature control with two phases: linear warm-up ($t \leq E_{\text{linear}}$) that gradually increases temperature from $\mathcal{T}_{\text{start}}$ to $\mathcal{T}_{\text{peak}}$ to encourage exploration, followed by cosine decay that progressively reduces temperature to $\mathcal{T}_{\text{final}}$ for sharp cluster assignments. The shared temperature parameter $\mathcal{T}(t)$ ensures consistent pseudo-label quality across both primary and over-segmentation branches through coordinated similarity scaling and iterative Sinkhorn-Knopp balancing.

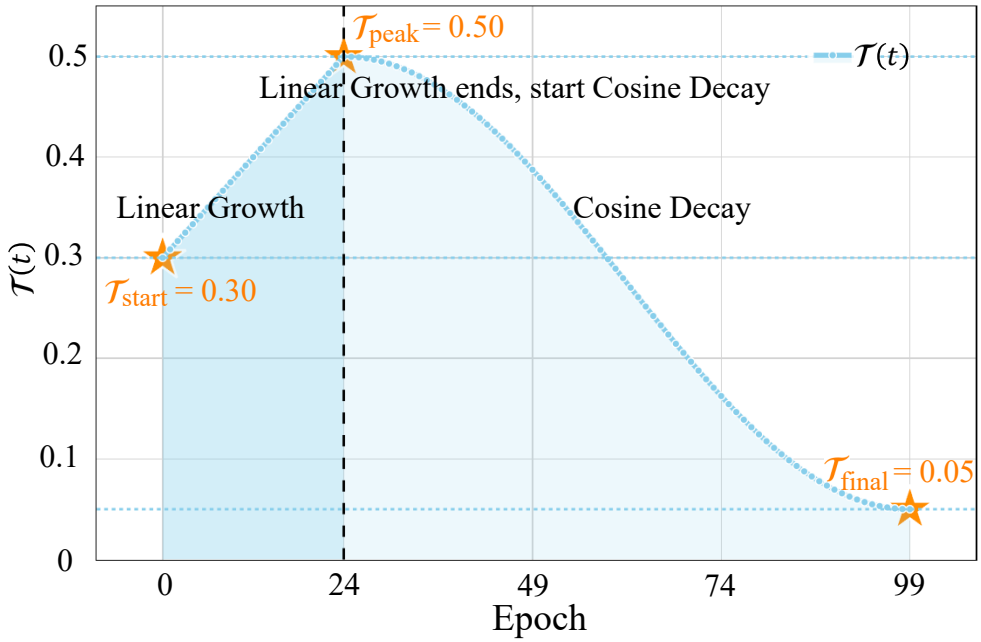

Figure 5.13: Temporal schedule of the temperature parameter $\mathcal{T}(t)$ used in the Sinkhorn–Knopp balancing module. In the initial phase (epochs 0–24), $\mathcal{T}(t)$ increases linearly from $\mathcal{T}_{\text{start}} = 0.30$ to $\mathcal{T}_{\text{peak}} = 0.50$, ensuring smooth pseudo-label assignments during early training. In the subsequent cosine decay phase (epochs 25–99), $\mathcal{T}(t)$ gradually decreases to $\mathcal{T}_{\text{final}} = 0.05$, thereby sharpening the assignment distributions as the feature representations stabilize. This adaptive schedule prevents premature overconfidence in the pseudo-labels and enhances the robustness of novel class discovery.

where T_{total} represents the total number of training epochs and $t \geq T_{\text{warm}}$. By the final epoch, the over-segmentation branch operates with a prototype capacity of C_u.

This dual-branch approach enables the network to initially over-segment the unlabeled data through a high value of $\rho(t)$, thereby capturing fine-grained and transient structures, and subsequently consolidate these into a stable and interpretable partition through the progressive scheduling mechanism, enhancing overall segmentation performance in large-scale remote sensing applications.

5.3.3 Temporal Sinkhorn-Knopp Balancing

Reliable pseudo-label generation is crucial for effective novel class discovery in unlabeled data. Our framework employs the Sinkhorn-Knopp algorithm to transform raw similarity

scores into balanced probability distributions, with a key innovation being the introduction of temporal temperature scheduling that adapts the assignment sharpening behavior throughout training. This adaptive mechanism ensures robust pseudo-label quality across different learning phases, and the complete workflow is illustrated in Figure 5.12.

Given the complete feature set $F = \{\mathbf{f}_i\}_{i=1}^{N}$ extracted by the MinkUNet34C, we select only the unlabeled features $F_u = \{\mathbf{f}_i : i \in \mathcal{I}_u\}$ based on the label mask, where $\mathcal{I}_u$ denotes the indices of unlabeled points, $N_u = |\mathcal{I}_u|$ represents the number of unlabeled points, and δ denotes the feature dimension. These unlabeled features are processed together with prototype sets from both branches (primary: $\{\omega_k^{(P)}\}_{k=1}^{C_u}$, over-segmentation: $\{\omega_k^{(O)}\}_{k=1}^{\rho(t)}$) to compute similarity matrices for each branch using inner products between unlabeled features and prototypes. For the primary branch, the similarity matrix $\mathbf{S}^{(P)} \in \mathbb{R}^{N_u \times C_u}$ is computed as:

$$S_{ik}^{(P)} = \langle \mathbf{f}_i, \omega_k^{(P)} \rangle, \tag{5.3}$$

where $S_{ik}^{(P)}$ represents the similarity between the i-th unlabeled feature $\mathbf{f}_i \in \mathbb{R}^{\delta}$ and the k-th primary prototype $\omega_k^{(P)} \in \mathbb{R}^{\delta}$, $\langle \cdot, \cdot \rangle$ denotes the inner product operation, $i = 1, \ldots, N_u$, and $k = 1, \ldots, C_u$. Similarly, for the over-segmentation branch, the similarity matrix $\mathbf{S}^{(O)} \in \mathbb{R}^{N_u \times \rho(t)}$ is computed using the same operation between features $\mathbf{f}_i$ and over-segmentation prototypes $\omega_k^{(O)} \in \mathbb{R}^{\delta}$, where $k = 1, \ldots, \rho(t)$.

The temperature parameter $\mathcal{T}(t)$ follows a two-phase piecewise schedule shown in Figure 5.13. During early training epochs, linear warm-up encourages exploration:

$$\mathcal{T}(t) = \mathcal{T}_{\text{start}} + (\mathcal{T}_{\text{peak}} - \mathcal{T}_{\text{start}}) \frac{t}{E_{\text{linear}}}, \quad 0 \leq t < E_{\text{linear}} \tag{5.4}$$

where $\mathcal{T}(t)$ is the temperature parameter at epoch t, $\mathcal{T}_{\text{start}}$ is the initial temperature value, $\mathcal{T}_{\text{peak}}$ is the peak temperature value, $E_{\text{linear}} = \lfloor T_{\text{total}}/4 \rfloor$ represents the linear growth phase duration, and t is the current epoch index. For subsequent epochs, cosine decay gradually sharpens assignments:

$$\mathcal{T}(t) = \mathcal{T}_{\text{final}} + \tfrac{1}{2}\big[\mathcal{T}_{\text{peak}} - \mathcal{T}_{\text{final}}\big]\big(1 + \cos\big(\pi \tfrac{t - E_{\text{linear}}}{T_{\text{total}} - E_{\text{linear}}}\big)\big), \quad E_{\text{linear}} \leq t < T_{\text{total}}, \tag{5.5}$$

where $\mathcal{T}_{\text{final}}$ is the final temperature value and T_{total} is the total number of training epochs. In our implementation, we use default values $\mathcal{T}_{\text{start}} = 0.30$, $\mathcal{T}_{\text{peak}} = 0.50$, and $\mathcal{T}_{\text{final}} = 0.05$.

The Sinkhorn-Knopp normalization begins by exponentially scaling the similarity matrices using the shared temperature parameter. For the primary branch:

$$Z_{ik}^{(P,0)} = \exp\left(\frac{S_{ik}^{(P)}}{\mathcal{T}(t)}\right), \tag{5.6}$$

where $Z_{ik}^{(P,0)}$ is the initial exponentially scaled similarity value for the primary branch. For the over-segmentation branch, $Z_{ik}^{(O,0)}$ is computed analogously using $S_{ik}^{(O)}$ and the same temperature parameter.

Subsequently, r iterations of alternating row and column normalization are performed independently for each branch, ensuring that the final matrices are doubly stochastic with balanced prototype assignments. This iterative process transforms the initial exponentially scaled matrices $\mathbf{Z}^{(P,0)}$ and $\mathbf{Z}^{(O,0)}$ into the final probability matrices through repeated normalization steps, where each iteration normalizes rows (ensuring each point's assignment probabilities sum to 1) followed by column normalization (ensuring balanced load across prototypes). After r iterations, the resulting matrices $\mathbf{Q}^{(P)} \in \mathbb{R}^{N_u \times C_u}$ and $\mathbf{Q}^{(O)} \in \mathbb{R}^{N_u \times \rho(t)}$ are doubly stochastic and provide balanced pseudolabel distributions. Discrete pseudo-labels are obtained by selecting the maximum probability prototype:

$$\hat{y}_i^{(P)} = \arg\max_k Q_{ik}^{(P)}, \tag{5.7}$$

$$\hat{y}_i^{(O)} = \arg\max_k Q_{ik}^{(O)}, \tag{5.8}$$

where $\hat{y}_i^{(P)}$ and $\hat{y}_i^{(O)}$ are the discrete pseudo-labels for the i-th unlabeled point ($i = 1, \ldots, N_u$) in each branch, $Q_{ik}^{(P)}$ and $Q_{ik}^{(O)}$ are the probability values, and $\arg\max_k$ selects the prototype index with highest probability.

This adaptive scheduling strategy ensures smooth distributions during early training to prevent premature overconfident assignments, while later phases employ sharper distributions to capture well-differentiated class boundaries. The module generates pseudolabel matrices for both branches that serve as surrogate ground truth for subsequent loss computations.

5.3.4 Dynamic Weighting Loss

Urban point clouds exhibit severe class imbalance, where dominant categories like ground surfaces may contain millions of points while rare but environmentally significant features

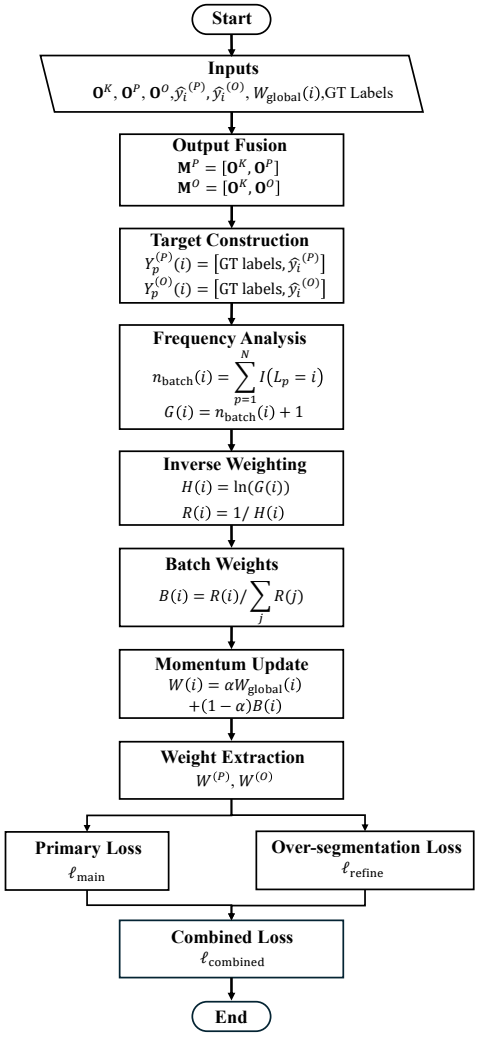

Figure 5.14: Dynamic weighting strategy for addressing class imbalance in urban point cloud analysis. The system combines supervised logits ($\mathbf{O}^K$) with unsupervised outputs ($\mathbf{O}^P$, $\mathbf{O}^O$) to form unified prediction matrices $\mathbf{M}^P$ and $\mathbf{M}^O$. Target distributions incorporate ground truth labels and Sinkhorn-Knopp pseudo-labels ($\mathbf{Q}^P$, $\mathbf{Q}^O$). Through frequency analysis, inverse weighting, and momentum-based updates ($\alpha = 0.8$), the algorithm generates adaptive class weights that emphasize underrepresented categories while maintaining training stability, ultimately producing balanced weighted cross-entropy losses for effective novel class discovery.

are represented by only sparse samples. To address this challenge, we propose a dynamic weighting strategy that adaptively balances local batch variations with long-term distributional stability, as illustrated in Figure 5.14.

Our approach operates on a unified prediction space encompassing both supervised and unsupervised branches. At each training iteration, we process N points with predictions organized into two combined output matrices. For the primary branch, we construct $\mathbf{M}^P \in \mathbb{R}^{N\times(C_k+C_u)}$ by concatenating known class logits $\mathbf{O}^K \in \mathbb{R}^{N\times C_k}$ from the supervised head with primary novel class logits $\mathbf{O}^P \in \mathbb{R}^{N\times C_u}$ from the primary unsupervised head. Similarly, for the over-segmentation branch, we construct $\mathbf{M}^O \in \mathbb{R}^{N\times(C_k+\rho(t))}$ by concatenating the same known class logits $\mathbf{O}^K$ with over-segmentation logits $\mathbf{O}^O \in \mathbb{R}^{N\times\rho(t)}$ from the over-segmentation head. Here, C_k represents the number of labeled classes, C_u denotes the number of unlabeled classes, and $\rho(t)$ indicates the dynamic over-segmentation prototype count at epoch t.

For each point p in the current batch, we determine its target label L_p, where labeled points use ground truth class indices and unlabeled points use pseudo-labels from the Sinkhorn-Knopp algorithm. The weighting mechanism begins by analyzing the distribution of these target assignments within each batch through a multi-step transformation process. We first compute occurrence frequencies $n_{\text{batch}}(i)$ for each prediction index i by counting across all N points in the batch using the indicator function $I(\cdot)$. To handle zero frequencies and amplify rare classes, we apply frequency regularization by computing $G(i)$, which adds one to each raw frequency count. This is followed by logarithmic transformation $H(i)$ and inverse weighting $R(i)$, where the reciprocal operation ensures that rarely occurring categories receive higher weights while moderating frequent ones. Finally, batch weights $B(i)$ are computed by normalizing the inverse values $R(i)$ over all indices i within the current branch, ensuring that the weights sum to unity and maintain proportional emphasis on underrepresented classes.

To ensure training stability across iterations, these batch-level statistics are integrated with historical information through momentum updates:

$$W(i) = \alpha W_{\text{global}}(i) + (1-\alpha)B(i), \tag{5.9}$$

where $\alpha = 0.8$ provides the momentum factor, $W_{\text{global}}(i)$ represents global weights from the previous iteration, and $W(i)$ denotes the updated global weights after smoothing. This exponential moving average prevents erratic weight changes and maintains long-term stability while adapting to current batch characteristics.

Since the dual unsupervised branches operate with different prototype dimensions, we extract branch-specific weights from the unified weight vector. The primary branch re-

ceives weights $W^{(P)}$ compatible with its combined prediction output $\mathbf{M}^P$, while the over-segmentation branch uses weights $W^{(O)}$ aligned with its output $\mathbf{M}^O$. These weights are applied during loss computation to ensure dimensional consistency between weight vectors and prediction matrices.

The network predictions $\mathbf{M}^P$ and $\mathbf{M}^O$ are converted to probability distributions $\hat{m}_p^P(i)$ and $\hat{m}_p^O(i)$ through softmax normalization. These predictions are compared against branch-specific target distributions $Y_p^{(P)}(i)$ and $Y_p^{(O)}(i)$, which combine one-hot encoded true labels for known classes with one-hot encoded discrete pseudo-labels $\hat{y}_i^{(P)}$ and $\hat{y}_i^{(O)}$ for novel classes. The weighted cross-entropy losses incorporate the dynamic weights to ensure balanced gradient emphasis across all prediction categories:

$$\ell_{\text{main}} = -\sum_{p=1}^{N} \sum_{i=0}^{C_k+C_u-1} Y_p^{(P)}(i) W^{(P)}(i) \ln(\hat{m}_p^P(i)), \tag{5.10}$$

$$\ell_{\text{refine}} = -\sum_{p=1}^{N} \sum_{i=0}^{C_k+\rho(t)-1} Y_p^{(O)}(i) W^{(O)}(i) \ln(\hat{m}_p^O(i)), \tag{5.11}$$

where ℓ_{main} corresponds to the primary branch loss and ℓ_{refine} represents the over-segmentation branch loss. The final training objective ℓ_{combined} combines both losses equally by averaging them to maintain balanced learning across both prototype learning strategies. This dynamic weighting mechanism ensures that underrepresented urban features receive enhanced gradient emphasis proportional to their rarity, achieving robust and balanced learning across all categories essential for effective land cover discovery mapping.

5.4 Experiments

5.4.1 Experimental Configuration and Train-test Partitioning

The proposed framework employs sparse voxelization with a voxel size of 0.2 and down-samples to 40,000 points per point cloud during training for computational efficiency. We use a batch size of 4 with 8 worker threads for data loading. For DALES, we strictly follow the official train-test split provided by the dataset authors. For H3D, we apply a temporal split by using the point clouds acquired in March 2016, March 2018, and March 2019 for training, and those acquired in November 2018 for testing. We utilize the complete training and validation data from each temporal period. This temporal separation strategy

serves the dual purpose of ensuring temporal independence for robust model evaluation while assessing the framework's capability to discover novel classes across different temporal contexts within the same geographic area. Specifically, November 2018 was selected as the test epoch because it provides both seasonal diversity and temporal diversity, enabling comprehensive evaluation of novel class discovery robustness across varying seasonal and multi-year acquisition conditions.

5.4.2 Experimental Protocol and Real-world Equivalence

A key concern in novel class discovery evaluation is whe-ther experimental settings truly reflect real-world deployment scenarios. Our experimental protocol addresses this challenge by creating conditions that are procedurally equivalent to encountering genuinely unseen urban features in practice.

In our experiments, certain classes are designated as "novel" by excluding their labels from the training supervision while retaining their point cloud data in the dataset. The model must discover these classes through unsupervised clustering mechanisms operating on feature representations learned from all point cloud data. This process mirrors real-world scenarios where new urban infrastructure appears: the model encounters point cloud data from novel urban classes and must identify them through unsupervised clustering.

Ground truth labels for hidden classes only enable quantitative evaluation of discovery accuracy and do not participate in the training or inference processes. These labels remain completely isolated from the model during all learning and discovery phases, serving purely as evaluation benchmarks. The model's clustering decisions are made entirely through unsupervised mechanisms applied to learned feature representations, ensuring authentic replication of real-world discovery conditions.

The model's discovery mechanism operates under identical conditions in both experimental and real-world scenarios. Whether discovering a "hidden" power line in our DALES experiment or identifying a newly emerged urban feature in operational deployment, the model relies on the same feature learning and unsupervised clustering processes. In practical deployment scenarios, the method processes identical inputs and generates identical outputs as in our experimental setup. The input consists of point clouds with labeled known classes and unlabeled novel classes, while the output provides semantic predictions for both known and novel categories. The dual-branch architecture processes the mixed labeled-unlabeled input and generates the same type of semantic segmentation outputs through identical feature learning, prototype clustering, and Sinkhorn-Knopp normalization processes. Whether applied to experimental datasets or real-world urban monitoring

Table 5.1: Land cover discovery mapping results on DALES dataset (%). Pink-highlighted values indicate the classes designated as novel in each split. The "Novel" column reports the mIoU for the newly discovered classes, "Base" corresponds to the mIoU for the predefined (known) classes, and "All" denotes the overall mIoU across all classes.

Split	Model	Ground	Vegetation	Car	Pole	Building	Truck	Fence	Power Line	Novel	Base	All
DALES-2^0	NOPS	64.90	86.55	58.88	0.00	93.16	**27.80**	25.97	56.59	32.45	58.16	51.73
	CHNCD	54.61	81.33	**71.16**	0.00	91.93	24.04	48.42	62.31	27.31	63.20	54.22
	Ours	**85.66**	**86.65**	70.78	**0.49**	**94.99**	20.72	**51.70**	**70.33**	**43.08**	**65.86**	**60.17**
DALES-2^1	NOPS	95.08	71.42	65.42	35.60	91.88	2.36	40.57	62.72	36.89	65.21	58.13
	CHNCD	94.79	52.79	68.67	59.96	91.28	**5.13**	47.35	68.89	28.96	71.82	61.11
	Ours	**95.65**	**86.48**	**74.78**	**66.58**	**94.60**	3.08	**54.83**	**75.40**	**44.78**	**76.97**	**68.93**
DALES-2^2	NOPS	95.18	88.14	69.76	35.78	58.79	24.83	37.50	0.13	29.46	58.53	51.26
	CHNCD	94.74	82.77	67.70	44.04	**78.94**	30.04	44.95	0.00	39.47	60.71	55.40
	Ours	**95.53**	**91.10**	**71.92**	**61.16**	76.30	**30.15**	**49.72**	**12.46**	**44.38**	**66.60**	**61.04**
DALES-2^3	NOPS	95.17	90.44	41.48	49.63	93.80	**26.07**	42.32	76.99	41.90	72.02	64.49
	CHNCD	94.99	86.62	53.71	68.58	92.03	22.70	46.24	70.86	49.98	72.63	66.96
	Ours	**95.66**	**91.61**	**69.54**	**73.09**	**94.70**	21.59	**48.15**	**79.25**	**58.85**	**75.98**	**71.70**

data, the model produces the same format of predictions: class labels for known categories and discovered novel classes.

The fundamental input-output paradigm remains unchanged between controlled evaluation and operational deployment, demonstrating that our experimental framework directly represents real-world application conditions without requiring any methodological modifications. The only difference lies in evaluation approaches: experiments use pre-existing ground truth for quantitative assessment, while real-world scenarios may utilize available ground truth or human verification depending on application requirements.

5.4.3 Evaluation and Visualization Results on DALES

Table 5.1 presents the land cover discovery mapping performance on the DALES dataset across four experimental configurations. The proposed 3DLCDM framework demonstrates consistent improvements over both baseline methods [71, 18] across all evaluation metrics. The novel class mIoU increases from 27.31%-49.98% (baseline methods) to 43.08%-58.85% across all splits, while base class performance advances from 58.16%-72.63% to 65.86%-76.97%. Overall mIoU improves from 51.26%-66.96% to 60.17%-71.70%, confirming the framework's effectiveness for adaptive land cover mapping in dynamic urban environments.

In the DALES-2^0 split, ground and pole constitute the novel classes. The framework achieves substantial improvements in ground segmentation, advancing from 64.90%

(NOPS) and 54.61% (CHNCD) to 85.66%. This enhancement is critical for land cover discovery mapping as accurate ground delineation forms the foundation for distinguishing impervious surfaces from permeable areas, directly affecting urban hydrology modeling and stormwater management planning. Among base classes, power line segmentation exhibits notable improvement from 56.59% and 62.31% to 70.33%, while fence mapping advances from 25.97% and 48.42% to 51.70%. These infrastructure elements are essential for comprehensive urban morphology characterization in evolving landscapes.

For DALES-2^1 where vegetation and truck represent novel classes, vegetation segmentation shows remarkable advancement from 71.42% (NOPS) and 52.79% (CHNCD) to 86.48%. Enhanced vegetation mapping supports dynamic monitoring of urban green infrastructure evolution, enabling timely assessment of ecosystem service changes as cities adapt to climate challenges. The pole category demonstrates the most significant base class improvement, increasing from 35.60% and 59.96% to 66.58%, followed by car segmentation which advances from 65.42% and 68.67% to 74.78%. These transportation-related improvements facilitate better characterization of mobility infrastructure changes in urban land cover discovery mapping applications.

The DALES-2^2 configuration designates building and po-wer line as novel classes, presenting challenges for segmenting both dominant and linear urban features. Building segmentation achieves 76.30% compared to 58.79% (NOPS), while power line segmentation improves substantially from 0.13% and 0.00% to 12.46%. The enhanced building mapping capability is particularly valuable for monitoring urban densification patterns and architectural evolution in land cover discovery mapping frameworks. Vegetation shows the highest base class improvement, advancing from 88.14% and 82.77% to 91.10%, contributing to more comprehensive characterization of green space distribution within built environments.

In DALES-2^3, car and fence novel classes represent fine-scale urban elements critical for detailed land use analysis. Car segmentation advances significantly from 41.48% (NOPS) and 53.71% (CHNCD) to 69.54%, enabling better characterization of parking areas and transportation zones within the urban fabric. The pole base class exhibits exceptional improvement from 49.63% and 68.58% to 73.09%, while power line mapping increases from 76.99% and 70.86% to 79.25%. The substantial improvements in both dynamic elements (cars) and linear infrastructure (poles, power lines) indicate the framework's effectiveness for land cover discovery mapping in complex urban environments where multiple infrastructure types coexist and evolve simultaneously.

Figures 5.15–5.18 present qualitative comparisons between the NOPS method and the proposed 3DLCDM framework under different experimental splits that simulate the challenge of novel class discovery in dynamic land cover mapping. For example, in the DALES-2^0 split (Figure 5.15), the novel classes are defined as ground and pole. The NOPS method



Figure 5.15: Visualization of land cover discovery mapping results on the DALES-2^0 split, where ground and pole are defined as the novel classes. The figure shows two representative urban scenes with zoom-in views: (a,d) ground truth annotations, (b,e) NOPS segmentation results, and (c,f) our 3DLCDM framework outputs.

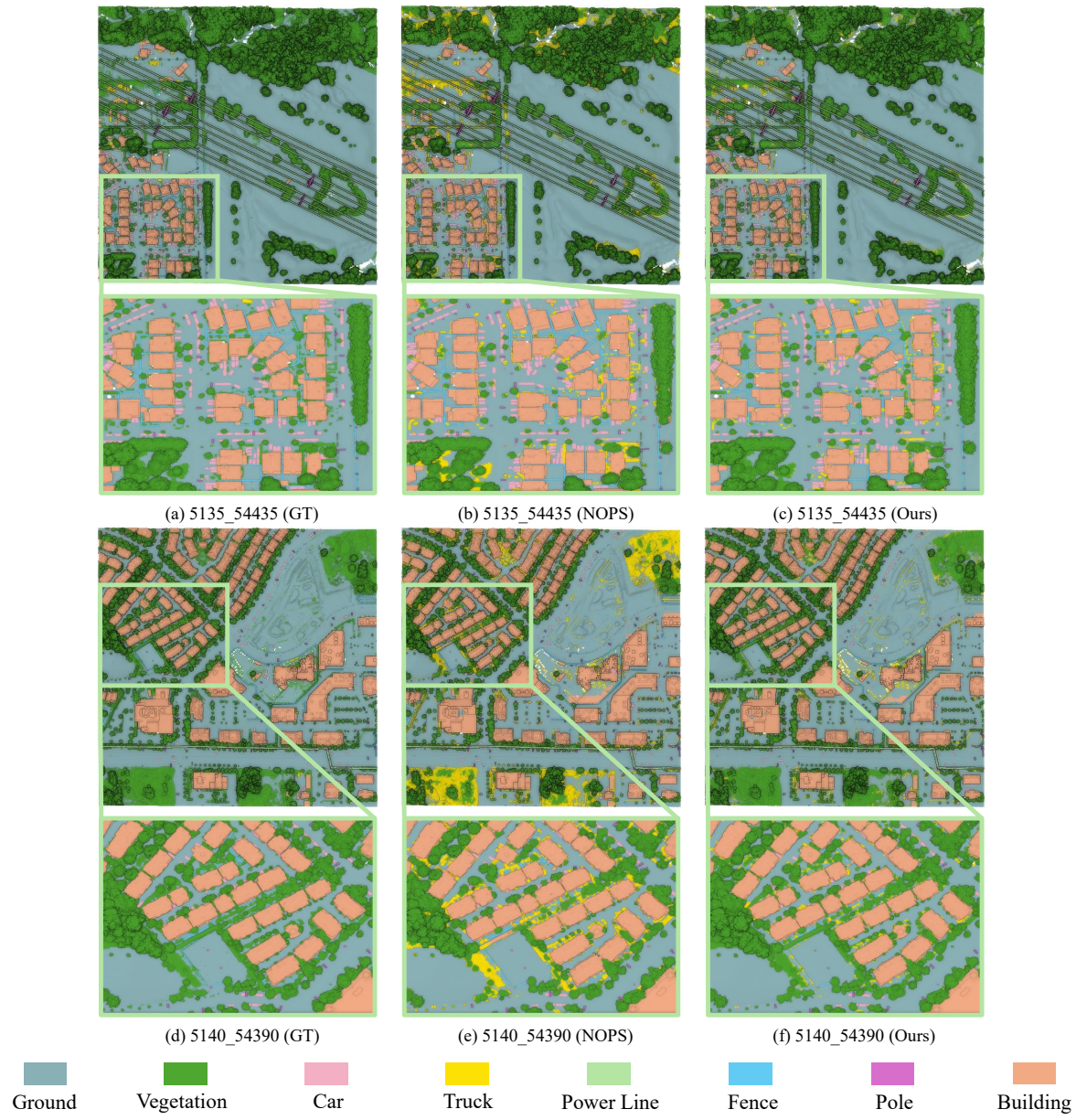

Figure 5.16: Visualization of land cover discovery mapping results on the DALES-2[1] split, where vegetation and truck are defined as the novel classes. The figure shows two representative urban scenes with zoom-in views: (a,d) ground truth annotations, (b,e) NOPS segmentation results, and (c,f) our 3DLCDM framework outputs.



Figure 5.17: Visualization of land cover discovery mapping results on the DALES-2^2 split, where building and power line are defined as the novel classes. The figure shows two representative urban scenes with zoom-in views: (a,d) ground truth annotations, (b,e) NOPS segmentation results, and (c,f) our 3DLCDM framework outputs.

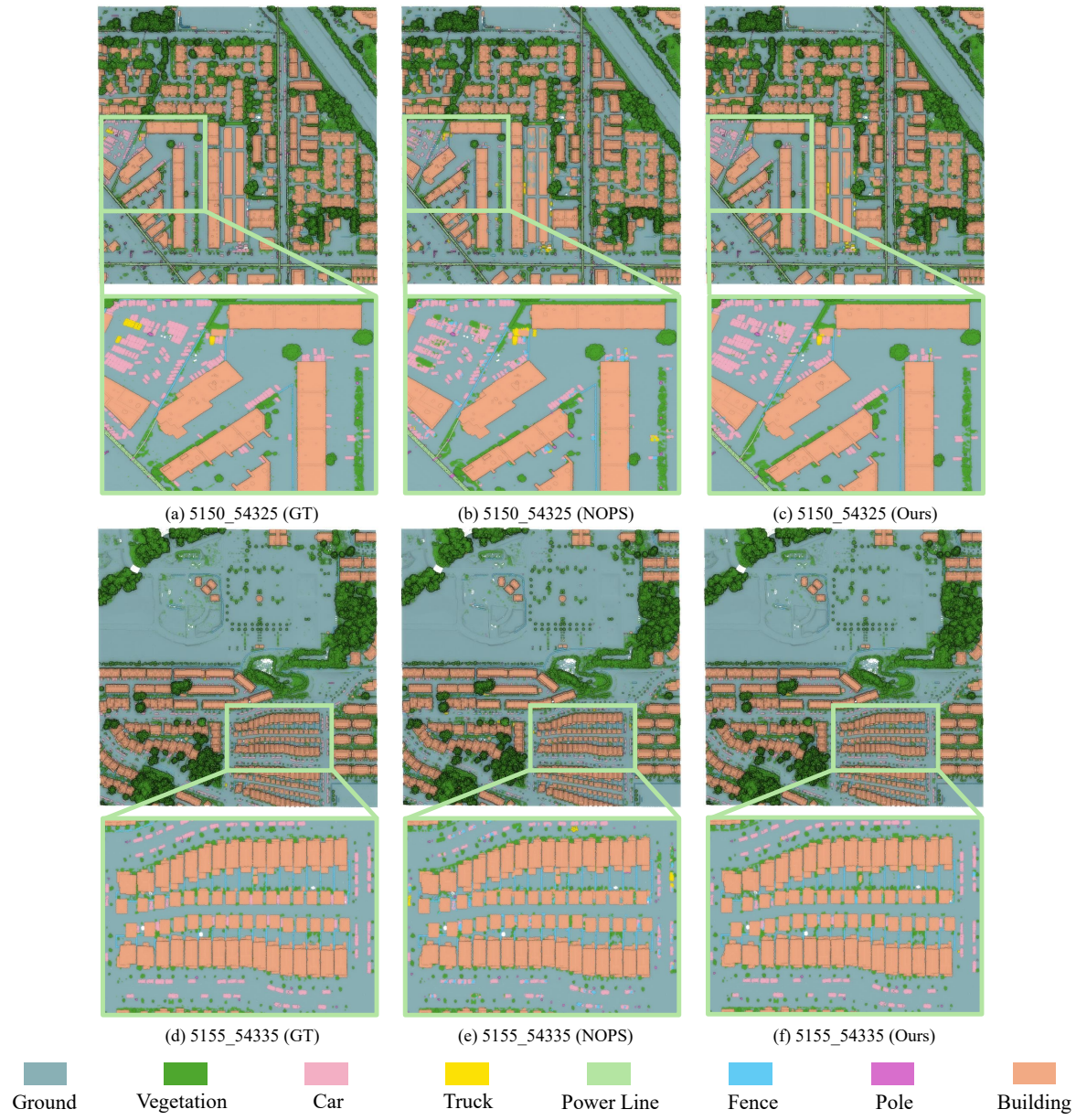

Figure 5.18: Visualization of land cover discovery mapping results on the DALES-2[3] split, where car and fence are defined as the novel classes. The figure shows two representative urban scenes with zoom-in views: (a,d) ground truth annotations, (b,e) NOPS segmentation results, and (c,f) our 3DLCDM framework outputs.

tends to confuse these two classes, frequently mis-segmenting a substantial portion of ground as pole. In contrast, the 3DLCDM framework accurately segments most of the ground. However, pole segmentation remains challenging for both methods, with extremely low IoU values achieved across all approaches. As observed in Figure 5.15, both methods struggle to accurately identify pole structures, frequently mis-segmenting them as vegetation or power line categories. This confusion stems from the minimal number of pole points in the dataset, making it difficult to learn their distinctive features during novel class discovery. Additionally, poles and power lines are spatially connected in many regions, further complicating their accurate segmentation.

Similarly, in the DALES-2^1 split (Figure 5.16), where vegetation and truck are treated as novel classes, NOPS often misassigns parts of vegetation to the truck category. The proposed method, however, clearly distinguishes and captures vegetation, which is crucial for monitoring urban green spaces and assessing ecological health. However, truck segmentation remains challenging for both methods. As observed in Figure 5.16, both methods struggle to accurately segment truck structures, frequently confusing them with roof or vegetation categories. The truck-vegetation confusion occurs because the unsupervised clustering mechanism has difficulty learning distinct features to separate these categories during novel class discovery. Additionally, trucks share geometric similarities with roof structures due to their planar surface characteristics, and their sparse and scattered placement throughout the scene further complicates the formation of coherent prototype representations.

In the DALES-2^2 split (Figure 5.17), with building and power line as the novel classes, NOPS exhibits confusion by misclassifying some building regions as power line and incorrectly assigning portions of power line to the fence category. The 3DLCDM approach achieves sharper separation between these urban infrastructural elements, thereby providing more reliable mapping of critical structures. Finally, in the DALES-2^3 split (Figure 5.18), where car and fence are defined as novel classes, the proposed framework demonstrates superior performance in fine-scale feature delineation. NOPS frequently misclassifies car structures as fence categories and exhibits boundary blurring for these fine-scale elements, whereas the 3DLCDM framework maintains precise edge detail preservation and accurate class separation. These visual improvements translate directly to enhanced land cover map accuracy for dynamic urban monitoring applications, where precise class boundaries are essential for environmental impact assessment and sustainable planning initiatives.

Table 5.2: Land cover discovery mapping results on H3D dataset (%). Pink-highlighted values indicate the classes designated as novel in each split. The "Novel" column reports the mIoU for the newly discovered classes, "Base" corresponds to the mIoU for the predefined (known) classes, and "All" denotes the overall mIoU across all classes.

Split	Model	Low Vegetation	Impervious Surface	Vehicle	Urban Furniture	Shrub	Tree	Vertical Surface	Roof	Façade	Chimney	Soil	**Novel**	**Base**	**All**
H3D-3^0	NOPS	21.62	28.37	76.71	53.72	60.64	91.93	52.24	92.20	80.66	**17.74**	25.55	22.58	66.71	54.67
	CHNCD	38.87	45.66	77.76	59.61	65.84	93.10	51.84	92.57	81.44	4.60	23.99	29.71	68.27	57.75
	Ours	**68.03**	**62.03**	**78.04**	**59.76**	**69.76**	**94.02**	**56.46**	**93.09**	**83.00**	0.01	**26.40**	**43.36**	**70.07**	**62.78**
H3D-3^1	NOPS	73.32	75.18	1.37	55.64	63.72	60.02	53.10	42.44	74.79	36.39	25.57	34.61	57.21	51.05
	CHNCD	**78.13**	**78.63**	6.92	**61.85**	**68.50**	62.41	57.71	25.26	**81.59**	49.86	**28.61**	31.53	63.11	54.50
	Ours	77.83	77.83	**10.28**	59.64	67.12	**79.32**	**58.07**	**78.27**	80.65	**59.24**	24.92	**55.96**	**63.16**	**61.20**
H3D-3^2	NOPS	74.87	75.69	79.58	57.76	36.98	92.15	5.12	91.15	51.67	51.41	**28.00**	31.26	68.83	58.58
	CHNCD	77.69	76.64	**79.71**	60.08	48.28	93.06	**22.08**	91.21	29.05	63.52	25.61	33.14	70.94	60.63
	Ours	**79.01**	**80.72**	78.97	**62.03**	**59.80**	**94.48**	8.96	**93.29**	**69.72**	**65.49**	27.59	**46.16**	**72.70**	**65.46**
H3D-2^3	NOPS	75.30	76.04	80.57	39.96	66.86	92.50	53.24	91.70	81.05	50.68	12.81	26.39	74.22	65.52
	CHNCD	78.00	77.37	79.99	33.80	69.23	93.45	55.03	93.09	81.92	64.11	11.93	22.87	76.91	67.08
	Ours	**80.90**	**81.52**	**81.09**	**52.90**	**73.11**	**95.32**	**60.91**	**94.16**	**84.15**	**68.11**	**18.66**	**35.78**	**79.92**	**71.89**

5.4.4 Evaluation and Visualization Results on H3D

Table 5.2 presents the quantitative evaluation results on the H3D dataset across four experimental configurations. The proposed 3DLCDM framework demonstrates consistent improvements over both baseline methods, with novel class mIoU increasing from 22.58%-34.61% (NOPS) and 22.87%-33.14% (CHNCD) to 35.78%-55.96% across all splits. Base class mIoU improves from 57.21%-74.22% (NOPS) and 63.11%-76.91% (CHNCD) to 63.16%-79.92%, while overall mIoU advances from 51.05%-65.52% (NOPS) and 54.50%-67.08% (CHNCD) to 61.20%-71.89%.

The H3D-3^0 split achieves a novel mIoU of 43.36% compared to 22.58% (NOPS) and 29.71% (CHNCD). Low vegetation shows the most substantial improvement, advancing from 21.62% to 68.03%, demonstrating the framework's effectiveness in discovering ground-level vegetation critical for comprehensive land cover mapping. Base class performance benefits significantly, with shrub segmentation improving from 60.64% and 65.84% to 69.76% and urban furniture advancing from 53.72% to 59.76%, contributing to the base mIoU of 70.07% versus 66.71% (NOPS) and 68.27% (CHNCD). These improvements enable automatic segmentation of emerging vegetation patterns and infrastructure modifications that are essential for monitoring evolving urban green infrastructure in land cover discovery mapping applications.

H3D-3^1 achieves a novel mIoU of 55.96%, substantially exceeding NOPS (34.61%) and CHNCD (31.53%). Roof segmentation exhibits exceptional improvement from 42.44% (NOPS) and 25.26% (CHNCD) to 78.27%, demonstrating effective discovery of building structures essential for land cover change monitoring. Base class chimney segmentation

shows the most notable advancement from 36.39% to 59.24%, while façade segmentation improves from 74.79% to 80.65%, maintaining base mIoU at 63.16%. This capability supports continuous tracking of urban development and architectural changes without requiring manual annotation updates in dynamic land cover mapping systems.

For H3D-3^2, façade segmentation shows remarkable pro-gress from 51.67% (NOPS) and 29.05% (CHNCD) to 69.72%, indicating successful discovery of vertical architectural elements often overlooked in conventional land cover classification. The novel mIoU reaches 46.16%, surpassing both baseline methods. Base class performance achieves 72.70% mIoU, with impervious surface advancing from 75.69% to 80.72% and urban furniture improving from 57.76% to 62.03%. The enhanced recognition of building envelope components facilitates detailed urban morphology analysis and supports land cover discovery mapping that adapts to evolving architectural practices and building design trends.

In H3D-2^3, urban furniture segmentation reaches 52.90% compared to 39.96% (NOPS) and 33.80% (CHNCD), demonstrating capability in discovering fine-scale infrastructure elements. Base class performance excels with a mIoU of 79.92% versus 74.22% (NOPS) and 76.91% (CHNCD), exemplified by vertical surface improving from 53.24% to 60.91% and shrub segmentation advancing from 66.86% to 73.11%. This reflects the framework's ability to incorporate emerging urban amenities and public infrastructure into comprehensive land cover maps, supporting planning applications that require detailed inventory of evolving built environment features.

Figure 5.19 presents visualization results for land cover discovery mapping on the H3D dataset across four experimental splits, demonstrating the framework's performance in complex urban environments with fine-grained semantic categories. In the H3D-3^0 split (Figure 5.19 panels a-c), where low vegetation, impervious surface, and chimney constitute novel classes, NOPS exhibits substantial misclassification by incorrectly assigning impervious surface regions to chimney categories and misclassifying extensive low vegetation areas as impervious surface. Additionally, vehicle boundary regions are incorrectly segmented as low vegetation, resulting in overall boundary degradation. The proposed 3DLCDM framework achieves enhanced class separation and boundary preservation across low vegetation and impervious surface categories. However, a notable failure case emerges in chimney segmentation within this split. Visual analysis reveals that our method misclassifies chimney structures as low vegetation, as evidenced by multiple chimneys on rooftops being incorrectly segmented as low vegetation category in Figure 5.19 panel (c). This failure primarily stems from the extremely small number of chimney points in the dataset and the limited spatial footprint of individual chimney structures, which may challenge the geometric feature preservation at the current voxel resolution. The unsupervised clustering mechanism faces difficulties in forming coherent representations for such rare fine-scale urban elements

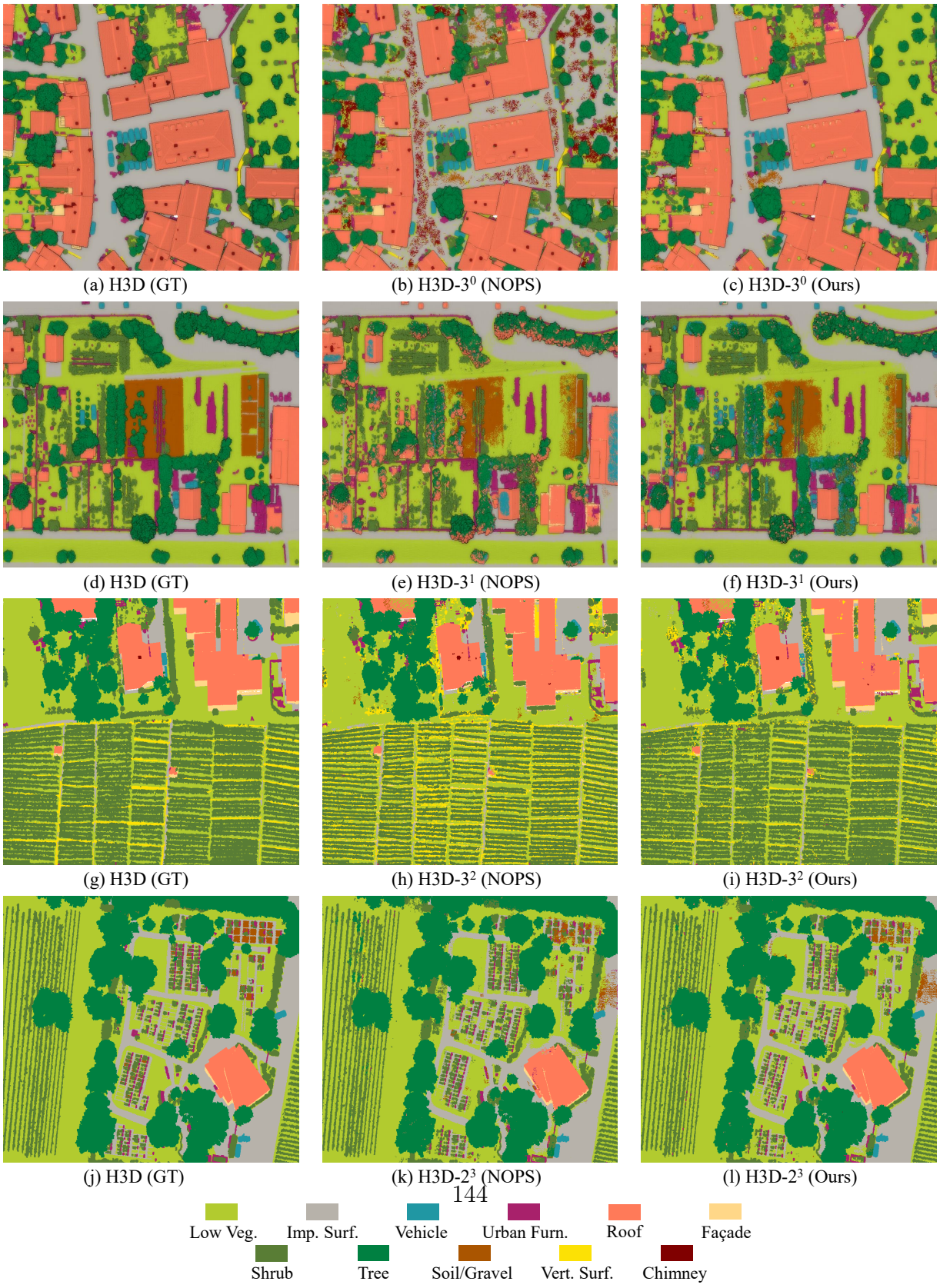

Figure 5.19: Visualization of land cover discovery mapping results on the H3D dataset for four splits.

during the novel class discovery process.

For the H3D-3^1 split (Figure 5.19 panels d-f), where tree, roof, and vehicle represent novel classes, NOPS demonstrates significant confusion among these three categories, with frequent misclassification between tree and roof regions as well as vehicle-roof confusion in multiple scene areas. The 3DLCDM approach provides relatively more precise segmentation of all three novel classes, maintaining distinct feature boundaries. However, vehicle segmentation within this split reveals certain limitations in our approach. While Figure 5.19 panel (f) shows that our method can identify most vehicle instances better than NOPS, it also exhibits confusion by incorrectly segmenting small portions of tree and roof structures as vehicles, and conversely misclassifying some vehicle regions as urban furniture. This cross-category confusion stems from the geometric similarity between vehicle tops and roof surfaces, as well as the spatial proximity between vehicles and urban furniture elements, which challenges the unsupervised clustering mechanism in distinguishing between these compact urban objects when they must be discovered simultaneously.

In the H3D-3^2 configuration (Figure 5.19 panels g-i), both methods encounter challenges with shrub and vertical surface novel classes. The confusion between shrub and vertical surface stems from both geometric and spatial factors. Geometrically, both categories exhibit similar vertical height profiles in 3D point cloud data. Additionally, as observed in Figure 5.19 panels g-i, shrub and vertical surface elements are spatially adjacent in the lower regions of the scene, further complicating their distinction. Both methods exhibit mutual confusion between these two categories due to these combined challenges. However, our 3DLCDM framework achieves better boundary delineation and detail preservation, attributed to progressive prototype refinement and the temporal Sinkhorn-Knopp normalization that maintains finer clustering granularity during the discovery process. Notably, the 3DLCDM method accurately identifies façade novel class regions, while NOPS frequently misclassifies façade areas as shrub categories.

The H3D-2^3 split (Figure 5.19 panels j-l) reveals that both approaches achieve reasonable segmentation for major categories including roof, vehicle, low vegetation, impervious surface, and tree. However, NOPS misassigns portions of shrub to tree and vertical surface categories. While both methods exhibit some soil/gravel misclassification, the proposed framework demonstrates improved boundary delineation for soil/gravel regions. These qualitative results confirm the quantitative performance improvements and validate the framework's enhanced capability for discovering fine-grained urban features in high-resolution 3D remote sensing data.

Table 5.3: Performance analysis on DALES dataset (%). The table compares three approaches: (1) traditional fully supervised methods with complete class knowledge, (2) our 3DLCDM framework demonstrating novel class discovery across different split configurations, and (3) backbone architecture impact using MinkUNet50 and MinkUNet101 on DALES-2^0. Pink-highlighted values indicate classes designated as novel in each split.

Category	Method	Ground	Vegetation	Car	Pole	Building	Truck	Fence	Power Line	mIoU
Traditional (All Known)	PointNet++ [69]	94.1	91.2	75.4	40.0	89.1	30.3	46.2	79.9	68.3
	ConvPoint [4]	96.9	91.9	75.5	40.3	96.3	21.7	29.6	86.7	67.4
	SPG [49]	94.7	87.9	62.9	28.5	93.4	18.7	33.6	65.2	60.6
	KPConv [79]	97.1	94.1	85.3	75.0	96.6	41.9	63.5	95.5	81.1
	Point Transformer [104]	96.8	93.2	82.2	70.4	96.8	38.9	60.6	95.9	79.3
	Point Transformer v2 [91]	96.0	93.3	84.6	71.5	94.9	39.6	61.5	95.3	79.6
	SPT [72]	96.7	93.1	86.1	65.3	96.7	52.4	52.7	94.0	79.6
	Point Transformer v3 [90]	96.7	93.6	87.0	74.2	96.8	45.7	66.4	97.1	82.2
	3DLST [61]	97.6	95.4	80.2	83.2	94.3	41.3	66.7	83.1	80.2
	LGMamba [50]	96.7	93.5	87.2	76.0	96.9	45.6	65.9	96.8	82.3
LCDM (Novel Discovery)	DALES-2^0 (Ours)	85.7	86.7	70.8	0.5	95.0	20.7	51.7	70.3	60.2
	DALES-2^1 (Ours)	95.7	86.5	74.8	66.6	94.6	3.1	54.8	75.4	68.9
	DALES-2^2 (Ours)	95.5	91.1	71.9	61.2	76.3	30.2	49.7	12.5	61.0
	DALES-2^3 (Ours)	95.7	91.6	69.5	73.1	94.7	21.6	48.2	79.3	71.7
	DALES-2^0 (MinkUNet50)	77.1	88.3	71.4	0.6	90.1	27.7	49.5	80.4	60.6
	DALES-2^0 (MinkUNet101)	88.4	88.1	66.4	0.0	94.3	39.8	53.9	64.3	61.9

5.4.5 Comparison with Traditional Semantic Segmentation Methods

To address the relationship between our land cover discovery mapping approach and existing semantic segmentation techniques, we compare our 3DLCDM framework with state-of-the-art traditional methods on the DALES dataset, as shown in Table 5.3. This comparison involves fundamentally different problem settings: traditional methods operate with complete supervision for all predefined classes, while our approach tackles the more challenging scenario of discovering novel classes.

The quantitative results reveal that traditional fully-supe-rvised methods achieve higher overall performance, with mIoU ranging from 60.6% (SPG [49]) to 82.3% (LGMamba [50]). In comparison, our framework achieves 60.2%-71.7% mIoU across different experimental configurations. However, when examining individual class performance where our method receives supervised signals, the performance gap becomes much smaller. For ground surfaces, our method achieves 95.7% IoU compared to the best traditional result of 97.6% (3DLST [61]), representing only a 1.9% difference. For buildings, we reach 95.0% versus 96.9% (LGMamba), a gap of 1.9%. These results demonstrate that our hybrid supervision strategy maintains competitive accuracy on known classes while simultaneously providing novel discovery capabilities.

The critical distinction lies in our framework's novel class discovery performance, which traditional methods cannot provide by design. We successfully discover classes like ground (85.7%), vegetation (86.5%), and cars (69.5%) when treated as novel during training. Traditional methods operate under closed-set assumptions and lack mechanisms for discovering novel classes.

In practical terms, while traditional methods achieve higher accuracy under ideal conditions with complete annotations, our framework provides essential adaptability for real-world urban mapping where new infrastructure, land use types, or building materials continuously emerge. The moderate accuracy trade-off enables a critical capability that traditional approaches fundamentally lack: the ability to discover and integrate previously unseen urban features into land cover maps without requiring extensive retraining or manual annotation. The competitive performance on known classes combined with novel discovery capabilities establishes our framework as a practical advancement for adaptive land cover mapping applications.

5.4.6 Impact of Backbone Architecture on LCDM Performance

An important observation from our experimental results is that LCDM inevitably involves some compromise in known class performance, as it requires allocating computational resources and feature learning capacity between supervised segmentation of established categories and unsupervised discovery of novel classes. To investigate whether this performance compromise can be mitigated through enhanced feature extraction capabilities, we conducted additional experiments using more advanced backbone networks as shown in the bottom rows of Table 5.3. Specifically, we evaluated MinkUNet50 and MinkUNet101 architectures on the DALES-2^0 split while keeping our progressive dual-branch prototype learning framework and all other components unchanged. The results demonstrate that more sophisticated backbone architectures can indeed improve overall performance: MinkUNet50 achieves 60.6% overall mIoU and MinkUNet101 reaches 61.9% overall mIoU, compared to 60.2% with the MinkUNet34C backbone. Notably, MinkUNet101 shows substantial improvement in truck segmentation, advancing from 20.7% to 39.8%, while MinkUNet50 demonstrates enhanced power line segmentation, increasing from 70.3% to 80.4%. These findings suggest that the inherent performance compromise between known and novel class segmentation can be partially addressed through the adoption of more powerful feature extraction networks. Furthermore, this demonstrates the flexibility of our hybrid supervision framework, which can readily accommodate advances in backbone network design to achieve enhanced overall performance while preserving robust novel class discovery capabilities. This architectural adaptability ensures that future improvements in 3D point

Table 5.4: Land cover discovery mapping results on SemanticPOSS dataset (%). Pink-highlighted values indicate the classes designated as novel in each split. The "Novel" column reports the mIoU for the newly discovered classes, "Base" corresponds to the mIoU for the pre-defined (known) classes, and "All" denotes the overall mIoU across all classes.

Split	Model	Bike	Building	Car	Cone	Fence	Ground	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
	CHNCD	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
	ours	**43.95**	41.76	0.48	**29.11**	**45.82**	**73.87**	**63.91**	54.43	**33.25**	**58.98**	**30.45**	7.64	20.35	42.64	**37.05**	**38.77**
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
	CHNCD	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
	ours	25.12	**82.16**	**43.99**	**36.70**	**35.51**	**79.09**	25.63	**79.30**	**31.97**	**53.86**	**28.80**	3.41	17.94	28.75	**45.72**	**41.81**
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
	CHNCD	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
	ours	**44.14**	**80.20**	**43.74**	**28.28**	**45.49**	**78.69**	**62.55**	**77.88**	**26.16**	**56.10**	**12.13**	2.19	6.93	**15.07**	**51.93**	**43.42**
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
	CHNCD	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
	ours	**44.93**	**80.00**	**39.51**	6.87	**45.99**	**78.77**	**62.66**	**77.22**	**33.47**	**39.46**	28.58	**5.87**	17.03	**17.40**	**50.82**	**43.11**

cloud processing networks can be effectively incorporated into our land cover discovery mapping framework.

5.4.7 Evaluation on SemanticPOSS Dataset with Alternative Splitting Strategies

To further validate the robustness and generalizability of the 3DLCDM framework, we conduct comprehensive experiments on the SemanticPOSS dataset collected in China. SemanticPOSS presents various and complicated scenes collected in a densely populated urban environment using a vehicle-equipped Pandora sensor module that differs from the sensor configurations used in DALES and H3D datasets. The dataset captures structurally complex scenarios with diverse urban elements, providing a complementary evaluation environment that tests the framework's adaptability across varying urban complexity levels.

Different from the largest-smallest alternation pattern employed in the DALES and H3D experiments, the SemanticPOSS evaluation follows the splitting protocol established by NOPS. This approach first ranks all classes according to their point frequencies in descending order, then divides the ranked list into four consecutive segments in sequential order. Each split designates one of these sequential segments as novel classes while treating the remaining classes as known categories.

The experimental results in Table 5.4 demonstrate the 3DLCDM framework's performance across all four SemanticPOSS splits. The framework achieves novel class mIoU

Table 5.5: Ablation study of key components in 3DLCDM on DALES-2^0 split (%). We remove each core component to evaluate their individual contributions: (1) removing the over-segmentation branch and using only the primary branch, (2) using fixed temperature $\mathcal{T} = 0.3$ instead of adaptive temporal scheduling in Sinkhorn-Knopp, and (3) using fixed global class frequency-based weights without batch-level dynamic updates.

Component	**Configuration**	Ground	Vegetation	Car	Pole	Building	Truck	Fence	Power Line	**Novel**	**Base**	**All**
w/o Dual-Branch	Single Branch	73.37	82.63	68.59	0.00	92.51	**25.70**	46.68	63.00	36.69	63.19	56.56
w/o Temporal SK	Fixed Temperature	66.24	**89.40**	66.21	0.02	93.43	25.22	49.79	67.12	33.13	65.20	57.18
w/o Dynamic Weight	Static Weights	79.86	88.50	**72.59**	0.00	94.10	20.56	41.35	58.64	39.93	62.62	56.95
Full Model (Ours)	All Components	**85.66**	86.65	70.78	**0.49**	**94.99**	20.72	**51.70**	**70.33**	**43.08**	**65.86**	**60.17**

ranging from 15.07% to 42.64%, with competitive performance compared to NOPS (8.95% to 35.70%) and CHNCD (10.30% to 43.68%) baselines. Base class performance shows consistent improvements across all configurations, with mIoU advancing from 26.57%-44.17% (NOPS) and 28.41%-48.18% (CHNCD) to 37.05%-51.93%. Overall mIoU increases from 29.38%-36.35% (NOPS) and 33.11%-39.44% (CHNCD) to 38.77%-43.42%, demonstrating the framework's effectiveness in structurally complex urban environments.

These comprehensive evaluations across the SemanticPOSS dataset with frequency-ordered splitting validate the framework's robustness across different experimental methodologies and urban complexity levels. The framework consistently achieves substantial base class improvements while maintaining competitive novel class discovery performance, with notable enhancements in critical land cover categories such as ground surface mapping and building segmentation. The performance gains across varying splitting strategies confirm the effectiveness of the dual-branch prototype learning mechanism and dynamic weighting strategy in adapting to diverse urban contexts, thereby supporting reliable land cover discovery mapping across different experimental configurations and urban complexity scenarios.

5.4.8 Ablation Study on Key Components

To evaluate the contribution of each core component in the 3DLCDM framework, we conduct comprehensive ablation experiments on the DALES-2^0 split, where ground and pole are designated as novel classes. Table 5.5 presents the quantitative results of systematically removing individual components while maintaining all other aspects of the framework unchanged.

The removal of the over-segmentation branch, retaining only the primary branch for novel class discovery, results in a notable degradation in performance. The novel mIoU

Table 5.6: Ablation study on base-novel class splitting strategies using DALES dataset (%). We evaluate four alternative splitting schemes to assess the robustness of our 3DLCDM framework: (1) Single major class (vegetation only), (2) Traffic-related categories (cars and trucks), (3) Elongated structures (power lines, fences, and poles), and (4) Diversified split (vegetation, buildings, power lines, and trucks).

Splitting Strategy	**Method**	Ground	Vegetation	Car	Pole	Building	Truck	Fence	Power Line	**Novel**	**Base**	**All**
Single Major Class	NOPS	95.09	87.39	65.27	41.28	91.46	**22.82**	30.86	67.26	87.39	59.15	62.68
	Ours	**95.25**	**90.64**	**72.23**	**65.14**	**94.79**	21.67	**51.31**	**74.08**	**90.64**	**67.78**	**70.64**
Traffic Categories	NOPS	95.18	90.71	45.83	52.04	**94.53**	18.80	40.34	73.97	32.32	74.46	63.92
	Ours	**95.64**	**91.43**	**67.09**	**69.29**	94.45	**20.14**	**51.72**	**77.58**	**43.62**	**80.02**	**70.92**
Elongated Structures	NOPS	95.14	90.31	69.18	0.12	94.36	**25.83**	37.39	45.44	27.65	74.96	57.22
	Ours	**95.65**	**91.62**	**76.37**	**18.14**	**94.76**	22.48	**40.47**	**53.71**	**37.44**	**76.18**	**61.65**
Diversified Split	NOPS	93.89	50.67	**69.61**	35.84	58.91	2.08	40.80	0.00	27.92	60.03	43.97
	Ours	**95.44**	**71.15**	67.18	**59.47**	**78.80**	**2.28**	**50.64**	**0.10**	**38.08**	**68.18**	**53.13**

decreases from 43.08% to 36.69%, a reduction of 6.39%. This decline is particularly pronounced for the ground class, where the IoU drops from 85.66% to 73.37%. The performance reduction demonstrates the importance of progressive prototype refinement through over-segmentation, which enables the model to capture fine-grained distinctions among emerging urban features that might otherwise be overlooked by a single prototype clustering approach. The dual-branch architecture facilitates more comprehensive exploration of the unlabeled feature space, thereby enhancing the discovery and characterization of novel land cover classes in complex urban environments.

Replacing the adaptive temporal temperature scheduling with a fixed temperature parameter ($\mathcal{T} = 0.3$) leads to substantial performance degradation, particularly affecting novel class discovery. The novel mIoU decreases by 9.95% (from 43.08% to 33.13%), while the overall mIoU drops by 2.99%. The ground class experiences a significant IoU reduction from 85.66% to 66.24%. This degradation underscores the critical role of adaptive temperature control in generating high-quality pseudo-labels. The temporal scheduling mechanism ensures smooth pseudo-label assignments during early training phases while progressively sharpening the distributions as feature representations stabilize, thereby preventing premature overconfident assignments that could mislead the novel class discovery process.

The substitution of the batch-level dynamic weighting mechanism with static frequency-based weights results in a 3.15% decrease in novel mIoU (from 43.08% to 39.93%) and a 3.22% reduction in overall mIoU. This degradation highlights the importance of adaptive class balancing in urban point cloud analysis. The dynamic weighting strategy effectively addresses the severe class imbalance inherent in urban remote sensing data by adaptively

emphasizing underrepresented classes during training, thereby ensuring that environmentally significant but rare features receive adequate attention in the learning process.

The ablation results collectively demonstrate that all three components contribute synergistically to the framework's effectiveness in land cover discovery mapping. Each component addresses distinct challenges: dual-branch prototype learning ensures comprehensive exploration of the unlabeled feature space, temporal Sinkhorn-Knopp balancing provides robust pseudo-label generation, and dynamic weighting strategy effectively handles class imbalance in urban environments. These findings validate the design choices made in developing the 3DLCDM framework and confirm that the integrated approach is essential for discovering novel urban classes from high-resolution 3D remote sensing data.

5.4.9 Evaluation of Alternative Base-novel Class Splitting Strategies

To evaluate the robustness of our 3DLCDM framework across diverse novel class discovery scenarios, we conduct comprehensive experiments using four alternative splitting strategies on the DALES dataset, as presented in Table 5.6. These experiments are designed to simulate the heterogeneous nature of novel class emergence in real-world urban environments and assess the framework's adaptability to varying land cover discovery mapping challenges.

The single major class strategy designates vegetation as the sole novel class, representing scenarios where large-scale natural features emerge as previously unannotated elements in urban mapping systems. Under this configuration, our method achieves a novel class IoU of 90.64% compared to 87.39% for the baseline NOPS approach, demonstrating an improvement of 3.25%. The overall mIoU increases from 62.68% to 70.64%, indicating enhanced performance across both novel and base classes. This scenario reflects situations where extensive green infrastructure or urban forestry initiatives introduce substantial natural components that require integration into existing land cover classification systems.

The traffic-related categories strategy evaluates the discovery of semantically coherent but spatially sparse urban elements by treating cars and trucks as novel classes. Our framework achieves a novel mIoU of 43.62% compared to 32.32% for NOPS, representing an improvement of 11.30%. This configuration simulates the emergence of new mobility infrastructure or vehicle types that necessitate updating of transportation-related land cover categories, reflecting the dynamic nature of urban mobility systems and their environmental implications.

The elongated structures strategy presents a challenging scenario by simultaneously

Table 5.7: Ablation study on partitioning strategies for DALES-2^0 split (%). Pink-highlighted values indicate novel classes. ★ indicates the strategy used in main experiments.

Method	Partitioning	Ground	Vegetation	Car	Pole	Building	Truck	Fence	Power Line	Novel	Base	All
NOPS	FPS+KNN	62.23	**89.71**	**65.43**	0.00	90.36	24.49	25.52	42.65	31.12	56.36	50.05
	Spatial Grid★	**64.90**	86.55	58.88	0.00	**93.16**	**27.80**	**25.97**	**56.59**	**32.45**	**58.16**	**51.73**
3DLCDM	FPS+KNN	83.48	**88.18**	**72.80**	0.04	94.70	19.71	51.05	67.34	41.76	65.63	59.66
	Spatial Grid★	**85.66**	86.65	70.78	**0.49**	**94.99**	**20.72**	**51.70**	**70.33**	**43.08**	**65.86**	**60.17**

Table 5.8: Ablation study on partitioning strategies for H3D-3^0 split (%). Pink-highlighted values indicate novel classes. ★ indicates the strategy used in main experiments.

Method	Partitioning	Low Vegetation	Impervious Surface	Vehicle	Urban Furniture	Shrub	Tree	Vertical Surface	Roof	Façade	Chimney	Soil	Novel	Base	All
NOPS	Spatial Grid	**54.26**	**41.06**	60.91	38.96	56.50	90.61	42.34	86.76	68.94	0.00	23.10	**31.77**	58.52	51.22
	FPS+KNN★	21.62	28.37	**76.71**	**53.72**	**60.64**	**91.93**	**52.24**	**92.20**	**80.66**	**17.74**	**25.55**	22.58	**66.71**	**54.67**
3DLCDM	Spatial Grid	47.21	53.54	73.76	58.73	65.87	92.89	51.11	92.13	80.99	**0.10**	25.58	33.62	67.63	58.36
	FPS+KNN★	**68.03**	**62.03**	**78.04**	**59.76**	**69.76**	**94.02**	**56.46**	**93.09**	**83.00**	0.01	**26.40**	**43.36**	**70.07**	**62.78**

treating power lines, fences, and poles as novel classes. These elements are characterized by linear geometric properties and sparse point distributions, representing the discovery of fine-scale infrastructure components. Despite the inherent difficulty in identifying such features, our method achieves a novel mIoU of 37.44% compared to 27.65% for NOPS, showing an improvement of 9.79%. This strategy mirrors real-world situations where utility networks, communication infrastructure, or boundary demarcation systems evolve and require incorporation into land cover mapping frameworks.

The diversified split strategy constitutes the most complex evaluation scenario, treating vegetation, buildings, power lines, and trucks as novel classes. This configuration challenges the framework to simultaneously discover major natural features, primary built structures, and minor infrastructure elements, reflecting comprehensive urban development scenarios where multiple land cover types emerge concurrently. Our approach achieves a novel mIoU of 38.08% compared to 27.92% for NOPS, demonstrating an improvement of 10.16%. The overall mIoU increases from 43.97% to 53.13%, indicating maintained effectiveness even when substantial portions of the urban environment are treated as novel classes.

These experiments demonstrate consistent improvements across diverse splitting configurations, validating the robustness of our 3DLCDM framework for land cover discovery mapping applications. The results confirm that the framework's dual-branch prototype learning mechanism, temporal Sinkhorn-Knopp balancing, and dynamic weighting strategy effectively adapt to varying novel class discovery scenarios, providing reliable performance for monitoring evolving urban environments across different land cover compositions and spatial distributions.

5.4.10 Evaluation of Dataset-specific Partitioning Strategies

The employment of distinct partitioning strategies for the DALES and H3D datasets reflects both dataset-specific optimization and methodological validation objectives. From a technical perspective, the DALES dataset's regular tile-based structure and relatively balanced class distribution favor spatial grid partitioning that preserves large-scale geometric relationships, while the H3D dataset's complex semantic composition and irregular spatial distribution of minority classes necessitate adaptive partitioning to preserve local geometric characteristics across diverse urban features. The spatial grid partitioning strategy is specifically designed to leverage DALES' inherent tile-based organization, where the dataset is naturally structured as regular tiles, maximizing the preservation of original geometric relationships and spatial continuity across large-scale urban features. Conversely, the FPS+KNN partitioning strategy is tailored to address H3D's higher class complexity and the inherent spatial heterogeneity of fine-grained urban features.

Furthermore, to validate the robustness of our method, we conduct additional ablation experiments based on partitioning strategies by applying both preprocessing approaches to each dataset and performing experiments to obtain comprehensive LCDM results. Tables 5.7 and 5.8 present these cross-partitioning experimental results, demonstrating consistent effectiveness of our 3DLCDM framework across different preprocessing methodologies. On the DALES dataset, our method consistently outperforms the NOPS baseline under both partitioning strategies: achieving novel mIoU improvements from 32.45% to 43.08% with spatial grid partitioning and from 31.12% to 41.76% with FPS+KNN partitioning, while maintaining competitive base class performance. Similarly, on the H3D dataset across different partitioning configurations, our framework demonstrates substantial novel class discovery improvements, advancing from 22.58% to 43.36% with FPS+KNN partitioning and from 31.77% to 33.62% with spatial grid partitioning, with consistent base class performance enhancements. The consistent performance gains across diverse preprocessing paradigms validate that our core algorithmic contributions, including progressive dual-branch prototype learning, temporal Sinkhorn-Knopp normalization, and dynamic weighting mechanisms, maintain robustness across varying data characteristics and preprocessing methodologies. This robustness is particularly significant for practical deployment scenarios where preprocessing constraints may be dictated by dataset characteristics, computational resources, or domain-specific requirements, ensuring reliable novel class discovery performance across varying operational environments.

Table 5.9: Comprehensive computational efficiency comparison on DALES and H3D datasets with different partitioning strategies. ★ indicates the strategy used in main experiments.

Split	Method	Partitioning	Initial Memory (GB)	Peak Memory (GB)	Memory Increase (GB)	Test Time (min)
DALES-2^0	NOPS	FPS+KNN	0.14	9.78	9.64	3.15
		Spatial Grid★	0.14	6.54	6.40	2.63
	3DLCDM	FPS+KNN	0.00	9.62	9.62	3.50
		Spatial Grid★	0.00	6.49	6.49	2.67
H3D-3^0	NOPS	Spatial Grid	0.14	9.15	9.00	1.92
		FPS+KNN★	0.14	4.47	4.33	3.03
	3DLCDM	Spatial Grid	0.00	12.16	12.16	1.94
		FPS+KNN★	0.00	4.48	4.48	2.83

5.4.11 Computational Efficiency Analysis

To evaluate computational scalability for large-scale point cloud processing and validate efficiency across different partitioning strategies, we conduct comprehensive efficiency analysis comparing our 3DLCDM framework with the baseline NOPS method during inference on Tesla V100-SXM2-32GB under identical experimental configurations. All measurements are performed on the original dataset sizes to ensure representative evaluation. Table 5.9 presents detailed computational metrics across both datasets with both spatial grid and FPS+KNN partitioning strategies. The analysis reveals distinct computational patterns: FPS+KNN partitioning requires higher memory consumption and longer inference times for the DALES dataset due to the disruption of natural tile-based organization, while spatial grid partitioning shows increased memory requirements for the H3D dataset due to incompatibility with irregular spatial distributions. Despite these partitioning-induced variations, our framework achieves substantial novel class discovery improvements while maintaining computational efficiency comparable to the baseline across all configurations.

The computational requirements show that 3DLCDM maintains practical feasibility for operational deployment across different preprocessing paradigms. Memory consumption remains within standard GPU limits, and inference times stay comparable to baseline methods regardless of the partitioning strategy employed. The observed efficiency variations between partitioning strategies are primarily driven by dataset-specific characteristics, with both approaches maintaining practical computational requirements for operational deployment. This demonstrates that the framework can be deployed in resource-constrained environments while providing enhanced novel class discovery capabilities for dynamic land cover mapping applications, confirming the scalability and practical viability of our approach across varying data preprocessing requirements.

5.5 Discussion

5.5.1 Temporal Implications for LCDM

Our experimental design inherently incorporates temporal analysis through cross-temporal validation using the H3D dataset, addressing a critical aspect of land cover monitoring. By training on multiple historical epochs (March 2016, March 2018, and March 2019) and testing on a different temporal period (November 2018), we evaluate the framework's ability to discover novel classes across different temporal contexts and acquisition conditions. The selection of November 2018 as the test epoch is motivated by dual temporal considerations: it provides autumn acquisition conditions that contrast with the spring training data while ensuring temporal separation from training epochs, creating an optimal scenario for assessing both seasonal and multi-year robustness across varying acquisition conditions. This approach addresses the fundamental challenge of reliably identifying emerging land cover features regardless of temporal variations in data acquisition conditions, seasonal changes, or environmental fluctuations.

The temporal separation between training and testing data ensures that our novel class discovery mechanism can generalize across different seasons and acquisition conditions, which is essential for operational land cover discovery mapping systems. In dynamic environments, such temporal robustness enables continuous processing of new acquisitions while maintaining the ability to identify both established and newly emerging features. The stable performance across temporal epochs validates the framework's applicability for long-term monitoring applications where data may be acquired under varying environmental conditions throughout different seasons and years.

The H3D dataset enables comprehensive evaluation of the temporal robustness of our novel class discovery mechanism across different acquisition periods. Classes designated as "novel" in our experiments (e.g., low vegetation, impervious surfaces, chimneys) serve as proxies to validate our framework's discovery mechanisms under controlled conditions that are procedurally equivalent to encountering genuinely unseen urban features in operational deployment. The ability to consistently discover these feature types across temporal epochs demonstrates the framework's capability to identify and adapt to changes as they occur in dynamic environments, providing a foundation for monitoring land cover evolution over extended time periods.

5.5.2 Practical Applications in Environmental Remote Sensing

From an environmental perspective, our experimental results demonstrate that the 3DLCDM framework substantially benefits LCDM through its novel class discovery capability, which addresses a fundamental limitation in current remote sensing practice: the inability of traditional land cover mapping systems to adapt when new urban features emerge over time. For instance, in the DALES dataset, the IoU of ground segmentation is boosted from 54.61% to 85.66% when treated as a novel class, while the car category shows a substantial increase from 41.48% to 69.54% when withheld from supervision. In the H3D dataset, our method achieves notable gains in discovering low vegetation (from 21.62% to 68.03%) and impervious surface (from 28.37% to 62.03%) when these are treated as novel categories. Our experimental validation demonstrates that the framework can successfully discover and segment classes withheld from supervision, achieving novel class mIoU improvements of up to 16.95% on the DALES dataset and up to 24.43% on the H3D dataset. These improvements demonstrate the framework's core value: enabling automatic identification and accurate segmentation of emerging urban features without requiring extensive manual re-annotation efforts, thereby supporting continuous updating of land cover inventories as urban environments evolve.

This adaptive capability has significant implications for operational environmental monitoring applications, directly addressing real-world scenarios where urban development introduces previously unobserved land cover types that are not represented in existing training datasets. Traditional remote sensing approaches require complete dataset re-annotation and model retraining when new urban features appear, creating significant operational challenges and costs for long-term monitoring programs. Our framework overcomes this limitation by automatically incorporating novel classes into the land cover mapping system through unsupervised discovery mechanisms. The framework's demonstrated ability to discover novel classes across different urban categories (ground surfaces, vegetation types, building structures, and vehicle classifications) provides evidence for its potential to automatically adapt to emerging urban features as they appear in real landscapes, enabling incremental incorporation of new environmental features through its unsupervised discovery mechanism. The temporal robustness validated through cross-temporal experiments on the H3D dataset further supports the framework's applicability for long-term environmental monitoring, where reliable performance across different acquisition conditions is essential for maintaining data continuity and reliability in longitudinal studies.

5.6 Chapter Summary

This chapter introduces 3DLCDM, which bridges research frameworks with operational deployment through hybrid supervision for LCDM. The framework integrates supervised segmentation of known classes with unsupervised discovery of emerging urban features, employing a shared MinkUNet34C backbone that processes features through complementary pathways: a supervised head for established categories and dual unsupervised heads for novel class discovery. The dual-branch design comprises a primary branch with fixed prototypes corresponding to expected novel classes and an over-segmentation branch with progressive prototype scheduling. Temporal Sinkhorn-Knopp normalization transforms raw similarity scores into balanced probability distributions, generating discrete pseudo-labels that stabilize discovery under noisy clustering conditions. Frequency-based weighting dynamically combines per-iteration batch statistics with long-term momentum to mitigate severe class imbalance between dominant urban categories and rare but environmentally significant features.

Experimental evaluation on DALES and H3D datasets demonstrates consistent improvements across all performance metrics, with novel class mIoU improving by up to 16.95% and 24.43% respectively, base class mIoU advancing by up to 11.76% and 5.95%, and overall mIoU increasing by up to 10.80% and 10.15%. Block partitioning strategies adapted to dataset characteristics (spatial grid for regular tile-based DALES, FPS+KNN for heterogeneous H3D) enable scalable processing while preserving geometric fidelity necessary for discovering fine-grained novel structures. The framework's ability to automatically discover and segment novel urban features enables continuous generation of updated land cover maps that accurately reflect dynamic urban morphology, establishing 3DLCDM as a practical solution for LCDM in evolving urban environments.

This work situates discovery within land cover mapping applications through hybrid supervision and operational-scale processing. Taken together with the geometry-aware organization established in Chapter 3 and the adaptive representation and transfer advanced in Chapter 4, the dissertation develops a unified research program addressing the fundamental challenges of novel class discovery in large-scale 3D point cloud segmentation: organizing spatial information effectively, integrating complementary representations adaptively, and bridging the gap between research frameworks and operational deployment requirements.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This dissertation addresses the fundamental challenge of discovering and segmenting novel semantic categories in large-scale 3D point clouds. Urban environments continuously evolve with emerging infrastructure types, architectural innovations, and adaptive land use patterns that existing fixed-taxonomy systems cannot accommodate. Three interconnected research questions have guided this work, progressing from foundational geometric organization through adaptive representation learning to practical deployment integration.

RQ1: How can discovery systems effectively leverage spatial organization to preserve fine-grained geometric structure for novel category recognition? Novel class discovery in 3D point clouds requires distinguishing categories with subtle geometric differences, yet spatial relationships that define these categories can be lost through inappropriate data organization. Without explicit consideration of spatial proximity and local structure, discovery systems may fail to capture the fine-grained geometric patterns essential for identifying novel categories. Chapter 3 establishes that cluster-based spatial organization with proximity-aware neighbor construction enables preservation of geometric fidelity for effective discovery. CHNCD constructs compact local neighborhoods within cluster-defined regions and applies spatial attention mechanisms to model fine-grained relationships, maintaining boundary fidelity for small structures across large-scale datasets. Evaluation across diverse point cloud types (indoor S3DIS, street-level Toronto-3D, airborne SemanticSTF, and mobile mapping SemanticPOSS) demonstrates consistent improvements in boundary fidelity and novel class recognition, with SemanticPOSS-4^0 achieving 43.68% novel class mIoU compared to 35.70% for prior state-of-the-art meth-

ods. This work reveals that explicit spatial organization through hierarchical clustering is fundamental to preserving the geometric characteristics that define novel categories.

RQ2: How can complementary geometric and semantic representations be fused adaptively, with principled knowledge transfer from known to novel categories? Urban sensing generates point clouds with extreme density variations, from sparse airborne scans to dense terrestrial acquisitions, posing challenges for uniform sampling strategies. Furthermore, relying solely on geometric features limits semantic discriminability for functionally distinct but geometrically similar structures. Chapter 4 demonstrates that density-aware geometric sequence modeling combined with gated multi-source fusion enables robust discovery across diverse sensing modalities. AGDNet extracts meaningful local patterns despite point distribution variations through density-based sequences, while dynamic Gaussian embeddings and gated fusion mechanisms allow the network to adaptively weight geometric versus semantic features based on scene characteristics. Multi-level knowledge transfer (operating at category, instance, and distribution levels) provides a principled framework for propagating learned representations from well-annotated known classes to underrepresented novel categories. Comprehensive evaluation demonstrates average improvements of 6.47%/3.48% (Toronto-3D), 4.61%/3.12% (SemanticSTF), and 6.64%/4.24% (SemanticPOSS) in novel class mIoU over NOPS/CHNCD baselines, validating that adaptive multi-source modeling outperforms fixed feature extractors in heterogeneous discovery scenarios.

RQ3: How can supervised segmentation for established classes be coupled with discovery mechanisms for emerging classes to support continuously updated land cover mapping? Operational land cover mapping faces a unique challenge: maintaining consistent, accurate labels for established categories across temporal acquisitions while simultaneously identifying emerging feature types. This requirement differs from typical discovery settings, where the focus is on identifying novel categories without the constraint of maintaining precise labels for known classes. Chapter 5 establishes that hybrid supervision, coupling full labels for known classes with prototype-driven discovery for novel categories enables continuous mapping without sacrificing accuracy. 3DLCDM integrates temporal pseudo-label stabilization to prevent prediction oscillation for gradually emerging features, while frequency-aware dynamic weighting addresses extreme class imbalance common in land cover data. Compatibility with standard sparse-convolutional architectures ensures deployment feasibility, as demonstrated through evaluation on airborne datasets (DALES, H3D) where novel class mIoU improvements reach 16.95% and 24.43% respectively. This work demonstrates that discovery capabilities can be effectively integrated into operational mapping pipelines through carefully designed hybrid supervision strategies.

6.1.1 Contributions

This dissertation advances 3D novel class discovery and its deployment for urban land cover mapping through the following contributions:

- **Scalable geometry-aware organization.** A representation and neighborhood strategy that preserves fine-grained geometry and boundary fidelity while remaining efficient at large scale, improving stability of novel/base separation across indoor and outdoor datasets.
- **Adaptive multi-source modeling and semantic transfer.** Complementary encoders for local geometry and global structure, fused through context-aware mechanisms and coupled with multi-level transfer objectives to bridge the semantic gap between known and novel classes.
- **Hybrid supervision for operational mapping.** A unified formulation that jointly segments established classes and discovers emerging categories under class imbalance, enabling continuously updated 3D land cover products.
- **Comprehensive validation across diverse settings.** Systematic evaluation on representative indoor, street-level, and high-resolution remote-sensing datasets demonstrates robustness across sensors, acquisition geometries, and urban complexity levels.

6.2 Future Work

While this dissertation advances novel class discovery for point cloud segmentation, several promising research directions remain underexplored. These directions reflect fundamental shifts in how discovery systems might reason about, learn from, and adapt to evolving urban environments.

6.2.1 Large Language Model Reasoning for Contextual Scene Understanding

Current discovery systems primarily leverage geometric and visual features for category identification, with limited integration of contextual reasoning and world knowledge that

could enhance interpretation of complex urban scenes. Large language models offer significant capabilities in spatial reasoning, multi-step inference, and common-sense understanding that could transform how discovery systems identify and categorize novel features. Recent advances demonstrate their capacity to process 3D point clouds through multimodal interfaces, enabling instruction-driven segmentation, scene-level reasoning, and contextual disambiguation of ambiguous structures.

Integrating language models with discovery frameworks presents opportunities at multiple levels. Natural language instructions could replace fixed taxonomies, enabling domain experts to describe emerging features contextually rather than through predefined labels. At the scene understanding level, language models demonstrate strong capabilities in compositional reasoning, decomposing complex structures into known primitives and inferring novel categories through analogy. For instance, such systems could identify vertical solar farms by reasoning that they share photovoltaic function with ground-mounted arrays but exhibit distinct vertical geometry, without requiring explicit training examples. Successful integration would enable discovery systems that reason about function, context, and evolution patterns rather than relying solely on appearance similarity, facilitating natural language explanations for discovered features in urban planning workflows.

6.2.2 Multimodal Fusion for Robust Discovery Under Heterogeneous Data

Operational urban monitoring increasingly integrates diverse sensing modalities beyond LiDAR point clouds, including RGB imagery, thermal sensors, hyperspectral data, and crowdsourced annotations. Effectively fusing these heterogeneous sources presents both opportunities and challenges for novel class discovery. Textual labels, 2D images, and 3D geometry provide complementary signals where text encodes semantic priors, images capture appearance and texture, while point clouds preserve spatial structure and scale. However, naive fusion strategies suffer from modality imbalance where dominant signals suppress informative but sparse cues, and feature misalignment across coordinate systems and resolutions.

Emerging frameworks address these limitations through adaptive weighting and uncertainty-aware fusion. Cross-modal approaches employ uncertainty estimation to down-weight ambiguous features when projecting across modalities, preventing noisy embeddings from degrading representations. Correlation fusion modules learn to combine geometric, visual, and textual features based on scene complexity and sensor reliability, achieving robust performance when individual modalities degrade. Multimodal discovery systems would significantly enhance robustness in resource-constrained deployments where single-modality

systems face challenges. For instance, LiDAR-only processing has difficulty distinguishing vegetated green roofs from conventional roofs, but integrating vegetation indices from multispectral imagery disambiguates these functionally distinct structures. Similarly, thermal data reveals heat islands and energy-inefficient buildings, enabling discovery of climate-relevant features difficult to identify from geometry and color alone.

6.2.3 Agent-Based Discovery for Continuous Urban Intelligence

While the methods developed in this dissertation effectively discover novel categories within individual point cloud acquisitions, urban environments require continuous monitoring where infrastructure evolves through construction, renovation, and adaptive reuse. Agent-based paradigms, empowered by language models for planning and decision-making, offer a transformative approach where autonomous systems actively query environments, reason about changes, and refine taxonomies over time. Embodied intelligence in open city environments represents a critical frontier, emphasizing spatial reasoning, task planning, and multi-agent collaboration for large-scale outdoor settings.

Language model-driven agents could extend discovery capabilities through several mechanisms. Active sensing strategies enable agents to plan optimal viewpoints for disambiguating uncertain regions, reducing annotation costs by prioritizing high-information acquisitions. Temporal reasoning allows agents to detect gradual changes such as building retrofits adding solar panels versus sudden appearances like temporary structures, adapting discovery strategies accordingly. Multi-agent systems could distribute discovery tasks spatially, with specialized agents monitoring distinct urban districts and sharing learned prototypes through communication protocols, enabling city-wide knowledge aggregation. Such agent-based extensions would enable truly adaptive monitoring systems that maintain current inventories despite urban flux. For disaster response, agents could rapidly map emergent features by reasoning about pre/post-event differences. For policy compliance, agents could monitor permitted construction against approved plans, flagging deviations for review. For climate adaptation, agents could track green infrastructure proliferation and assess spatial equity in implementations, informing evidence-based interventions.

6.2.4 Scalability to Real-Time Systems and Deployment in Dynamic Environments

Operational deployment of novel class discovery systems introduces considerations beyond algorithmic accuracy, particularly regarding computational efficiency for real-time

or near-real-time processing and model adaptation mechanisms for long-term deployment. Enhancing computational efficiency would expand the applicability of novel class discovery toward real-time systems in broader application scenarios. Future research could explore lightweight backbone variants through knowledge distillation, preserving discovery capabilities while reducing computational overhead. Hierarchical processing strategies that prioritize regions with high uncertainty estimates offer another avenue for improving efficiency, while incremental inference mechanisms that update predictions as new acquisitions arrive would enable more responsive mapping workflows. Integration with hardware-accelerated sparse convolution implementations and model quantization techniques could further enhance throughput for city-scale or regional deployments.

Extending novel class discovery to long-term operational deployment requires addressing model adaptation over time. The temporal validation conducted on the H3D dataset demonstrates robustness across different acquisition periods, providing a foundation for longer-term monitoring applications. As urban landscapes accumulate changes over years, the underlying data distributions may shift beyond what initial training captures. Continual learning mechanisms that refine prototypes incrementally while preserving performance on established categories would enable sustained accuracy without complete retraining. Confidence calibration strategies could support human-in-the-loop workflows where domain experts validate discovered categories before integration into official land cover inventories. Developing longitudinal benchmark datasets capturing urban evolution across multi-year periods would facilitate systematic evaluation of these model adaptation capabilities, enabling novel class discovery frameworks to serve broader operational scenarios with sustained performance over extended deployment periods.

These future directions collectively point toward discovery systems that reason contextually through language, fuse multimodal evidence robustly, operate continuously through agent-based intelligence, and scale efficiently for real-time deployment in dynamic environments. Realizing this vision requires not only algorithmic advances but also new evaluation paradigms, benchmark datasets capturing temporal evolution, and frameworks for human-AI collaboration in operational urban intelligence systems.

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis K. Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proc. CVPR*, pages 1534–1543. IEEE, 2016.

[2] Simon Elias Bibri, Jeffrey Huang, Senthil Kumar Jagatheesaperumal, and John Krogstie. The synergistic interplay of artificial intelligence and digital twin in environmentally planning sustainable smart cities: A comprehensive systematic review. *Environ. Sci. Ecotechnol.*, 20:100433, 2024.

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proc. CVPR*, pages 11679–11689. IEEE, 2020.

[4] Alexandre Boulch. Convpoint: Continuous convolutions for point cloud processing. *Comput. Graph.*, 88:24–34, 2020.

[5] Dengfeng Chai, Shawn Newsam, Hankui K. Zhang, Yifan Qiu, and Jingfeng Huang. Cloud and cloud shadow detection in landsat imagery based on deep convolutional neural networks. *Remote Sens. Environ.*, 225:307–316, 2019.

[6] Tzu-Hsin Karen Chen, Chunping Qiu, Michael Schmitt, Xiao Xiang Zhu, Clive E. Sabel, and Alexander V. Prishchepov. Mapping horizontal and vertical urban densification in denmark with landsat time-series from 1985 to 2018: A semantic segmentation solution. *Remote Sens. Environ.*, 251:112096, 2020.

[7] Yidong Chen, Guorong Cai, Qiming Xia, Zhaoliang Liu, Binghui Zeng, Zongliang Zhang, Jinhe Su, and Zongyue Wang. Not all points are balanced: Class balanced single-stage outdoor multi-class 3d object detector from point clouds. *Int. J. Appl. Earth Obs. Geoinf.*, 128:103766, 2024.

[8] Yujia Chen, Guo Zhang, Hao Cui, Xue Li, Shasha Hou, Jinhao Ma, Zhijiang Li, Haifeng Li, and Huabin Wang. A novel weakly supervised semantic segmentation framework to improve the resolution of land cover product. *ISPRS J. Photogramm. Remote Sens.*, 196:73–92, 2023.

[9] Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Meta discovery: Learning to discover novel classes given very limited data. In *Proc. ICLR*. OpenReview.net, 2022.

[10] Christopher B. Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proc. CVPR*, pages 3075–3084. IEEE, 2019.

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. NeurIPS*, pages 2292–2300, 2013.

[12] Hengming Dai, Xiangyun Hu, Jinming Zhang, Zhen Shu, Jiabo Xu, and Juan Du. Large-scale ALS point cloud segmentation via projection-based context embedding. *IEEE Trans. Geosci. Remote. Sens.*, 62:1–16, 2024.

[13] Jiacheng Deng, Jiahao Lu, and Tianzhu Zhang. Quantity-quality enhanced self-training network for weakly supervised point cloud semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–17, 2025.

[14] Jing Du, Guorong Cai, Zongyue Wang, Shangfeng Huang, Jinhe Su, José Marcato Junior, Julian Smit, and Jonathan Li. Resdlps-net: Joint residual-dense optimization for large-scale point cloud semantic segmentation. *ISPRS J. Photogramm. Remote Sens.*, 182:37–51, 2021.

[15] Jing Du, Guorong Cai, Zongyue Wang, Jinhe Su, Min Huang, John S. Zelek, José Marcato Jr., and Jonathan Li. Mtcloud: Multi-type convolutional linkage network for point cloud instance segmentation. *Expert Syst. Appl.*, 270:126432, 2025.

[16] Jing Du, Zuning Jiang, Shangfeng Huang, Zongyue Wang, Jinhe Su, Songjian Su, Yundong Wu, and Guorong Cai. Point cloud semantic segmentation network based on multi-scale feature fusion. *Sensors*, 21(5):1625, 2021.

[17] Jing Du, Lingfei Ma, Jing Li, Nannan Qin, John S. Zelek, Haiyan Guan, and Jonathan Li. Rdmknet & toronto-rdmk: Large-scale datasets for road marking classification and segmentation. *IEEE Trans. Intell. Transp. Syst.*, 25(10):13467–13482, 2024.

[18] Jing Du, Linlin Xu, Lingfei Ma, Kyle Gao, John Zelek, and Jonathan Li. 3d semantic segmentation: Cluster-based sampling and proximity hashing for novel class discovery. *ISPRS J. Photogramm. Remote Sens.*, 223:274–295, 2025.

[19] Jing Du, John Zelek, Dedong Zhang, and Jonathan Li. 3dlcdm: Hybrid supervision for land cover discovery mapping of emerging urban structures in 3d remote sensing. *Remote Sens. Environ.*, 331:115018, 2025.

[20] Jing Du, John Zelek, Dedong Zhang, and Jonathan Li. Gated multi-source fusion with geometric sequence modeling for novel urban structure discovery. *ISPRS J. Photogramm. Remote Sens.*, 230:495–523, 2025.

[21] Shouhang Du, Shihong Du, Bo Liu, and Xiuyuan Zhang. Mapping large-scale and fine-grained urban functional zones from vhr images using a multi-scale semantic segmentation network and object based approach. *Remote Sens. Environ.*, 261:112480, 2021.

[22] Hanna Elie and Comín Francisco A. Urban green infrastructure and sustainable development: A review. *Sustainability*, 13(20), 2021.

[23] Felix K. Abagale Etienne Umukiza and Thomas Apusiga Adongo. Characterization of landuse and landcover dynamics and their impact on runoff generation patterns in dam catchments of northern ghana. *Geocarto Int.*, 39(1):2335247, 2024.

[24] Mohammed A. Fadhel, Ali M. Duhaim, Ahmed Saihood, Ahmed Sewify, Mokhaled N. A. Al-Hamadani, Ahmed Shihab Albahri, Laith Alzubaidi, Ashish Gupta, Seyedali Mirjalili, and Yuantong Gu. Comprehensive systematic review of information fusion methods in smart cities and urban environments. *Inf. Fusion*, 107:102317, 2024.

[25] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proc. ICCV*, pages 9264–9272. IEEE, 2021.

[26] Mario Fuentes Reyes, Yuxing Xie, Xiangtian Yuan, Pablo d'Angelo, Franz Kurz, Daniele Cerra, and Jiaojiao Tian. A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. *ISPRS J. Photogramm. Remote Sens.*, 205:74–97, 2023.

[27] Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Mach. Learn.*, 113(7):4845–4901, 2024.

[28] Kai Chen Goh, Tonni Agustiono Kurniawan, Hui Hwang Goh, Dongdong Zhang, Meihui Jiang, Wei Dai, Muhammad Imran Khan, Mohd Hafiz Dzarfan Othman, Faissal Aziz, Abdelkader Anouzla, and Christia Meidiana. Harvesting valuable elements from solar panels as alternative construction materials: A new approach of waste valorization and recycling in circular economy for building climate resilience. *Sust. Mater. Technol.*, 41:e01030, 2024.

[29] Peiyan Gu, Chuyu Zhang, Ruijie Xu, and Xuming He. Class-relation knowledge distillation for novel class discovery. In *Proc. ICCV*, pages 16428–16437. IEEE, 2023.

[30] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *Proc. ICLR*, 2022.

[31] Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proc. ICLR*. OpenReview.net, 2020.

[32] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proc. ICCV*, pages 8400–8408. IEEE, 2019.

[33] Xu Han, Chong Liu, Yuzhou Zhou, Kai Tan, Zhen Dong, and Bisheng Yang. Whu-urban3d: An urban scene lidar point cloud dataset for semantic instance segmentation. *ISPRS J. Photogramm. Remote Sens.*, 209:500–513, 2024.

[34] Da He, Qian Shi, Jingqian Xue, Peter M. Atkinson, and Xiaoping Liu. Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation. *Remote Sens. Environ.*, 299:113884, 2023.

[35] Haixu He, Jining Yan, Dong Liang, Zhongchang Sun, Jun Li, and Lizhe Wang. Time-series land cover change detection using deep learning-based temporal semantic segmentation. *Remote Sens. Environ.*, 305:114101, 2024.

[36] Yongjun He, Jinfei Wang, Ying Zhang, and Chunhua Liao. An efficient urban flood mapping framework towards disaster response driven by weakly supervised semantic segmentation with decoupled training samples. *ISPRS J. Photogramm. Remote Sens.*, 207:338–358, 2024.

[37] Danfeng Hong, Bing Zhang, Hao Li, Yuxuan Li, Jing Yao, Chenyu Li, Martin Werner, Jocelyn Chanussot, Alexander Zipf, and Xiao Xiang Zhu. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation us-

ing high-resolution domain adaptation networks. *Remote Sens. Environ.*, 299:113856, 2023.

[38] Seyyed Shahabaddin Hosseini Dehshiri and Bahar Firoozabadi. Sustainable solar energy in urban areas: A novel framework for uncertainty, reliability, and 7e analysis. *J. Clean. Prod.*, 444:141136, 2024.

[39] Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. Global streetscapes — a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS J. Photogramm. Remote Sens.*, 215:216–238, 2024.

[40] Laura Jo Houchmand, Marcel Macarulla Martí, and Santiago Gassó-Domingo. Photovoltaics and green roofs: Holistic analysis in built environments. *Renew. Sustain. Energy Rev.*, 207:114987, 2025.

[41] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *Proc. ICLR*. OpenReview.net, 2018.

[42] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *Int. J. Comput. Vis.*, 130(2):316–343, 2022.

[43] Jeffrey Huang, Simon Elias Bibri, and Paul Keel. Generative spatial artificial intelligence for sustainable smart cities: A pioneering large flow model for urban digital twin. *Environ. Sci. Ecotechnol.*, 24:100526, 2025.

[44] Ali Jamali, Swalpa Kumar Roy, Danfeng Hong, Peter M. Atkinson, and Pedram Ghamisi. Spatial-gated multilayer perceptron for land use and land cover mapping. *IEEE Geosci. Remote. Sens. Lett.*, 21:1–5, 2024.

[45] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single- and multi-modal data. In *Proc. ICCV*, pages 590–599. IEEE, 2021.

[46] K. J. Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N. Balasubramanian. Novel class discovery without forgetting. In *Proc. ECCV*, volume 13684, pages 570–586. Springer, 2022.

[47] Firoozeh Karimi and Selima Sultana. Urban expansion prediction and land use/land cover change modeling for sustainable urban development. *Sustainability*, 16(6), 2024.

[48] Michael Kölle, Dominik Laupheimer, Stefan Schmohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. The hessigheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. *ISPRS Open J. Photogramm. Remote Sens.*, 1:100001, 2021.

[49] Loïc Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proc. CVPR*, pages 4558–4567, 2018.

[50] Dilong Li, Jing Zhao, Chongkei Chang, Ziyi Chen, and Jixiang Du. Lgmamba: Large-scale ALS point cloud semantic segmentation with local and global state-space model. *IEEE Geosci. Remote. Sens. Lett.*, 22:1–5, 2025.

[51] Huchen Li, Haiyan Guan, Lingfei Ma, Xiangda Lei, Yongtao Yu, Hanyun Wang, Mahmoud Reza Delavar, and Jonathan Li. Mvpnet: A multi-scale voxel-point adaptive fusion network for point cloud semantic segmentation in urban scenes. *Int. J. Appl. Earth Obs. Geoinf.*, 122:103391, 2023.

[52] Maosu Li, Yijie Wu, Anthony G. O. Yeh, and Fan Xue. HRHD-HK: A benchmark dataset of high-rise and high-density urban scenes for 3d semantic segmentation of photogrammetric point clouds. 1:3714–8, 2023.

[53] Mengtian Li, Shaohui Lin, Zihan Wang, Yunhang Shen, Baochang Zhang, and Lizhuang Ma. Class-imbalanced semi-supervised learning for large-scale point cloud semantic segmentation via decoupling optimization. *Pattern Recognit.*, 156:110701, 2024.

[54] Wenbin Li, Zhichen Fan, Jing Huo, and Yang Gao. Modeling inter-class and intra-class constraints in novel class discovery. In *Proc. CVPR*, pages 3449–3458. IEEE, 2023.

[55] Yansheng Li, Yuhan Zhou, Yongjun Zhang, Liheng Zhong, Jian Wang, and Jingdong Chen. Dkdfn: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. *ISPRS J. Photogramm. Remote Sens.*, 186:170–189, 2022.

[56] Chun Liu, Doudou Zeng, Akram Akbar, Hangbin Wu, Shoujun Jia, Zeran Xu, and Han Yue. Context-aware network for semantic segmentation toward large-scale point clouds in urban environments. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–15, 2022.

[57] Jiaxu Liu, Zhengdi Yu, Toby P. Breckon, and Hubert P. H. Shum. U3ds^3: Unsupervised 3d semantic scene segmentation. In *proc. WACV*, pages 3747–3756. IEEE, 2024.

[58] Songlin Liu, Linwei Chen, Li Zhang, Jun Hu, and Ying Fu. A large-scale climate-aware satellite image dataset for domain adaptive land-cover semantic segmentation. *ISPRS J. Photogramm. Remote Sens.*, 205:98–114, 2023.

[59] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel CNN for efficient 3d deep learning. In *NeurIPS*, pages 963–973, 2019.

[60] Hélder Silva Lopes, Diogo Guedes Vidal, Nadhima Cherif, Lígia Silva, and Paula C. Remoaldo. Green infrastructure and its influence on urban heat island, heat risk, and air pollution: A case study of porto (portugal). *J. Environ. Manage.*, 376:124446, 2025.

[61] Dening Lu, Linlin Xu, Jun Zhou, Kyle (Yilin) Gao, and Jonathan Li. 3dlst: 3d learnable supertoken transformer for lidar point cloud scene segmentation. *Int. J. Appl. Earth Obs. Geoinf.*, 140:104572, 2025.

[62] Dening Lu, Jun Zhou, Kyle (Yilin) Gao, Jing Du, Linlin Xu, and Jonathan Li. Dynamic clustering transformer network for point cloud segmentation. *Int. J. Appl. Earth Obs. Geoinf.*, 128:103791, 2024.

[63] Hao Lu, Bowen Li, Gang Yang, Guangpeng Fan, Han Wang, Yong Pang, Zuyuan Wang, Yining Lian, Haifeng Xu, and Huan Huang. Towards a point cloud understanding framework for forest scene semantic segmentation across forest types and sensor platforms. *Remote Sens. Environ.*, 318:114591, 2025.

[64] Yuchi Ma, Shuo Chen, Stefano Ermon, and David B. Lobell. Transfer learning in environmental remote sensing. *Remote Sens. Environ.*, 301:113924, 2024.

[65] Iaroslav Melekhov, Anand Umashankar, Hyeong-Jin Kim, Vladislav Serkov, and Dusty Argyle. ECLAIR: A high-fidelity aerial lidar dataset for semantic segmentation. In *Proc. CVPR Workshops*, pages 7627–7637. IEEE, 2024.

[66] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *Proc. IV*, pages 687–693. IEEE, 2020.

[67] Hanh Nguyen Pham, Kinh Bac Dang, Thanh Vinh Nguyen, Ngoc Cuong Tran, Xuan Quy Ngo, Duc Anh Nguyen, Thi Thanh Hai Phan, Thu Thuy Nguyen, Wenshan Guo, and Huu Hao Ngo. A new deep learning approach based on bilateral semantic segmentation models for sustainable estuarine wetland ecosystem management. *Sci. Total Environ.*, 838:155826, 2022.

[68] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. CVPR*, pages 77–85. IEEE, 2017.

[69] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proc. NeurIPS*, pages 5099–5108, 2017.

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, volume 139, pages 8748–8763, 2021.

[71] Luigi Riz, Cristiano Saltori, Elisa Ricci, and Fabio Poiesi. Novel class discovery for 3d point cloud semantic segmentation. In *Proc. CVPR*, pages 9393–9402. IEEE, 2023.

[72] Damien Robert, Hugo Raguet, and Loïc Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *Proc. ICCV*, pages 17149–17158. IEEE, 2023.

[73] Subhankar Roy, Mingxuan Liu, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Class-incremental novel class discovery. In *Proc. ECCV*, volume 13693 of *Lecture Notes in Computer Science*, pages 317–333. Springer, 2022.

[74] Muhamad Risqi U. Saputra, Irfan Dwiki Bhaswara, Bahrul Ilmi Nasution, Michelle Ang Li Ern, Nur Laily Romadhotul Husna, Tahjudil Witra, Vicky Feliren, John R. Owen, Deanna Kemp, and Alex M. Lechner. Multi-modal deep learning approaches to semantic segmentation of mining footprints with multispectral satellite imagery. *Remote Sens. Environ.*, 318:114584, 2025.

[75] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. In *Proc. CVPR*, pages 10526–10535. IEEE, 2020.

[76] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pac. J. Math.*, 21(2):343–348, 1967.

[77] Yuan Sun, Yiqun Wu, Huifang Yu, and Yonghua Li. Spatiotemporal evolution of urban landscapes in chinese historic water towns (1918–2021). *Landscape Res.*, 49(4):568–583, 2024.

[78] Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In *Proc. CVPR*, pages 797–806. IEEE, 2020.

[79] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proc. ICCV*, pages 6410–6419. IEEE, 2019.

[80] Ali Tighnavard Balasbaneh, Willy Sher, Aziman Madun, and Ashraf Ashour. Life cycle sustainability assessment of alternative green roofs – a systematic literature review. *Build. Environ.*, 248:111064, 2024.

[81] Colin Troisemaine, Vincent Lemaire, Stéphane Gosselin, Alexandre Reiffers-Masson, Joachim Flocon-Cholet, and Sandrine Vaton. Novel class discovery: an introduction and key concepts. *CoRR*, abs/2302.12028, 2023.

[82] Nina M. Varney, Vijayan K. Asari, and Quinn Graehling. DALES: A large-scale aerial lidar data set for semantic segmentation. In *Proc. CVPR Workshops*, pages 717–726, 2020.

[83] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Unsupervised semantic segmentation of urban scenes via cross-modal distillation. *Int. J. Comput. Vis.*, pages 1–23, 2025.

[84] C. Wang, S. Leisz, L. Li, X. Shi, J. Mao, Y. Zheng, and A. Chen. Historical and projected future runoff over the mekong river basin. *Earth Syst. Dyn.*, 15(1):75–90, 2024.

[85] Junjue Wang, Ailong Ma, Yanfei Zhong, Zhuo Zheng, and Liangpei Zhang. Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery. *Remote Sens. Environ.*, 277:113058, 2022.

[86] Libo Wang, Rui Li, Dongzhi Wang, Chenxi Duan, Teng Wang, and Xiaoliang Meng. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote. Sens.*, 13(16):3065, 2021.

[87] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.*, 38(5):146:1–146:12, 2019.

[88] Marc Wieland, Sandro Martinis, Ralph Kiefl, and Veronika Gstaiger. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sens. Environ.*, 287:113452, 2023.

[89] Fan Wu, Chao Wang, Hong Zhang, Juanjuan Li, Lu Li, Weirong Chen, and Bo Zhang. Built-up area mapping in china from gf-3 sar imagery based on the framework of deep learning. *Remote Sens. Environ.*, 262:112515, 2021.

[90] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: simpler, faster, stronger. In *Proc. CVPR*, pages 4840–4851. IEEE, 2024.

[91] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer V2: grouped vector attention and partition-based pooling. In *Proc. NeurIPS*, 2022.

[92] Binbin Xiang, Maciej Wielgosz, Theodora Kontogianni, Torben Peters, Stefano Puliti, Rasmus Astrup, and Konrad Schindler. Automated forest inventory: Analysis of high-density airborne lidar point clouds with 3d deep learning. *Remote Sens. Environ.*, 305:114078, 2024.

[93] Xiaoyang Xiang, Zhihong Zhai, Chengliang Fan, Yunfei Ding, Lifei Ye, and Jiangbo Li. Modelling future land use land cover changes and their impacts on urban heat island intensity in guangzhou, china. *J. Environ. Manage.*, 366:121787, 2024.

[94] Aoran Xiao, Jiaxing Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El-Saddik, Shijian Lu, and Eric P. Xing. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *Proc. CVPR*, pages 9382–9392. IEEE, 2023.

[95] Muli Yang, Liancheng Wang, Cheng Deng, and Hanwang Zhang. Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In *Proc. CVPR*, pages 3459–3468. IEEE, 2023.

[96] Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, and Cheng Deng. Divide and conquer: Compositional experts for generalized novel class discovery. In *Proc. CVPR*, pages 14248–14257. IEEE, 2022.

[97] Qiquan Yang, Yi Xu, TC Chakraborty, Meng Du, Ting Hu, Ling Zhang, Yue Liu, Rui Yao, Jie Yang, Shurui Chen, Changjiang Xiao, Renrui Liu, Mingjie Zhang, and Rui Chen. A global urban heat island intensity dataset: Generation, comparison, and analysis. *Remote Sens. Environ.*, 312:114343, 2024.

[98] Zelin Zang, Lei Shang, Senqiao Yang, Fei Wang, Baigui Sun, Xuansong Xie, and Stan Z. Li. Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In *Proc. ICCV*, pages 11824–11833. IEEE, 2023.

[99] Ziyin Zeng, Qingyong Hu, Zhong Xie, Bijun Li, Jian Zhou, and Yongyang Xu. Small but mighty: Enhancing 3d point clouds semantic segmentation with u-next framework. *Int. J. Appl. Earth Obs. Geoinf.*, 136:104309, 2025.

[100] Ziyin Zeng, Huan Qiu, Jian Zhou, Zhen Dong, Jinsheng Xiao, and Bijun Li. Pointnat: Large-scale point cloud semantic segmentation via neighbor aggregation with transformer. *IEEE Trans. Geosci. Remote. Sens.*, 62:1–18, 2024.

[101] Huang Zhang, Xin Ning, Changshuo Wang, Enhao Ning, and Lusi Li. Deformation depth decoupling network for point cloud domain adaptation. *Neural Netw.*, 180:106626, 2024.

[102] Yachao Zhang, Yuxiang Lan, Yuan Xie, Cuihua Li, and Yanyun Qu. Cross-cloud consistency for weakly supervised point cloud semantic segmentation. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–12, 2025.

[103] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, pages 22982–22994, 2021.

[104] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proc. ICCV*, pages 16239–16248. IEEE, 2021.

[105] Yuyang Zhao, Zhun Zhong, Nicu Sebe, and Gim Hee Lee. Novel class discovery in semantic segmentation. In *Proc. CVPR*, pages 4330–4339. IEEE, 2022.

[106] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proc. CVPR*, pages 10867–10875. IEEE, 2021.

[107] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proc. CVPR*, pages 9462–9470. IEEE, 2021.

[108] Shiqi Zhou, Haifeng Diao, Mo Wang, Weiyi Jia, Yuankai Wang, Zhiyu Liu, Wei Gan, Mimi Zhou, Zhiqiang Wu, and Zichen Zhao. Knowledge mapping and emerging trends of urban resilient infrastructure research in urban studies: Precedent work, current progress and future perspectives. *J. Clean. Prod.*, 452:142087, 2024.

[109] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proc. CVPR*, pages 4490–4499. IEEE, 2018.

[110] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proc. CVPR*, pages 9939–9948. IEEE, 2021.

APPENDICES

Appendix A

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.

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

$$\mathrm{IoU}_i = \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \tag{A.1}$$

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:

$$\mathrm{mIoU}_{\mathrm{Novel}} = \frac{1}{N} \sum_{j=1}^{N} \mathrm{IoU}_j, \tag{A.2}$$

$$\mathrm{mIoU}_{\mathrm{Base}} = \frac{1}{B} \sum_{k=1}^{B} \mathrm{IoU}_k, \tag{A.3}$$

$$\text{mIoU}_{\text{All}} = \frac{1}{C} \sum_{i=1}^{C} \text{IoU}_i, \tag{A.4}$$

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.

Appendix B

Hyperparameter Configurations

This appendix summarizes the hyperparameter configurations for the three proposed methods. Table B.1 presents the configuration for CHNCD (Chapter 3), Table B.2 details the settings for AGDNet (Chapter 4), and Table B.3 lists the parameters for 3DLCDM (Chapter 5).

Table B.1: Hyperparameter configuration for CHNCD.

Parameter	**Value**
Learning rate	1×10^{-2}
Batch size	4
Epochs (S3DIS, Toronto-3D)	500
Epochs (SemanticSTF)	100
Voxel size	0.05
Points per sample	60,000

Table B.2: Hyperparameter configuration for AGDNet.

Parameter	**Value**
Optimizer	SGD
Learning rate	1×10^{-2}
Momentum	0.9
Weight decay	1×10^{-4}
Warm-up epochs	10
Final learning rate	1×10^{-5}
Batch size	4
Worker threads	8
Epochs (Toronto-3D)	500
Epochs (SemanticSTF)	100
Voxel size	0.05
Points per sample	60,000
DropPath initial ratio	0.1
Number of Gaussians	1,000
Loss Weights (linearly increasing)	
Prototype Relation Loss λ_{PR}	0.3 (max)
Contrastive Alignment Loss λ_{CA}	0.2 (max)
Semantic Transfer Loss λ_{ST}	0.1 (max)

Table B.3: Hyperparameter configuration for 3DLCDM.

Parameter	**Value**
Batch size	4
Worker threads	8
Total epochs	100
Voxel size	0.2
Points per sample	40,000
Dynamic weighting momentum α	0.8
Temperature Scheduling	
Initial temperature T_{start}	0.30
Peak temperature T_{peak}	0.50
Final temperature T_{final}	0.05
Linear warm-up duration	$T_{\text{total}}/4$

Appendix C

Copyright Acknowledgment for IEEE Materials

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of *University of Waterloo*'s products or services. Internal or personal use of this material is permitted.

If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please visit IEEE Permissions to learn how to obtain a License from RightsLink.

The use of IEEE copyrighted material in this thesis is in accordance with IEEE's policies and guidelines. The author has ensured that the material is used responsibly, with proper acknowledgment and without any modifications that would misrepresent the original work.