electronics

MDPI

Article

MDGroup: Multi-Grained Dual-Aware Grouping for 3D Point Cloud Instance Segmentation

Wenyun Sun * and Ruifeng Han

School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202312620009@nuist.edu.cn

* Correspondence: wenyunsun@nuist.edu.cn

Abstract

Instance segmentation of 3D point clouds is a fundamental task for scene understanding in applications such as autonomous driving, robotics, and augmented reality. The inherent irregularity and sparsity of point clouds, compounded by scale variations and instance adhesion, pose significant challenges to accurate segmentation. Existing grouping-based methods are often limited by the loss of geometric details in single-path backbones and by error propagation near complex boundaries. To address these issues, a Multi-grained Dual-aware Grouping algorithm (MDGroup) is proposed, which explicitly integrates multi-grained feature representation with dual awareness of class and boundary. The algorithm features a Dual-Resolution 3D U-Net (DRNet) that preserves local geometric details while aggregating global semantics through adaptive alignment. A four-branch prediction scheme enhances semantic and offset estimation with boundary and directional cues, enabling fine-grained boundary modeling. Furthermore, a Hierarchical Adaptive Multi-grained Feature fusion framework (HAMF) achieves efficient cross-scale alignment by combining Class-Aware Dynamic Voxelization and Class-Aware Pyramid Scaling. Finally, a Boundary-Aware Weighted Aggregation mechanism (BAWA) refines instance grouping by dynamically weighting semantic confidence, geometric distance, boundary probability, and directional consistency. To extend the model to dynamic scenes, a Temporal Adaptive Gating (TAG) module is introduced to leverage historical frame correlations. Extensive experiments on the ScanNet v2, S3DIS, STPLS3D, SemanticKITTI, LiDAR-Net, and OCID benchmarks demonstrate that MDGroup achieves state-of-the-art performance among grouping-based methods, particularly on small objects, complex boundaries, and dynamic environments.

Keywords: 3D point cloud instance segmentation; grouping-based methods; dual-resolution network; multi-grained feature fusion; class-aware representation; boundary-aware aggregation

Received: 15 January 2026
Revised: 12 February 2026
Accepted: 17 February 2026
Published: 24 February 2026

1. Introduction

3D point cloud instance segmentation is a core task in scene understanding and serves as a critical foundation for applications such as autonomous driving, robotic navigation, augmented reality, and digital twins. Unlike images, point clouds are inherently sparse and irregular, making it difficult to balance receptive field organization with the preservation of geometric details. Additionally, variations in scale, occlusion, and tightly connected instances introduce further uncertainty in segmentation.

In practical engineering scenarios, grouping-based paradigms are widely adopted due to their scalability and efficiency, demonstrating strong throughput and stability in

large-scale environments. Early grouping-based instance segmentation methods [1–3] typically cluster points with similar semantic labels and close geometric proximity using semantic scores and center offset vectors. More recent approaches, such as SoftGroup [4] and its extended version SoftGroup++ [5], allow each point to be associated with multiple categories and perform clustering based on soft semantic scores. Compared to hard one-hot semantic predictions, this strategy offers improved robustness and accuracy, making SoftGroup one of the most representative grouping-based instance segmentation methods to date.

However, the single-path downsampling U-Net backbone used in these methods irreversibly loses high-frequency geometric details during repeated compression and upsampling. Offset vectors are prone to cross-instance propagation in complex boundary and adjacent regions. Moreover, the misalignment between long-range semantic context and local geometric details creates performance bottlenecks in recognizing small objects and accurately locating boundaries.

Beyond the spatial irregularities observed in static scans, practical applications often involve dynamic scenes and sensors with varying imaging resolutions. Traditional grouping-based methods typically process data as independent spatial frames, overlooking the temporal correlations inherent in continuous sensor streams. This limitation hampers performance in dynamic environments such as autonomous driving. Furthermore, existing feature fusion strategies often ignore the physical imaging resolution of optical sensors and the varying quality of signal features across different granularities, leading to suboptimal feature alignment and noise accumulation.

To overcome these limitations, this paper proposes a novel Multi-grained Dual-aware Grouping algorithm (MDGroup) for point cloud instance segmentation. The backbone design introduces a Dual-Resolution 3D U-Net (DRNet), which retains a downsampled semantic branch for global context aggregation and adds a parallel high-resolution geometric branch. Sparse convolution is applied only to non-empty voxels to prevent sparsity diffusion, explicitly preserving local geometric information that is particularly sensitive to small objects and complex boundaries. To mitigate the channel and memory overhead introduced by the high-resolution geometric stream, DRNet employs adaptive channel compression and spatial alignment strategies. These mechanisms enable lightweight fusion of global semantics and local geometry at the same scale, generating multi-granularity features that enhance semantic consistency while maintaining controllable computational cost.

Building on this foundation, MDGroup extends the semantic and offset branches of SoftGroup by introducing boundary and direction branches. This allows the network to explicitly estimate the probability of geometric boundaries between instances and predict direction vectors pointing toward instance interiors in boundary regions. The multi-granularity features are processed through four parallel branches to generate semantic, offset, boundary, and direction predictions, providing complementary information for subsequent feature fusion and aggregation.

Considering the limited receptive field of high-resolution features and the memory cost of multi-level high-resolution branches, this paper further proposes a Hierarchical Adaptive Multi-grained Feature fusion framework (HAMF). HAMF includes a Class-Aware Dynamic Voxelization (CADV) module which incorporates sensor imaging resolution constraints for extracting class-level local structural features and a Class-Aware Pyramid Scaling (CAPS) module for capturing scene-level global semantic features. These features are progressively injected into the high-resolution geometric stream and aligned element-wise with local geometric details at corresponding scales. Compared to directly stacking additional high-resolution branches, this hierarchical cross-scale injection continuously supplies long-range context to the detail stream at lower cost, significantly reducing sensitivity to local noise and

improving boundary stability in complex topologies and long-range dependency scenarios. This framework further integrates a signal quality-aware fusion strategy to balance feature activation magnitude and signal confidence.

To address missegmentation caused by adjacent instances at the aggregation level, this paper introduces a Boundary-Aware Weighted Aggregation mechanism (BAWA). BAWA incorporates boundary and direction information from the four-branch predictions. Boundary probability is used to suppress cross-instance propagation, while direction vectors in boundary regions guide consistent aggregation toward instance interiors. This complements the offset branch's guidance toward instance centers and reduces incorrect clustering and boundary omissions at the source. In terms of supervision, class imbalance and directional consistency losses are designed for boundary and direction learning, ensuring stable convergence of the main task and enabling joint optimization of global semantics, local geometry, and boundary priors.

To validate the effectiveness and generalizability of MDGroup, extensive evaluations are conducted on three challenging 3D instance segmentation datasets: ScanNet v2, S3DIS, and STPLS3D. MDGroup achieves significant improvements over existing grouping-based methods. Ablation studies further demonstrate that the proposed backbone architecture, hierarchical fusion framework, and aggregation strategy effectively address performance bottlenecks in complex boundaries and small object scenarios, offering new insights for future backbone and aggregation mechanism designs in point cloud instance segmentation.

In summary, the contributions of this work are as follows:

- A Dual-Resolution 3D U-Net, DRNet, is designed to preserve local geometric details while extracting global semantic context. It introduces boundary and direction branches alongside semantic and offset branches to form a four-branch prediction structure that supports feature fusion and aggregation that supports spatial-temporal feature fusion. The TAG module is designed to extend the architecture to dynamic scenes by effectively aggregating historical temporal cues while suppressing outdated information.
- A Hierarchical Adaptive Multi-grained Feature fusion framework, HAMF, is proposed. It incorporates a sensor resolution-aware CADV module for class-level local structure extraction and CAPS for global semantic feature extraction. A signal quality-aware gating mechanism is further integrated to adaptively fuse these features with high-resolution geometric details based on feature activation confidence.
- A Boundary-Aware Weighted Aggregation mechanism, BAWA, is introduced, which integrates semantic confidence, geometric distance, boundary probability, and directional consistency through a dynamic multi-factor weighting function. This enables fine-grained boundary refinement and robust intra-instance aggregation while suppressing cross-instance errors.
- A novel Multi-grained Dual-aware Grouping algorithm, MDGroup, is proposed and extensively evaluated across six benchmarks covering static, dynamic, and real-world scenarios. The results show significant improvements over state-of-the-art grouping-based methods.

The remainder of this paper is organized as follows. Section 2 reviews recent advances in point cloud instance segmentation. Section 3 details the proposed model and its components. Section 4 presents comparative and ablation experiments. Finally, Section 5 concludes the paper.

2. Related Work

3D instance segmentation aims to identify individual object instances within a point cloud and assign a semantic label to each point. Unlike semantic segmentation, which only distinguishes different classes, instance segmentation further differentiates between

instances of the same class. Existing methods can be broadly categorized into four groups: proposal-based, transformer-based, kernel-based, and grouping-based approaches.

2.1. Proposal-Based Methods

Proposal-based methods follow a top-down segmentation paradigm by decomposing instance segmentation into two sub-tasks: 3D object detection and instance mask prediction. These approaches first generate 3D region proposals and then refine and segment the point cloud within each proposal. This strategy was initially inspired by the success of 2D Mask R-CNN [6], where 3D Region Proposal Network (3D-RPN) and 3D Region-of-Interest (3D-RoI) layers are employed to predict bounding box locations, class labels, and instance masks. For example, the 3D-SIS network proposed by Hou et al. [7] integrates multi-view RGB image features with 3D geometric features for joint semantic and instance segmentation.

An alternative direction was introduced by Yi et al. [8], who proposed GSPN, which adopts an analysis-by-synthesis strategy to generate high-quality 3D object proposals for each potential instance, which are further refined using a region-based PointNet [9]. Similarly, Liu et al. [10] introduced GICN, which estimates object centers and scales using Gaussian distributions and performs probabilistic sampling of proposals to regress bounding boxes and instance masks. Engelmann et al. [11] proposed 3D-MPA, which predicts instance centers and employs graph convolutional networks to aggregate features from surrounding points, enhancing proposal representation. Some methods also discard anchors and post-processing entirely; for example, Yang et al. [12] proposed 3D-BoNet, which directly regresses coarse 3D bounding boxes for all instances from global features extracted by PointNet++ [13]. It then performs point-wise binary classification to segment instance masks within proposals. The bounding box generation is formulated as an optimal matching problem and solved using the Hungarian algorithm [14], with multi-objective loss functions designed to regularize the predicted boxes. Recently, Kolodiazhnyi et al. [15] proposed TD3D, a point cloud instance segmentation method based entirely on sparse convolution. The approach first detects axis-aligned bounding boxes, then extracts point-level features within each box, followed by binary classification.

For large-scale outdoor LiDAR point clouds, Zhang et al. [16] proposed a network that utilizes self-attention modules to learn features in the bird's-eye view (BEV) and subsequently predicts horizontal centers and height boundaries to obtain instance labels. In the context of indoor scene layout prediction, Shi et al. [17] introduced a Variational Denoising Recursive Autoencoder (VDRAE) with hierarchical awareness, which iteratively aggregates contextual information to generate and refine object proposals.

Although proposal-based methods are intuitive and yield high-quality instance segmentation results, they typically involve multi-stage training procedures, require pruning of redundant proposals, and introduce considerable computational and time overhead due to their complex processing pipelines.

2.2. Transformer-Based Methods

Transformer-based methods typically model each object instance as a learnable instance query and directly decode the semantic and instance labels of each point in the point cloud, without relying on explicit proposal generation. The Transformer was originally proposed by Vaswani et al. [18] in the field of natural language processing, and later inspired a series of successful vision models, such as ViT [19] for image classification, DETR [20] for 2D object detection, and Mask2Former [21,22] for 2D segmentation.

However, the application of Transformers to 3D point clouds began relatively late. Early works primarily focused on 3D object detection [23–25] or semantic segmentation [26–28], often requiring dedicated adjustments to address the high computational

cost of attention mechanisms. Although some studies [23,24] attempted to incorporate vanilla Transformers in feature backbones or refinement stages, or used attention to enlarge the receptive field in the bottleneck layers [29], it was not until the emergence of Mask3D and SPFormer that fully Transformer-based architectures were adopted for end-to-end 3D instance segmentation.

Schult et al. [30] proposed Mask3D, which refines instance queries using masked cross-attention on hierarchical point features, enabling the queries to focus on specific object instances. Sun et al. [31] introduced SPFormer, which relies on precomputed superpoint features and iteratively updates queries through a Transformer decoder, significantly reducing computational overhead. Recognizing the limitations of mask-centric attention, Lai et al. [32] designed MAFT, which replaces the mask-based query initialization with position queries derived from object localization, thereby improving the precision of instance-aware attention. Building upon Mask3D, Lu et al. [33] proposed QueryFormer, which introduces a query initialization module and an auxiliary Transformer decoder to suppress background noise and enhance the quality of the final predictions.

Transformer-based instance segmentation methods demonstrate powerful feature learning capabilities through a unified attention mechanism. However, their performance often hinges on query initialization strategies that rely on farthest point sampling or random selection, which limits the model's ability to dynamically adapt to varying numbers of instances. The quadratic complexity of attention computation further necessitates the use of local feature aggregation or spatial constraints to remain computationally feasible. Moreover, the high data dependency of these methods poses challenges for generalization in few-shot scenarios. The strong reliance on the quality of initial queries, coupled with the high computational overhead, highlights two pressing issues in this line of research: improving data-efficient learning and developing more effective query assignment mechanisms.

2.3. Kernel-Based Methods

Kernel-based methods typically begin by generating or assigning a set of dynamic convolutional kernels [34] from the scene, where each kernel encodes the geometric and semantic information of a specific instance. These kernels then aggregate point features via convolution operations to predict instance masks. DyCo3D [29] establishes the foundational framework by employing a lightweight 3D U-Net for feature extraction and adopting PointGroup's [1] coarse clustering algorithm to construct discriminative kernels. PointInst3D [35] enhances flexibility by directly selecting sampling points as candidate kernels using farthest point sampling. DKNet [36] introduces a candidate localization strategy that identifies local maxima from a center heatmap as candidate kernels and aggregates their neighboring information to enhance discrimination, thereby generating more distinguishable instance kernels. ISBNet [37] further proposes an instance-aware sampling encoder that enables fast generation and robust optimization of dynamic kernels on sampled points.

Compared with other paradigms, kernel-based methods replace clustering with farthest point sampling, candidate localization, or high-recall instance-aware point sampling, and also circumvent the challenge of query initialization in Transformer-based approaches. However, a central challenge lies in the discriminative power of the convolutional kernels, which heavily depends on the sampling quality. Inefficient sampling may lead to feature ambiguity among kernels, and the reliance on prior localization limits the adaptability to complex instance geometries.

2.4. Grouping-Based Methods

Grouping-based methods typically adopt a bottom-up pipeline. These approaches begin by predicting low-level attribute information for each point in the point cloud, such

as semantic class labels, geometric offset vectors pointing to instance centers, or latent feature embeddings used to distinguish between different instances. Then, based on the predicted information, the algorithm aggregates and groups points with similar properties—such as converging offset targets, similar embeddings, or identical semantic labels—into clusters. Finally, independent instance masks are generated through filtering and merging of these clusters.

Early works in this class include SGPN proposed by Wang et al. [38], which constructs a feature similarity matrix between points and performs grouping based on feature similarity to form instances. Pham et al. [39] introduced JSIS3D, which utilizes a multi-value conditional random field model to jointly optimize semantic and instance labels, enabling instance acquisition. Lahoud et al. [40] proposed MTML, which learns both feature embeddings and directional embeddings: it first applies mean-shift clustering on the feature embeddings to generate object fragments and then scores and filters these fragments based on directional feature consistency. Han et al. [41] introduced OccuSeg, which performs graph-based clustering guided by object occupancy signals to achieve more precise segmentation. Zhang et al. [42] explored a probabilistic approach by modeling each point as a 3D Gaussian distribution and obtaining instances through distributional clustering.

Recent approaches have introduced various innovations in grouping strategies. A representative method is PointGroup by Jiang et al. [1], which predicts a 3D offset vector from each point to its corresponding instance center and performs clustering on a dual set of coordinates—original points and offset-shifted points. A scoring network, ScoreNet, is then applied to evaluate and filter the candidate clusters, gradually grouping nearby points with the same semantic label. Liang et al. [2] proposed SSTNet, which adopts a different structure by constructing a superpoint-based tree network from precomputed superpoints within the scene. Instances are then generated efficiently through node traversal and splitting. Chen et al. [3] extended PointGroup with a hierarchical clustering strategy in their HAIS framework, which merges fragmented regions around large clusters and refines instance boundaries based on intra-instance predictions. To address the impact of semantic prediction inaccuracies on grouping, Vu et al. [4] introduced SoftGroup, which replaces hard semantic labels with soft semantic scores. This allows each point to associate with multiple clusters representing different semantic categories. The subsequent refinement phase enhances positive samples and suppresses negative ones, thereby mitigating errors caused by semantic ambiguity. Its extended version, SoftGroup++ [5], further improves algorithmic efficiency and search space utilization, significantly accelerating the clustering process. Other enhancements include binary clustering strategies [43] and similar techniques designed to optimize the grouping stage.

Despite the significant progress and various innovative strategies achieved by grouping-based methods, they still face several inherent challenges. These methods rely heavily on the quality of per-point center offset predictions and feature similarity, making it difficult to maintain robustness across objects of varying spatial scales. Moreover, such approaches often depend on manually selected or fine-tuned geometric attributes and grouping parameters. In complex and diverse point cloud scenarios, balancing segmentation accuracy with computational efficiency remains a critical challenge for this line of research. It is worth noting that the challenge of preserving fine-grained details amidst environmental interference is not unique to point cloud analysis but is a universal issue in computer vision. A pertinent parallel exists in underwater image semantic segmentation, where medium scattering and absorption cause severe boundary blurring and information degradation, which is analogous to the difficulties introduced by sparsity and irregularity in 3D point clouds. To mitigate such issues, recent research [44] has successfully enhanced the DeepLab v3+ architecture by incorporating multi-scale feature aggregation strategies.

These cross-domain findings underscore the universality of multi-grained representation learning as a pivotal solution for refining boundaries in complex environments and provide broad empirical support for the feature fusion design philosophy adopted in this paper.

3. Materials and Methods

This section details the design of the proposed model. Section 3.1 presents the overall architecture, Section 3.2 introduces the backbone network, Section 3.3 describes the multi-grained feature fusion framework, and Section 3.4 explains the boundary-aware weighted aggregation mechanism.

3.1. Architecture Overview

The proposed MDGroup integrates multi-grained features and achieves 3D point cloud instance segmentation through a dual-aware grouping and clustering mechanism, which incorporates both boundary-aware and class-aware strategies. The overall architecture of MDGroup is illustrated in Figure 1. The algorithm first employs a DRNet to extract dual-grained features that balance global contextual understanding and local geometric details. Subsequently, four prediction branches are used to infer semantic scores, offset vectors, direction vectors, and boundary probabilities. These outputs are then fed into the HAMF to extract and fuse multi-grained features. Within HAMF, CAPS captures global scene-level semantic representations, and CADV generates class-level local structural features. Combined with the local geometric features from DRNet, the resulting multi-grained features are passed to an octree-based *k*-NN neighborhood search and soft grouping module.

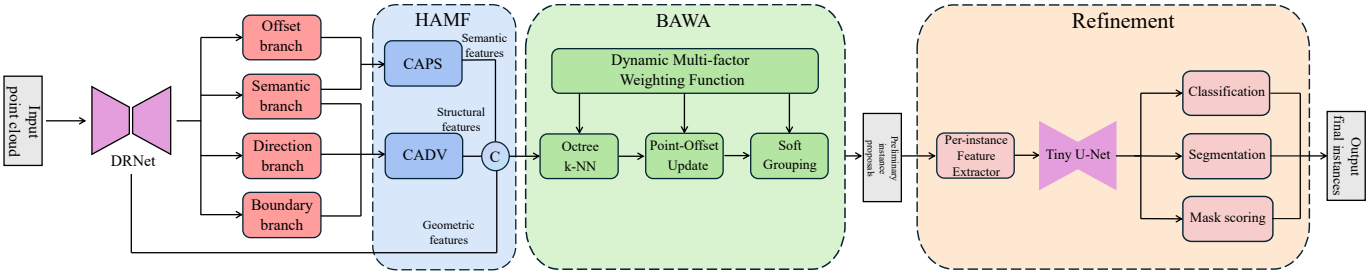

Figure 1. Overview of MDGroup.

Meanwhile, the algorithm introduces a dynamic multi-factor weighting function into the neighborhood relationships to seamlessly integrate the predictions from the four branches. This guides the clustering process through boundary-aware weighted aggregation and a point-guided update mechanism. Instance proposals are generated by adaptively grouping points based on their respective classes. Finally, for each instance, corresponding point features are extracted and refined through a lightweight U-Net, followed by a three-branch prediction head to perform classification, segmentation, and mask scoring, ultimately yielding the final instance proposals.

3.2. Backbone Network

The conventional U-Net [45] is capable of acquiring a large receptive field to capture contextual information; however, fine-grained feature information is inevitably lost during the progressive downsampling process. In consideration of the importance of high-resolution details, this subsection proposes the addition of a parallel high-resolution branch while retaining the downsampling branch. Unlike the downsampling branch, the high-resolution branch focuses on detailed features and is particularly sensitive to small objects and boundary information. Through this high-resolution branch, local geometric

information can be effectively preserved. The dual-resolution branch structure is designed to balance the extraction of global semantics while explicitly retaining fine-grained features, thereby improving the segmentation precision for small objects and complex boundaries.

Each point in the input point cloud data is represented by its coordinates and color and subsequently voxelized to be converted into a sparse voxel network. This voxelized representation is then used as the input to the enhanced backbone network to extract point features. The enhanced backbone network adopts a Dual-Resolution 3D U-Net (DRNet), wherein the downsampling branch retains the original U-Net structure, utilizing Submanifold Sparse Convolution (SubMConv3d) [46] and pooling layers to perform progressive downsampling. In contrast, the high-resolution branch is constructed by stacking multiple SubMConv3d layers without any pooling operations, thereby progressively extracting fine-grained features while maintaining the same spatial resolution as the input voxels. By leveraging the activation suppression property inherent in sparse convolutions, computations are performed exclusively on non-empty voxels, which effectively preserves local geometric details and prevents sparsity diffusion.

Several factors must be considered when introducing a high-resolution branch, including, but not limited to, the substantial computational burden incurred by extracting high-resolution features, particularly in dense prediction tasks such as segmentation. Furthermore, the direct concatenation of features at different resolutions may result in a significant increase in the number of channels, thereby further escalating the computational burden. In this subsection, an adaptive channel compression method is employed, whereby features from both branches are individually compressed via 1×1 convolutions prior to concatenation. Subsequently, the features from the downsampling branch are spatially aligned with those from the high-resolution branch through trilinear interpolation, after which channel concatenation is performed. This approach enables the effective fusion of global and local features, enhancing semantic consistency while simultaneously alleviating the computational burden.

Since individual feature vectors within the high-resolution feature maps possess a limited receptive field, simply concatenating features from the dual branches at the final stage is not the most effective approach for capturing long-range context enriched with high-level semantic information. Although a multi-scale high-resolution branch was initially considered, introducing a three-level branch would have significantly increased memory consumption, thereby violating the principle of efficient processing. Therefore, a multi-scale feature pyramid network (FPN) is constructed within the downsampling branch. Through lateral connections, global features at different scales are upsampled to match the corresponding scales of the high-resolution branch and are subsequently fused with the high-resolution features via element-wise addition, thus achieving multi-level semantic enhancement.

In addition to the aforementioned designs, Dice Loss is introduced at the final layer of the high-resolution branch as an auxiliary supervision to address class imbalance in boundary prediction tasks, while maintaining the cross-entropy loss on the main branch for semantic segmentation in order to avoid objective conflict between the two branches. This approach is adopted because directly imposing cross-entropy loss on the high-resolution branch may lead to premature overfitting, thereby interfering with the training of the main task. Furthermore, to inject global contextual information, global features are upsampled to the corresponding scale of the high-resolution branch and fused with the high-resolution features, which effectively prevents the high-resolution branch from becoming trapped in local noise.

In contrast to the single-resolution pipeline employed by the classical U-Net, the proposed approach explicitly decouples the learning processes of semantic and geometric fea-

tures through dual-resolution branches. Compared to multi-stage cascaded structures such as HRNet, the parallel design of the dual-branch architecture mitigates information degradation caused by serial processing, while maintaining computational efficiency through the inherent sparsity and effectiveness of sparse convolution. As a result, the backbone-extracted dual-granularity features significantly improve the segmentation quality along object boundaries, enhance the model's representational capacity, and provide richer candidate region information during RoI-level refinement.

To enhance the model's applicability to dynamic scenarios such as autonomous driving, where temporal signal correlation from optical sensors is critical, we integrate an optional Temporal Adaptive Gating (TAG) module as an extension. This lightweight, pluggable module is designed to operate on the shared low-level features F extracted by the DRNet. For a given point cloud sequence, the TAG module takes the current frame's features F_t and the memory features from the previous frame H_{t-1} as input. It first aligns the historical features to the current frame's coordinate system using the available ego-motion transformation, yielding H'_{t-1}. Subsequently, a self-attention gating mechanism, implemented as a lightweight sparse convolutional layer, computes an adaptive weighting map G_t. This map dynamically determines the contribution of historical information for each point, allowing the model to suppress outdated information in highly dynamic regions while preserving stable features in static areas. The final temporally-aware features F'_t are produced through a gated fusion $F'_t = (1 - G_t) \odot F_t + G_t \odot H'_{t-1}$, and are then passed to the downstream prediction branches. For all experiments conducted on the static datasets presented in this paper, the TAG module is bypassed, ensuring that our reported results are derived purely from the single-frame spatial processing capabilities of the core architecture. This design demonstrates a clear path for extending our method to temporal data without compromising its performance on static scenes.

Subsequently, four parallel branches are constructed on the shared low-level features F: the semantic branch $S = \mathrm{MLP}_{sem}(F)$, the offset branch $O = \mathrm{MLP}_{off}(F)$, the boundary branch $B = \mathrm{MLP}_{bnd}(F)$, and the direction branch $D = \mathrm{MLP}_{dir}(F)$, where MLP denotes the Multi-Layer Perceptron. In addition to the original semantic and offset branches, the boundary and direction branches are newly introduced in this section. The parameter count for these branches is significantly reduced due to the shared low-level features. Similarly, the semantic and offset branches are constructed using two-layer MLPs, where the semantic scores are grouped rather than converted into a single-shot semantic prediction [1,3], thereby preserving the advantages of SoftGroup [4]. It is important to note that the newly generated direction vectors differ from the offset vectors: the offset vectors represent the displacement from each point to the geometric center of its corresponding instance. By learning the offset vectors, points are shifted toward the centers of their respective instances, thereby facilitating more effective SoftGroup within instances. In contrast, the direction vectors learned by the direction branch are based on the boundary information provided by the boundary branch and act on points located near the instance boundaries.

The boundary branch predicts the probability of geometric boundaries between instances, denoted as $B_i \in [0, 1]$, where the binary supervision label $y_i \in 0, 1$ is generated by computing the Euclidean distance from each point to the nearest surface of a different instance. Since boundary points constitute only a small portion of the entire point cloud, a weighted Binary Cross-Entropy (BCE) loss based on class frequency is employed, formulated as

$$\mathcal{L}_{bnd} = -\frac{1}{N} \sum_{i=1}^{N} \left[w_{pos} y_i \log p_i + w_{neg} (1 - y_i) \log(1 - p_i) \right], \tag{1}$$

where $w_{pos} = \frac{1}{f_{pos}}$ and $w_{neg} = \frac{1}{f_{neg}}$, with f_{pos} and f_{neg} denoting the frequencies of boundary points and non-boundary points, respectively. This approach strengthens the learning of sparse boundary points. In the direction branch, a direction map $D \in \mathbb{R}^{N\times 3}$ is predicted, where each row (d_x, d_y, d_z) represents a unit direction vector pointing toward the interior of the corresponding homogeneous region based on boundary information. Through this design, the learned D provides a pointer for each point near the boundary, guiding it toward the interior of its respective instance. In this subsection, cosine similarity loss is adopted to supervise direction consistency, given by

$$\mathcal{L}_{dir} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{D_i \cdot D_i^*}{||D_i||_2||D_i^*||_2}\right), \tag{2}$$

where D_i^* denotes the ground-truth direction. The predicted direction vectors are further fused with the boundary predictions to guide the feature propagation for subsequent segmentation tasks.

The cross-entropy loss and the l_1 regression loss are utilized to train the semantic and offset branches, respectively. The losses are formulated as

$$\mathcal{L}_{sem} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{CE}(s_i, s_i^*), \tag{3}$$

$$\mathcal{L}_{off} = \frac{1}{\sum_{i=1}^{N}\mathbb{I}_{\{\boldsymbol{p}_i\}}}\sum_{i=1}^{N}\mathbb{I}_{\{\boldsymbol{p}_i\}}|\boldsymbol{o}_i - \boldsymbol{o}_i^*|_1, \tag{4}$$

where s^* denotes the semantic label, and $\boldsymbol{o}^*$ denotes the offset label, representing the vector from a point to the geometric center of its corresponding instance (similar to [1–3]). The indicator function $\mathbb{I}_{\{\boldsymbol{p}_i\}}$ is used to determine whether the point $\boldsymbol{p}_i$ belongs to any instance. Accordingly, the multi-task loss of the enhanced backbone network is formulated as

$$\mathcal{L}_{backbone} = \mathcal{L}_{sem} + \lambda_{off}\mathcal{L}_{off} + \lambda_{bnd}\mathcal{L}_{bnd} + \lambda_{dir}\mathcal{L}_{dir}, \tag{5}$$

where λ_* coefficients are empirically determined through sensitivity analysis to balance the contributions of each branch.

3.3. Multi-Grained Feature Fusion Framework

To address the issues of inter-class boundary ambiguity and the loss of class-specific detail commonly observed in existing methods, the Class-Aware Dynamic Voxelization (CADV) module is proposed in this subsection. By statistically analyzing the object size distributions across different classes and introducing dynamic voxel scaling factors, the CADV module enhances boundary precision and semantic consistency. This design enables differentiated feature extraction for objects of varying scales—for instance, small-scale objects (e.g., chairs) are processed using high-resolution voxels to preserve local geometric details, while large-scale objects (e.g., cabinets) are represented with low-resolution voxels to capture global topological structures. In doing so, the CADV module mitigates the coverage bias inherent in conventional fixed voxelization schemes while maintaining computational efficiency.

The distribution of object sizes across different classes is precomputed offline prior to training. Specifically, all samples in the training set are traversed to collect the bounding box dimensions (length, width, height) of each instance for every class. For a given class c, the mean dimensions of all instances are calculated as $\bar{l}_c$, $\bar{w}_c$, and $\bar{h}_c$, which together define the class-specific average size as $d_c = (\bar{l}_c, \bar{w}_c, \bar{h}_c)$. These values are stored in a dictionary $\{c : d_c\}$ for subsequent retrieval during model inference or training.

The offline precomputation provides the initial average size for each class, serving as the starting point for online adjustment. During training, the dimensions of instances in each batch are used to incrementally update the global average sizes, allowing them to better adapt to the current data distribution. Let d_c^{offline} denote the offline average size of class c. In the i-th batch, suppose B instances belonging to class c are observed, with their dimensions denoted as $d_c^{(i)}{}_{i=1}^{B}$. The class-wise online exponential moving average is then updated as

$$d_c^{\text{new}} = \beta \cdot d_c^{\text{old}} + (1 - \beta) \cdot \frac{1}{B} \sum_{i=1}^{B} d_c^{(i)}. \tag{6}$$

Here, $\beta \in [0, 1]$ is the momentum coefficient that controls the proportion of historical information retained and is adaptively adjusted based on batch characteristics. If no instances of class c are present in the current batch, d_c remains unchanged. $d_c^{(i)}$ denotes the size of the i-th instance in the current batch, and B is the number of instances of class c in the batch.

For example, suppose the initial offline average size for a certain class is $d_c^{\text{offline}} = (0.5\text{ m}, 0.5\text{ m}, 0.8\text{ m})$. In the current batch, there are two instances with dimensions $(0.6, 0.5, 0.9)$ and $(0.55, 0.52, 0.85)$. The average size for this batch is $\frac{1}{2}[(0.6 + 0.55)/2, (0.5 + 0.52)/2, (0.9 + 0.85)/2] = (0.575, 0.51, 0.875)$. If $\beta = 0.9$, the updated class-wise average size becomes $d_c^{\text{new}} = 0.9 \cdot (0.5, 0.5, 0.8) + 0.1 \cdot (0.575, 0.51, 0.875) = (0.5075, 0.501, 0.8075)$. Note that d_c^{new} does not represent the typical size of the current batch alone, but rather a weighted average between historical statistics and the current batch's observation.

To ensure the dynamic voxelization scale aligns with the physical limitations of optical sensors we determine the voxel size v_c for each class c through the formulation $v_c = \max(\alpha\, d_c, \rho)$. In this expression α represents the global scaling factor and ρ denotes the effective imaging resolution which reflects the intrinsic sampling density and precision of the point cloud data. Following the established protocols in recent grouping-based methods [5] the value of ρ is chosen based on the typical point density of the scanning modalities to ensure a balance between geometric fidelity and noise suppression. Specifically, ρ is set based on the sensor modality and scanning environment. It is set to 0.015 m for professional indoor laser scanning scenarios to match the high-fidelity data from MLS systems and 0.02 m for standard indoor scenes captured with consumer-grade RGB-D sensors. For outdoor autonomous driving environments we use 0.05 m while a value of 0.33 m is used for large-scale outdoor aerial surveys. These values correspond to the effective resolution at which the respective optical sensors can reliably resolve structural details. By introducing this physically grounded lower bound we prevent the voxelization process from excessively subdividing sparse regions of the point cloud where the sensor lacks sufficient sampling support. This mechanism, in turn, effectively suppresses the noise and uncertainty that often arise during the feature extraction of small objects. For most instances the term $\alpha \cdot d_c$ is significantly larger than ρ ensuring that the core class-aware dynamic scaling logic remains the primary driver of the voxelization process for the majority of the object classes. For example, small objects such as chairs are assigned finer voxel resolutions, while larger objects like cabinets use coarser voxel sizes.

The offline statistics provide stable priors by leveraging large-scale distributional information, while the online updates adaptively calibrate these priors during training based on intra-batch distributions, enabling the model to accommodate data distribution drift. This combination facilitates improved fine-grained perception across object classes of varying sizes. Outlier filtering is applied to discard extreme instance sizes within each class, further enhancing statistical stability.

To overcome the limitations of single-grained feature representation, this subsection constructs a Hierarchical Adaptive Multi-grained Feature fusion framework (HAMF). Within this framework, the Class-Aware Pyramid Scaling (CAPS) module, based on global downsampling, extracts scene-level semantic features to localize instances at a coarse-grained level, while the CADV module generates class-level structural features to reinforce geometric consistency among instances of the same class. While preserving global contextual information, class-specific local features are introduced. Furthermore, the high-resolution branch of the DRNet presented in the previous subsection retains fine-grained geometric details of the original point cloud via stacked SubMConv3d layers. The three-level hierarchical features are integrated through a dynamic weighting fusion mechanism to achieve complementary representation.

Similar to [5], the CAPS module takes as input the point cloud coordinates X, along with voxel-level offset vectors O and semantic scores S, and outputs voxel-level global features F^{global}, which are then interpolated to point-level features F_{global}. The CADV module proposed in this subsection takes as input the point cloud coordinates X, together with voxel-level semantic scores S, boundary probabilities B, and direction vectors Dir, and outputs voxel-level local features F_c^{local}, which are likewise interpolated to point-level features F_{local}. The globally and locally extracted point-level features are then fed into a magnitude-aware lightweight attention fusion module designed to explicitly account for the real-time quality of signals with different granularities. Drawing upon the correlation between feature activation magnitude and signal confidence in deep representations, the proposed method adopts the channel-wise L2 norms as proxies for signal reliability. In this context, weakly activated features typically correspond to ambiguous regions or background noise, whereas strongly activated features carry more discriminative semantic and geometric information. Specifically, the point-wise L2 norms of the global and local features are computed as $M_{\text{global}} = \|F_{\text{global}}\|_2$ and $M_{\text{local}} = \|F_{\text{local}}\|_2$, respectively. To quantify the quality discrepancy between granularities, a bias term $\psi = \tanh(M_{\text{global}} - M_{\text{local}})$ is derived, where the hyperbolic tangent function constrains the value range to $[-1, 1]$ to ensure numerical stability without introducing additional hyperparameters. This quality bias is injected into the gating generation process to modulate the learnable semantic weights, effectively enabling the network to jointly consider content relevance and real-time signal quality. The fusion weight is formulated as

$$\gamma = \sigma\Big(\text{MLP}\Big([F_{\text{global}} \parallel F_{\text{local}}]\Big) + \psi\Big), \tag{7}$$

where σ denotes the element-wise Sigmoid function and $\parallel$ indicates channel-wise concatenation. The multi-grained features are subsequently adaptively fused via

$$F_{\text{fused}} = \gamma \odot F_{\text{global}} + (1 - \gamma) \odot F_{\text{local}}, \tag{8}$$

and then passed through an MLP followed by ReLU activation to yield the final feature representation $F_{\text{final}} \in \mathbb{R}^{N \times D'}$, which is used for subsequent processing. The pseudocode of Algorithm 1 is as follows.

Algorithm 1 HAMF framework

Input: Original point cloud $X \in \mathbb{R}^{N\times 3}$; Voxel-level semantic score $S \in \mathbb{R}^{M\times C}$; Voxel-level offset vector $O \in \mathbb{R}^{M\times 3}$; Voxel-level boundary probability $B \in \mathbb{R}^{M}$; Voxel-level direction vector $Dir \in \mathbb{R}^{M\times 3}$; Class average size dictionary $\{\, c \mapsto d_c \mid c = 1 \dots C\}$; Imaging resolution ρ

Output: Fusion feature $F_{\text{final}} \in \mathbb{R}^{N\times D'}$

// CAPS
$\hat{X}^{(vox)} = X^{(vox)}\left[\text{Voxelize}(X)\right] + O$
$F^{\text{global}} = \text{CAPS}(\hat{X}^{(vox)}, S)$ ▷ $F^{\text{global}} \in \mathbb{R}^{M\times D}$
$F_{\text{global}} = \text{Interpolate}(F^{\text{global}}, X)$ ▷ $F_{\text{global}} \in \mathbb{R}^{N\times D}$
// CADV
for $c = 1 \dots C$ **do**
    $M_c = \{i \mid S[i,c] > \tau\}$
    $X_c = X[M_c]$, $B_c = B[M_c]$, $D_c = Dir[M_c]$
    Direction weighting $\hat{X}_c = X_c + \beta_c \cdot D_c \cdot B_c$
    $v_c = \max(\alpha\, d_c, \rho)$
    $V_c = \text{Voxelize}(\hat{X}_c, v_c)$
    $F_c^{\text{local}} = \text{SubMConv3d}(V_c)$ ▷ $F_c^{\text{local}} \in \mathbb{R}^{M_c\times D}$
end for
$F_{\text{local}} = \text{Interpolate}(\{F_c^{\text{local}}\}, X)$ ▷ $F_{\text{local}} \in \mathbb{R}^{N\times D}$
// Multi-grained dynamic fusion
$M_{\text{global}} = \|F_{\text{global}}\|_2, M_{\text{local}} = \|F_{\text{local}}\|_2$
$\psi = \tanh(M_{\text{global}} - M_{\text{local}})$
$\text{logits} = \text{MLP}([F_{\text{global}} \parallel F_{\text{local}}])$
$\gamma = \sigma(\text{logits} + \psi)$ ▷ $\gamma \in \mathbb{R}^{N\times D}$
$F_{\text{fused}} = \gamma \odot F_{\text{global}} + (1-\gamma) \odot F_{\text{local}}$
$F_{\text{final}} = \text{ReLU}(\text{MLP}(F_{\text{fused}}))$
return F_{final}

3.4. Boundary-Aware Weighted Aggregation

Traditional methods rely solely on geometric proximity and semantic scores, which can lead to incorrect groupings where adjacent instances are spatially close but semantically distinct. To address this issue, this subsection proposes a novel boundary-aware weighted aggregation mechanism, aiming to improve the accuracy of 3D point cloud instance segmentation by reducing cross-instance misgrouping at object boundaries. The proposed mechanism introduces a dynamic multi-factor weighting function into neighborhood aggregation, effectively integrating semantic confidence, geometric distance, directional consistency, and boundary probability. This enables fine-grained processing of boundary points, strengthening the aggregation of points within the same instance while effectively suppressing erroneous connections across different instances. As a result, the approach enhances boundary robustness and clustering accuracy in instance segmentation.

Following the idea adopted in previous point cloud instance segmentation works such as [4,5], which allow each point to be associated with multiple semantic classes to mitigate semantic uncertainty, this subsection adopts a similar strategy. An octree structure is first constructed for the point cloud, and a radius-constrained k-Nearest Neighbors (k-NN) search is performed for each point to obtain a set of neighboring points with potential connectivity, thereby establishing local neighborhood relationships. As pointed out in SoftGroup++ [5], traditional k-NN search often becomes a bottleneck in clustering due to its high computational cost, whereas using an octree structure reduces the time complexity from $O(n^2)$ to $O(n \log n)$. Building upon this, each point i and its neighbors j are iterated over to compute the boundary-aware weighting coefficient W_{ij}, which is then used in subsequent feature aggregation.

The boundary-aware weighting coefficient W_{ij} is formulated through a dynamic multi-factor weighting function that explicitly models semantic consistency, geometric proximity, and directional alignment, enabling robust grouping around object boundaries. The function is defined as

$$W_{ij} = \frac{S_i + S_j}{2} \cdot \exp\left(-\frac{||x_j - x_i||^2}{\sigma_c^2}\right) \cdot [B_i||D_i||_2 \cos(\theta_{ij}^D) + (1 - B_i)||O_i||_2 \cos(\theta_{ij}^O)], \quad (9)$$

where the semantic term $\frac{S_i+S_j}{2}$ promotes the aggregation of points with similar semantic scores while suppressing those with low confidence. The geometric term $\exp\left(-\frac{||x_j-x_i||^2}{\sigma_c^2}\right)$ emphasizes spatially close neighbors to constrain the influence to a local region. The directional term $B_i||D_i||_2 \cos(\theta_{ij}^D) + (1 - B_i)||O_i||_2 \cos(\theta_{ij}^O)$ adjusts the alignment based on boundary probability: it encourages or suppresses feature aggregation according to whether the point lies on a boundary or within an object, thereby refining instance boundaries through direction-aware modulation.

The semantic smoothing term adopts the mean $\frac{S_i+S_j}{2}$ rather than the product $S_i \cdot S_j$, offering a more relaxed strategy for aggregating semantically similar points. This formulation effectively mitigates the influence of low-confidence points, as a single low score does not immediately nullify the weight, thereby preserving partial information. In contrast to SoftGroup [4], which allows each point to be associated with multiple semantic classes, our approach emphasizes intra-group semantic consistency during grouping by directly incorporating semantic scores into the weighting function. As a result, points with high confidence and matching classes are assigned stronger connections, reinforcing semantic coherence in the instance clustering process.

Neighborhood search typically employs a fixed-radius constraint; however, in this subsection, the geometric proximity is dynamically adapted according to the class-specific average size d_c, with the distance decay parameter σ adjusted as $\sigma = \kappa\sqrt[3]{l_c \cdot w_c \cdot h_c}$, where κ is a global scaling factor that ensures a smooth decrease in weight with increasing distance. Building upon the class-aware design of HAMF in the previous subsection, an adaptive neighborhood strategy is adopted for different semantic classes: small-object classes are assigned finer-grained search distances, while large-object classes are given broader radii to improve grouping efficiency and accuracy. This enables the clustering parameters to adjust automatically according to object scale differences, enhancing adaptability across varying object classes.

The dynamic multi-factor weighting function introduces directional alignment and boundary constraints while maintaining semantic consistency. In the directional term, the magnitude of the directional vector, $|D_i|$, reflects the correction strength at boundary points—larger magnitudes indicate higher confidence in the predicted direction. Similarly, the magnitude of the offset vector, $|O_i|$, represents the distance from a point to the instance center—greater values imply a stronger need for intra-instance aggregation correction. These vector magnitudes carry clear physical meanings and act as independent weighting coefficients that modulate their respective contributions. The directional term is controlled by the boundary probability $B_i \in [0, 1]$, which determines whether the weight should rely more on the directional vector D_i or the offset vector O_i: a higher B_i implies a higher likelihood of the point being near an instance boundary, thus the weight emphasizes D_i. Conversely, a lower B_i favors the use of O_i.

The directional similarities are quantified by $\cos(\theta_{ij}^D) = \frac{D_i \cdot D_j}{\|D_i\|\|D_j\|}$ and $\cos(\theta_{ij}^O) = \frac{O_i \cdot O_j}{\|O_i\|\|O_j\|}$, which measure the cosine of the angles between the vector from point i to point j and the directional or offset vectors, respectively, thus capturing angular consistency. Higher cosine values indicate better alignment, leading to stronger reinforcement in the weighting.

When the angle exceeds 90°, the cosine value becomes negative, indicating a repulsive relationship. Specifically, a negative value of $\cos(\theta_{ij}^{D})$ implies that the directions are opposite, suggesting the points may belong to different instances, while a negative $\cos(\theta_{ij}^{O})$ indicates that the offset direction deviates from the instance center, potentially signaling a noisy point.

Based on the aforementioned weights, each point's position is updated through a weighted displacement, referred to in this subsection as the point-guided update mechanism. Specifically, for point i, the displacement is computed as $\Delta x_i = W_{ij}\left(B_i \hat{D}_i + (1 - B_i)\hat{O}_i\right)$, which can be interpreted as the cumulative result of summing over all neighbors $\hat{D}_i = \frac{D_i}{\|D_i\|+\epsilon}$ and $\hat{O}_i = \frac{O_i}{\|O_i\|+\epsilon}$ denote normalized unit vectors. This indicates that points are displaced along a composite vector direction determined by boundary probability and directional alignment information, as illustrated in Figure 2. The proposed mechanism seamlessly integrates four-branch predictions while extending the concept of soft grouping. This design applies varying degrees of offset vectors to points predicted by different branches, bringing points belonging to the same instance closer together in spatial distribution. To mitigate boundary ambiguity caused by object overlap, the boundary branch predicts the boundary probability of each point and combines it with explicit directional vectors predicted by the direction branch. These differentiated directional adjustments enable precise boundary point localization, effectively separating overlapping regions.

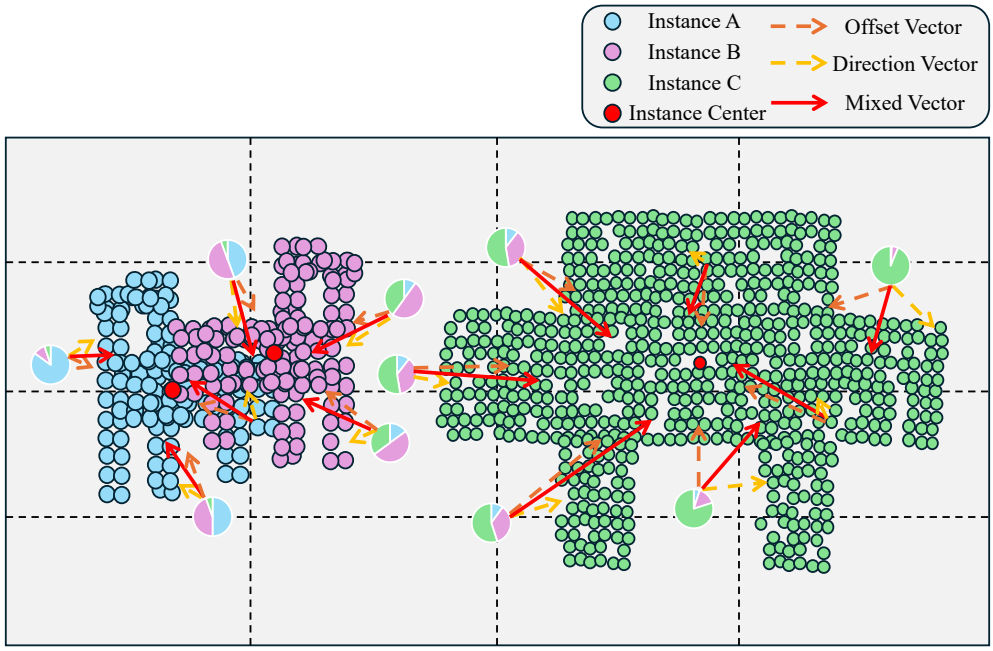

Figure 2. Illustration of the Point-Guided Update Mechanism.

Unlike the simple combination of a single offset vector and a direction vector shown in the figure, the proposed approach allows each point to be associated with multiple semantic categories, resulting in multiple vectors acting on the same point. A dynamic multi-factor weighting function is employed to compute the weight distribution for each point. This, combined with the point-guided update mechanism, further refines boundaries and optimizes spatial distribution, thereby achieving clustering-based semantic prediction and ensuring the aggregation of objects within the same category. Furthermore, by incorporating multi-granularity features fused through HAMF, the method effectively identifies overlapping instances and further distinguishes different instances within the same semantic category.

The point-guided update mechanism serves as a form of **pre-aggregation** in the grouping and clustering process, whereby points belonging to the same instance are drawn closer together through displacement, while boundary points near different instances are nudged toward their respective regions, thereby reducing ambiguity between object boundaries and avoiding incorrect connections across instances. Compared with conventional grouping approaches that directly displace each point toward a predicted center in one step, this weighted and localized displacement strategy enables a smoother and more gradual refinement of point cloud distributions during the clustering iterations. As a result, intra-cluster consistency is enhanced and inter-cluster separability is improved. After the displacement, points within the same instance become spatially closer and are more easily connected via Breadth-First Search (BFS) clustering, while points near the boundaries of different instances are effectively pulled apart, reducing over-clustering errors and improving the overall clustering accuracy.

The boundary-aware weighted aggregation mechanism ingeniously integrates the concepts of boundary awareness and direction-aligned aggregation, explicitly leveraging boundary information to guide clustering boundaries. By combining class-aware neighborhood scaling and grouping strategies, it dynamically adjusts the aggregation scheme according to different object classes. The detailed Algorithm 2 is presented in the following pseudocode.

Algorithm 2 Boundary-Aware Weighted Aggregation

Input: Point cloud X, semantic predictions S, offset predictions O, boundary predictions B, direction predictions D, class-specific average size $\{d_c\}_{c=1}^{C}$, global scaling factor κ, semantic threshold τ_{sem}
Output: Preliminary instance proposals $\mathcal{P}$

octree $=$ BuildOctree(X)
for $i = 1 \ldots N$ **do**
  $c = \arg\max_{c'} S[i, c']$, $(l_c, w_c, h_c) = d_c$, $\sigma_i = \kappa \sqrt[3]{l_c \cdot w_c \cdot h_c}$
end for
$W = \{\}$, $\Delta X = 0_{N \times 3}$
for all $i = 1 \ldots N$ **in parallel do**
  $N_i =$ RadiusSearch$(octree, X[i], \sigma_i)$
  $b = B[i]$, $v_{di} = D[i]$, $v_{oi} = O[i]$
  $m_{di} = \|D[i]\|_2$, $m_{oi} = \|O[i]\|_2$
  for all $j \in N_i$ **do**
    $w_s = (S[i, c] + S[j, c])/2$
    if $w_s > \tau_{\text{sem}}$ **then**
      $d^2 = \|X[j] - X[i]\|^2$, $w_o = e^{-d^2/\sigma_i^2}$
      $v_{dj} = D[j]$, $v_{oj} = O[j]$
      $m_{dj} = \|D[j]\|_2$, $m_{oj} = \|O[j]\|_2$
      $\cos\theta_{ij}^{D} = (v_{di} \cdot v_{dj})/(m_{di} \cdot m_{dj})$, $\cos\theta_{ij}^{O} = (v_{oi} \cdot v_{oj})/(m_{oi} \cdot m_{oj})$
      $W_{ij} = w_s\, w_o \left(b\, m_{di}\, cos_{ij}^{D} + (1 - b)\, m_{oi}\, cos_{ij}^{O}\right)$
      $\hat{d} = v_{di}/(m_{di} + \epsilon)$, $\hat{o} = v_{oi}/(m_{oi} + \epsilon)$
      $\Delta X[i] += W_{ij} \left((b\, \hat{d} + (1 - b)\, \hat{o}\right)$
    end if
  end for
end for
$X = X + \Delta X$
$\mathcal{P} =$ BFS_Cluster$(W > 0)$ ▷ connect edges with $w_{ij} > 0$
return $\mathcal{P}$

4. Experiments

This section presents a comprehensive evaluation of the proposed MDGroup method. We conduct experiments on six representative benchmark datasets, including ScanNet v2, S3DIS, STPLS3D, SemanticKITTI, LiDAR-Net, and OCID to validate performance across diverse environments ranging from static indoor rooms to dynamic outdoor driving scenes. Section 4.1 introduces these datasets, and Section 4.2 details the evaluation metrics. Section 4.3 describes the implementation details. Section 4.4 reports ablation studies to analyze the contribution of individual components. Sections 4.5–4.9 provide comparative experiments on the respective datasets, demonstrating the effectiveness, robustness and generalizability of MDGroup under various conditions, including complex boundaries dynamic sequences and real-world sensor artifacts.

4.1. Datasets

To validate the effectiveness and generalizability of the proposed model, we conduct 3D instance segmentation training and overall performance evaluation on six benchmark datasets that cover a diverse range of environments: ScanNet v2 [47], S3DIS [48], STPLS3D [49], SemanticKITTI [50], LiDAR-Net [51], and OCID [52].

ScanNet is a rich dataset of 3D indoor scenes. It comprises hundreds of diverse room types, such as hotel rooms, libraries, and offices. The ScanNet dataset includes 1613 scans, divided into 1202 for training, 312 for validation, and 100 for testing. Each scene is annotated with semantic and instance-level segmentation labels spanning 18 object categories.

S3DIS is a large-scale indoor dataset that covers six different areas from three distinct buildings on a university campus. It includes 271 scans with semantic and instance-level annotations across 13 object categories. In this paper, we adhere to the standard evaluation protocol by designating Area 5 as the independent validation set for single-scene evaluation and conducting 6-fold cross-validation across all areas to comprehensively verify the model's generalization ability.

STPLS3D is a synthetic outdoor dataset that closely simulates the photogrammetric point cloud generation process from aerial imagery. The dataset densely annotates 14 instance categories over more than 16 square kilometers of urban scenes across 25 regions. The point clouds are divided into non-overlapping 250 m^2 blocks, with each scene containing nearly one million points. Following the protocol in [49], we split the dataset into training and testing subsets for evaluation.

SemanticKITTI is a large-scale dataset for autonomous driving research, providing sequential and densely annotated LiDAR scans derived from the KITTI Vision Benchmark [50]. For our spatio-temporal evaluation, we adhere to the official protocol of the 4D Panoptic Segmentation task [53]. This task requires assigning a semantic label and a spatio-temporally unique instance ID to every point in the test sequences. Following the official guidelines, the evaluation excludes the classes *outlier*, *other-structure*, and *other-object*, focusing on the remaining 25 semantic categories. This benchmark serves to rigorously assess the model's ability to maintain instance consistency in dynamic scenes.

LiDAR-Net is a large-scale indoor dataset composed of raw point clouds captured by a mobile laser scanning system. It contains approximately 3.6 billion points with precise point-level annotations. Distinct from datasets derived from reconstructed meshes LiDAR-Net preserves authentic sensor characteristics. These characteristics include non-uniform point distributions manifesting as scanning holes and lines. Furthermore the dataset meticulously annotates scanning anomalies like reflection noise from specular surfaces and ghosts generated by moving objects.

OCID is another RGBD dataset designed to evaluate segmentation performance in cluttered environments. It was captured using real ASUS-PRO Xtion depth sensors. The dataset comprises 96 scenes that are incrementally built by adding one object at a time creating progressively complex arrangements. OCID uses a variety of common objects from the ARID [54] and YCB [55] datasets. Scene variations include different floor and table textures to simulate diverse indoor conditions. The dataset provides subsets such as ARID20 and YCB10 with up to 20 and 10 objects per scene respectively.

4.2. Evaluation Metrics

To evaluate the accuracy of 3D instance segmentation, we adopt Average Precision (AP), a widely used metric across various computer vision tasks. Specifically, we report mean Average Precision (mAP), mAP_{50}, and mAP_{25}. Here, mAP denotes the average of AP scores computed over multiple Intersection over Union (IoU) thresholds ranging from 50% to 95% with a step size of 5%, while mAP_{50} and mAP_{25} correspond to AP scores at fixed IoU thresholds of 50% and 25%, respectively. To evaluate the computational efficiency and deployment feasibility of the model, we report the number of parameters, inference time and Frames Per Second (FPS). The inference time is measured in milliseconds per scan, while FPS indicates the processing throughput.

In addition, for the evaluation on the S3DIS dataset, we further report mean Precision (mPrec) and mean Recall (mRec) to provide a more comprehensive performance analysis. For the spatio-temporal evaluation on the SemanticKITTI dataset, we adopt the official 4D Panoptic Segmentation metrics. The primary metric, Lidar Segmentation and Tracking Quality (LSTQ), holistically evaluates performance by combining instance segmentation accuracy and tracking consistency into a single score. To further analyze instance-level fidelity, we also report the mean Association Quality (mAQ), which averages the IoU over correctly matched instance pairs. Additionally, we include the mean Intersection over Union (mIoU) as an auxiliary metric to assess the semantic correctness of the points constituting each predicted instance, as this forms a foundation for accurate instance separation. Furthermore, for the cluttered indoor scenes in the OCID dataset, we utilize the F1-Score as a comprehensive metric that balances precision and recall to evaluate segmentation performance in complex physical configurations.

4.3. Implementation Details

The proposed MDGroup framework was implemented using the PyTorch 1.11.0 deep learning framework [56] and trained on a workstation equipped with a single NVIDIA RTX 3090 GPU. To enhance model robustness and mitigate overfitting, we employed standard data augmentation strategies during the training phase, including random rotation, random scaling, elastic distortion, and random flipping. The input point clouds were voxelized with resolutions tailored to each dataset's specific sensor characteristics and scale. For the indoor datasets ScanNet v2, S3DIS and OCID which were captured with consumer-grade RGB-D sensors we used a grid resolution of 0.02 m. Given the higher sensor precision of the LiDAR-Net dataset a finer resolution of 0.015 m was employed to better preserve its detailed geometric structures. For the outdoor datasets a resolution of 0.05 m was utilized for the autonomous driving scenes in SemanticKITTI while a coarser 0.33 m resolution was adopted for the large-scale aerial scans in STPLS3D to balance geometric fidelity with computational efficiency. These voxel resolutions are strictly aligned with the intrinsic sampling density and real-time data output characteristics of optical sensors, ensuring the inference efficiency of MDGroup matches the real-time data stream requirements of practical optical sensor applications. Additionally, we limited the maximum number of input points per scene to 250 k during training to effectively manage GPU memory consumption.

The network parameters were optimized using the Adam optimizer [57] with a batch size of 4. We utilized a Cosine Annealing learning rate scheduler [58] to gradually decay the initial learning rate from 0.001. Regarding the loss function configuration, the weights for the semantic, offset, boundary, and direction branches were empirically set to λ_{off} = 1.0, λ_{bnd} = 0.5, and λ_{dir} = 2.0, respectively, based on the sensitivity analysis presented in the following section. During the inference stage, the global scaling factor κ in the BAWA module was consistently set to 0.3. This setting enables the model to adaptively determine the grouping radius based on instance size, ensuring robust segmentation performance for objects of varying scales.

4.4. Ablation Study

To evaluate the contribution of each proposed component to the overall performance, an ablation study is conducted on the Area 5 split of the S3DIS dataset. The experiment progressively removes individual modules from MDGroup, including DRNet, HAMF, and BAWA. All comparisons are performed under identical experimental settings. Checkmarks (✓) indicate the inclusion of the corresponding module in the network. The results are presented in Table 1.

Table 1. Ablation study on each component of the proposed method in Area 5 of the S3DIS.

Baseline	DRNet	HAMF	BAWA	AP	AP_{50}	AP_{25}
✓	-	-	-	51.6	66.1	71.9
✓	✓	-	-	51.2	68.1	72.6
✓	-	✓	-	51.5	68.5	73.4
✓	-	-	✓	51.4	68.3	73.1
✓	✓	✓	-	51.7	68.6	73.4
✓	✓	-	✓	52.4	68.9	73.9
✓	-	✓	✓	52.0	68.8	73.7
✓	✓	✓	✓	**52.6**	**69.3**	**74.2**

As shown in Table 1, the baseline model SoftGroup achieves 51.6% AP, 66.1% AP_{50}, and 71.9% AP_{25}. When only the backbone is replaced with DRNet, AP increases to 51.2%. This improvement indicates that the high-resolution branch plays a positive role in preserving fine-grained geometric details, particularly for small objects and complex boundaries. Introducing HAMF alone raises AP to 51.5%, demonstrating that multi-granularity features effectively mitigate the sensitivity of high-resolution branches to local noise and enhance global semantic consistency. When BAWA is added independently, AP reaches 51.4%, confirming the importance of explicitly modeling boundary and directional priors for separating adjacent instances.

Combining DRNet with HAMF results in an AP of 51.7%. The combination of DRNet and BAWA further improves AP to 52.4%, showing that structural enhancement and boundary awareness are complementary in boosting performance. The dynamic multi-factor weighting function integrates predictions from four branches and guides clustering through a point shift refinement mechanism, significantly improving instance proposals. The combination of HAMF and BAWA also yields a 0.4 percentage point gain, further validating the synergy between multi-scale semantic class awareness and boundary-aware directional priors. When all three modules are enabled, MDGroup achieves 52.6% AP, 69.3% AP_{50}, and 74.2% AP_{25}. These results represent improvements of 1.9%, 4.8%, and 3.2% over SoftGroup, and the model shows the most stable performance in scenes with complex boundaries and closely adjacent instances.

To further verify the robustness of the proposed method and justify the hyperparameter selection, we conducted comprehensive sensitivity analyses on the ScanNet v2 validation set. We adopted a step-wise orthogonal grid search strategy to determine the optimal multi-task loss weights in Equation (5). First, fixing the backbone weights ($\lambda_{off} = 1.0$), we performed a joint grid search for the auxiliary weights λ_{bnd} and λ_{dir}. As shown in Table 2, the combination of $\lambda_{bnd} = 0.5$ and $\lambda_{dir} = 2.0$ yields the best performance. This confirms that boundary cues require moderate weighting to prevent the high-magnitude weighted BCE loss from dominating the gradients, while direction cues benefit from a larger weight to amplify the naturally small cosine similarity loss. Subsequently, we verified the sensitivity of the offset weight λ_{off} in Table 3, confirming that 1.0 is the optimal balance point for geometric center prediction, as an overly large weight would distort the semantic feature space, while an overly small one would lead to clustering divergence.

Table 2. Joint sensitivity analysis of auxiliary task weights λ_{bnd} and λ_{dir}.

λ_{bnd}	λ_{dir}	AP	AP_{50}	AP_{25}
0	0	49.8	66.8	78.2
0.2	0.5	50.3	67.6	78.9
0.2	1.0	50.5	67.8	79.1
0.2	2.0	50.8	68.2	79.4
0.2	5.0	50.4	67.5	78.8
0.5	0.5	50.6	68.0	79.3
0.5	1.0	50.9	68.4	79.6
0.5	**2.0**	**51.1**	**68.5**	**79.8**
0.5	5.0	50.5	67.7	79.0
1.0	0.5	50.4	67.5	78.8
1.0	1.0	50.7	68.0	79.4
1.0	2.0	50.3	67.2	78.6
1.0	5.0	49.9	66.4	77.9

Table 3. Sensitivity verification of the offset loss weight λ_{off}.

λ_{off}	AP	AP_{50}	AP_{25}
0.1	49.5	65.8	77.5
0.5	50.6	67.9	79.2
1.0	**51.1**	**68.5**	**79.8**
2.0	50.8	68.1	79.5
5.0	49.9	66.5	78.1

Furthermore, we evaluated the impact of the global scaling factor κ in the BAWA module, which controls the neighborhood search radius during inference. As presented in Table 4, we compared the proposed dynamic scaling strategy against a fixed radius baseline (i.e., 'None'). The results indicate that $\kappa = 0.3$ achieves the highest accuracy. Geometrically, this setting corresponds to a search radius of approximately 30% of the instance size, which effectively balances the coverage of intra-instance points with the separation of adjacent instances. Compared to the fixed radius baseline, the dynamic scaling strategy yields an absolute gain of 0.6% in AP_{50}, demonstrating its adaptability to objects of varying scales.

Having established the optimal architectural components and hyperparameter settings, we turn to the computational efficiency of the model. To investigate the overhead introduced by each module, we conduct a detailed complexity analysis in Table 5 by reporting the number of parameters, inference time, and FPS. As observed, the proposed DRNet incurs a moderate increase of 2.5 M parameters and 27 ms in latency compared

to the SoftGroup baseline. This represents a necessary trade-off to accommodate the dual-resolution architecture, which is pivotal for preserving fine-grained geometric details. However, the computational cost remains within a reasonable range thanks to the adaptive channel compression strategy employed in the high-resolution branch. The HAMF module proves to be lightweight, adding negligible overhead to the inference process. Crucially, while the BAWA mechanism involves complex multi-factor weighting calculations, its impact on inference speed is significantly mitigated by the integration of an Octree-based k-NN search. This efficient data structure reduces the neighborhood search complexity from $O(n^2)$ to $O(n \log n)$, effectively offsetting the computational burden of the weighting function. Consequently, the full MDGroup model achieves a total inference time of 642 ms, or 1.56 FPS, demonstrating that our method successfully balances substantial performance gains with practical computational feasibility.

Table 4. Sensitivity analysis of the global scaling factor κ.

κ	AP	AP_{50}	AP_{25}
None	50.5	67.9	79.2
0.1	50.1	67.4	78.7
0.2	50.8	68.2	79.5
0.3	**51.1**	**68.5**	**79.8**
0.4	50.9	68.3	79.6
0.5	50.4	67.5	78.9

Table 5. Component-wise complexity and real-time metrics analysis.

Component	Parameters (M)	Inference Time (ms)	FPS
baseline	30.9	588	1.70
+DRNet	33.4	615	1.63
+HAMF	31.5	602	1.66
+BAWA	31.1	608	1.64
MDGroup	34.2	642	1.56

Overall, DRNet enhances the preservation of fine-grained geometric details, improving segmentation accuracy for small objects and intricate boundaries. HAMF strengthens global semantic consistency and reduces sensitivity to local noise through multi-granularity feature injection. BAWA explicitly incorporates boundary and directional priors, significantly improving the separation of adjacent instances through dual-awareness of class and boundary. The integration of all three components not only leads to clearer boundary predictions and stable instance grouping but also maintains computational tractability. As evidenced by the complexity analysis, the proposed designs achieve the best quantitative results with a controlled increase in model size and latency, fully validating their effectiveness and complementarity.

4.5. Comparison with Related Work on S3DIS Dataset

This section presents the evaluation results of the proposed MDGroup method on the S3DIS benchmark dataset, as summarized in Table 6. The best performance for each metric is highlighted in **bold**. All metrics are computed using the official evaluation code provided by [4], which ensures consistency and fairness across all comparisons.

Table 6. The results of 3D instance segmentation on the S3DIS dataset. The performances on both Area-5 and 6-fold cross validation are reported.

Method	mAP	AP_{50}	$mPrec_{50}$	$mRec_{50}$
				Area 5 validation
SGPN [38]	-	-	36.0	28.7
ASIS [59]	-	-	55.3	42.4
PointGroup [1]	-	57.8	61.9	62.1
DyCo3D [29]	-	-	64.3	64.2
SSTNet [2]	42.7	59.3	65.5	64.2
DKNet [36]	-	-	70.8	65.3
HAIS [3]	-	-	71.1	65.0
TD3D [15]	48.6	65.1	74.4	64.8
SoftGroup [4]	51.6	66.1	73.6	66.6
SoftGroup++ [5]	50.9	67.8	73.8	67.6
MDGroup (ours)	**52.6**	**69.3**	**75.3**	**68.8**
				6-fold cross validation
SGPN [38]	-	-	38.2	31.2
ASIS [59]	-	-	63.6	47.5
3D-BoNet [12]	-	-	65.6	47.7
PointGroup [1]	-	64.0	69.6	69.2
HAIS [3]	-	-	73.2	69.4
SSTNet [2]	54.1	67.8	73.5	73.4
SoftGroup [4]	54.4	68.9	75.3	69.8
DKNet [36]	-	-	75.3	71.1
TD3D [15]	56.2	68.2	76.3	74.0
SPFormer [31]	-	69.2	74.0	71.1
SoftGroup++ [5]	56.6	71.3	75.9	**74.4**
MDGroup (ours)	**57.8**	**71.7**	**77.3**	73.2

Figure 3a illustrates the overall performance comparison. MDGroup achieves the highest accuracy among existing grouping-based instance segmentation methods, demonstrating its effectiveness in complex indoor environments. To further assess the model's generalization ability and robustness, two experimental settings are adopted, namely the Area 5 test split and the 6-fold cross-validation. The corresponding results are shown in Figure 3b,c. MDGroup consistently achieves the highest scores in both AP and F1-score, confirming its strong overall performance and reliability.

On the Area 5 test split, MDGroup significantly outperforms the state-of-the-art grouping-based baselines SoftGroup and SoftGroup++. For AP and AP_{50}, MDGroup achieves improvements of 1.9% and 4.8% over SoftGroup, and 3.3% and 2.2% over SoftGroup++. In terms of mPrec and mRec, MDGroup improves by 2.3% and 3.3% compared to SoftGroup, and by 2.0% and 1.8% compared to SoftGroup++.

Under the 6-fold cross-validation setting, MDGroup also performs better across most evaluation metrics, further confirming its stability and generalization capability in diverse indoor scenes. All comparative experiments are conducted under identical configurations without any additional tuning of heuristic hyperparameters.

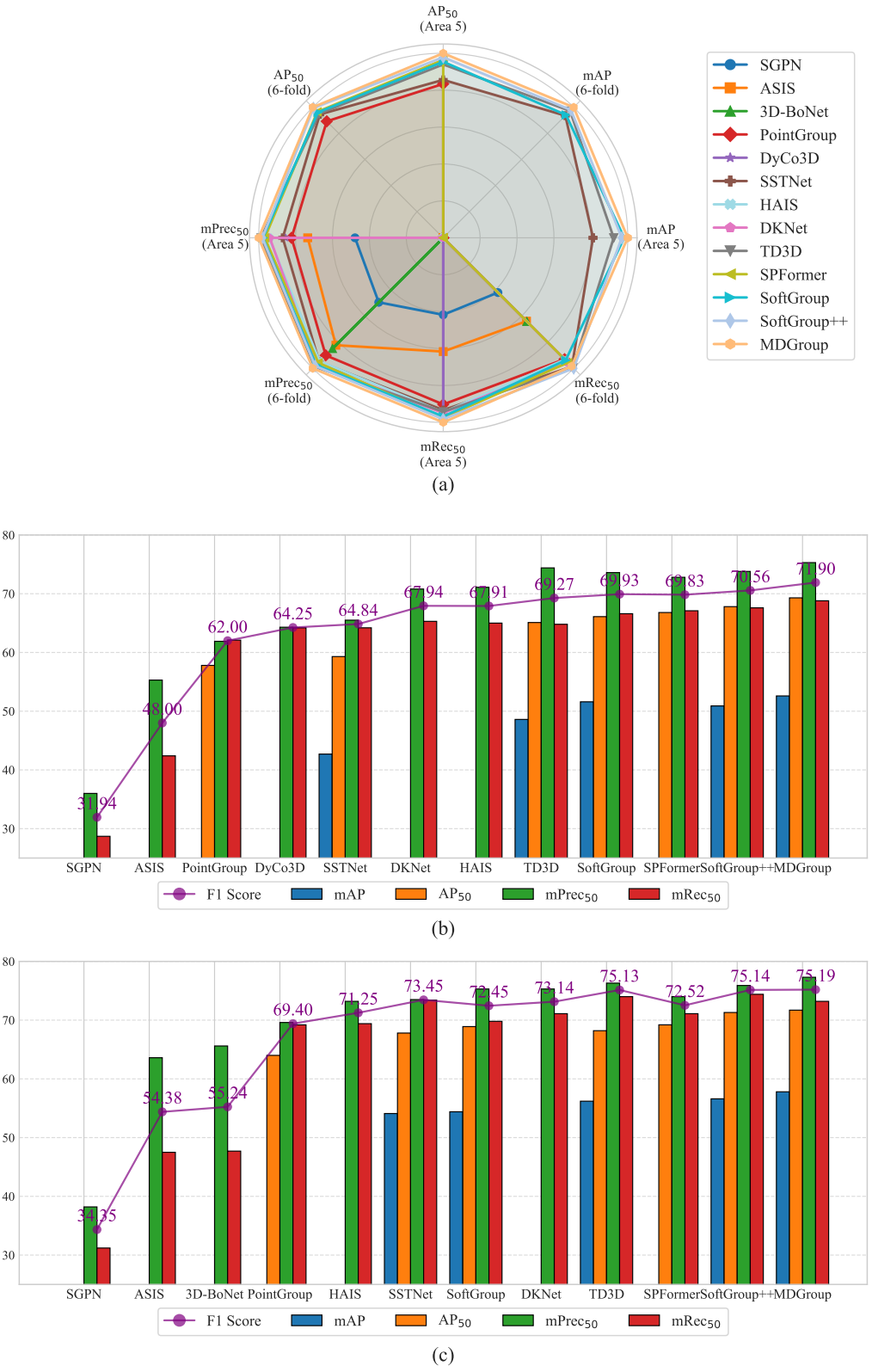

Figure 3. (**a**) Overall comparison on the S3DIS dataset. (**b**) Results under the Area 5 settings. (**c**) Results under 6-fold cross-validation.

4.6. Comparison with Related Work on ScanNet v2 Dataset

This section reports the instance segmentation results of the proposed MDGroup method on both the validation and hidden test sets of the ScanNet v2 dataset. Detailed

results are presented in Table 7, where the best performance for each metric is highlighted in bold.

On the validation set, MDGroup achieves improvements of 11.6% in mAP, 1.3% in AP_{50}, and 1.1% in AP_{25} compared to the state-of-the-art grouping-based method SoftGroup. The gain in mAP is particularly notable. On the hidden test set, MDGroup further demonstrates its superiority under more rigorous evaluation conditions, outperforming SoftGroup by 9.9% in mAP, 2.6% in AP_{50}, and 0.8% in AP_{25}.

Table 7. The results of 3D instance segmentation on the ScanNet v2 dataset. The performances on the validation and test set are reported.

Method	mAP	AP_{50}	AP_{25}
			Validation set
3D-SIS [7]	-	18.7	35.7
GSPN [8]	19.3	37.8	53.4
3D-MPA [11]	-	51.9	72.4
PointGroup [1]	34.8	56.9	71.3
DyCo3D [29]	35.4	57.6	72.9
NeuralBF [60]	36.0	55.5	71.1
HAIS [3]	43.5	64.4	75.6
OccuSeg [41]	44.2	60.7	71.9
SoftGroup [4]	45.8	67.6	78.9
SSTNet [2]	49.4	64.3	74.0
DKNet [36]	50.8	66.7	76.9
MDGroup (ours)	**51.1**	**68.5**	**79.8**
			Hidden test set
NeuralBF [60]	35.3	55.5	71.8
DyCo3D [29]	39.5	64.1	76.1
PointGroup [1]	40.7	63.6	77.8
HAIS [3]	45.7	69.9	80.3
SoftGroup [4]	50.4	76.1	86.5
SSTNet [2]	50.6	69.8	78.9
DKNet [36]	53.2	71.8	81.5
SPFormer [31]	54.9	77.0	85.1
ISBNet [37]	55.9	75.7	83.5
Mask3D [30]	56.6	78.0	87.0
MDGroup (ours)	**55.4**	**78.1**	**87.2**

In addition to comparisons with grouping-based methods, this section also includes a broader evaluation against other advanced approaches. The results show that while some methods achieve comparable scores to SoftGroup in AP_{50} and AP_{25}, they significantly outperform it in mAP. This indicates that these methods can meet accuracy requirements under lower IoU thresholds but struggle under higher thresholds. The main reason lies in their imprecise or blurry instance boundaries. For example, SoftGroup often produces coarse boundary predictions with frequent local errors. Although the overall instance shape may be roughly correct, small boundary deviations can lead to matching failures under high IoU thresholds. This issue is common in models that rely on voxel-based representations, have limited feature extraction capabilities, or use suboptimal clustering parameters during post-processing. Boundary precision is inherently constrained by voxel resolution. For small objects or thin structures, predictions tend to include excessive background points or miss foreground points. Since these structures contain fewer points, even a single-point error can result in a high error ratio, further destabilizing boundary predictions.

To address these challenges, MDGroup introduces a dynamic multi-factor weighting function that integrates semantic confidence, geometric distance, directional consistency, and boundary probability. This enables fine-grained boundary refinement, enhancing intra-instance point aggregation while effectively suppressing incorrect inter-instance connections. Additionally, the model incorporates DRNet to ensure effective fusion of high-level semantic features with low-level high-resolution geometric information, which is essential for accurate boundary prediction.

MDGroup also leverages HAMF to extract and fuse multi-scale features. By incorporating object size distributions obtained during preprocessing, the model adaptively extracts features for different categories and explicitly fuses them across multiple scales. This design allows the model to capture contextual information for large objects and fine details for small ones, mitigating boundary coarseness caused by inappropriate clustering parameters. These innovations contribute to improvements in AP_{50} and AP_{25} while significantly boosting mAP, reflecting enhanced boundary precision under higher IoU thresholds.

Beyond segmentation accuracy, practical deployment requires careful consideration of model complexity and inference efficiency. Table 8 presents a comparative analysis of MDGroup against representative methods on the ScanNet v2 dataset. Compared with the strong baseline SoftGroup, MDGroup introduces a moderate increase in parameters and latency while delivering a significant improvement in mAP. Notably, our method remains substantially faster than clustering-heavy approaches such as SSTNet. Due to the efficient Octree-based neighbor search, MDGroup achieves an inference time reduction of approximately 12% compared to the 729 ms required by SSTNet. Furthermore, in contrast to Transformer-based architectures like Mask3D, MDGroup maintains a more lightweight footprint with roughly 13.6% fewer parameters. While Mask3D achieves competitive inference speeds, it relies on a heavy decoder structure that may limit scalability in resource-constrained environments. By avoiding the quadratic complexity associated with self-attention mechanisms and optimizing the grouping strategy, MDGroup delivers improved segmentation performance with a favorable complexity-efficiency trade-off, making it well-suited for large-scale indoor scene understanding. The inference speed of 1.56 FPS with single RTX 3090 GPU matches well with the real-time data stream output frequency of mainstream optical sensors such as vehicle-mounted LiDAR and indoor 3D scanners in practical application scenarios and the lightweight parameter scale of 34.2 M facilitates convenient deployment on edge computing devices for optical sensor data processing.

Table 8. Comparison of model complexity and real-time inference metrics.

Method	Parameters (M)	Inference Time (ms)	FPS
HAIS [3]	30.9	578	1.73
SoftGroup [4]	30.9	588	1.70
SSTNet [2]	-	729	1.37
Mask3D [30]	39.6	578	1.73
MDGroup (ours)	34.2	642	1.56

Table 9 presents the benchmark results on the ScanNet v2 hidden test set based on AP_{50}. MDGroup achieves the best performance among grouping-based instance segmentation methods, reaching an average AP_{50} of 78.1%, which is 2.6% higher than SoftGroup and 1.6% higher than SoftGroup++. The advantage is clearly demonstrated. From a class-level perspective, MDGroup achieves the best performance in 11 out of 18 categories. The heatmap in Figure 4 further confirms its robustness across diverse object types, where the red boxes mark the optimal performance results for each class.

Table 9. 3D instance segmentation results on the ScanNet v2 hidden test set based on AP_{50}.

Method	AP_{50}	Bathtub	Bed	Bookshelf	Cabinet	Chair	Counter	Curtain	Desk	Door	Other	Picture	Fridge	S. Curtain	Sink	Sofa	Table	Toilet	Window
SGPN [38]	14.3	20.8	39.0	16.9	6.5	27.5	2.9	6.9	0.0	8.7	4.3	1.4	2.7	0.0	11.2	35.1	16.8	43.8	13.8
GSPN [8]	30.6	50.0	40.5	31.1	34.8	58.9	5.4	6.8	12.6	28.3	29.0	2.8	21.9	21.4	33.1	39.6	27.5	82.1	24.5
3D-SIS [7]	38.2	**100.0**	43.2	24.5	19.0	57.7	1.3	26.3	3.3	32.0	24.0	7.5	42.2	85.7	11.7	69.9	27.1	88.3	23.5
MASC [61]	44.7	52.8	55.5	38.1	38.2	63.3	0.2	50.9	26.0	36.1	43.2	32.7	45.1	57.1	36.7	63.9	38.6	98.0	27.6
PanopticFusion [62]	47.8	66.7	71.2	59.5	25.9	55.0	0.0	61.3	17.5	25.0	43.4	43.7	41.1	85.7	48.5	59.1	26.7	94.4	35.9
3D-Bonet [12]	48.8	**100.0**	67.2	59.0	30.1	48.4	9.8	62.0	30.6	34.1	25.9	12.5	43.4	79.6	40.2	49.9	51.3	90.9	43.9
MTML [40]	54.9	**100.0**	80.7	58.8	32.7	64.7	0.4	81.5	18.0	41.8	36.4	18.2	44.5	**100.0**	44.2	68.8	57.1	**100.0**	39.6
3D-MPA [11]	61.1	**100.0**	83.3	76.5	52.6	75.6	13.6	58.8	47.0	43.8	43.2	35.8	65.0	85.7	42.9	76.5	55.7	**100.0**	43.0
Dyco3D [29]	64.1	**100.0**	84.1	89.3	53.1	80.2	11.5	58.8	44.8	43.8	53.7	43.0	55.0	85.7	53.4	76.4	65.7	98.7	56.8
PE [42]	64.5	**100.0**	77.3	79.8	53.8	78.6	8.8	79.9	35.0	43.5	54.7	54.5	64.6	93.3	56.2	76.1	55.6	99.7	50.1
PointGroup [1]	63.6	**100.0**	76.5	62.4	50.5	79.7	11.6	69.6	38.4	44.1	55.9	47.6	59.6	**100.0**	66.6	75.6	55.6	99.7	51.3
GICN [10]	63.8	**100.0**	**89.5**	80.0	48.0	67.6	14.4	73.7	35.4	44.7	40.0	36.5	70.0	**100.0**	56.9	83.6	59.9	**100.0**	47.3
OccuSeg [41]	67.2	**100.0**	75.8	68.2	57.6	84.2	**47.7**	50.4	52.4	56.7	58.5	45.1	55.7	**100.0**	75.1	79.7	56.3	**100.0**	46.7
SSTNet [2]	69.8	**100.0**	69.7	88.8	55.6	80.3	38.7	62.6	41.7	55.6	58.5	**70.2**	60.0	**100.0**	**82.4**	72.0	69.2	**100.0**	50.9
HAIS [3]	69.9	**100.0**	84.9	82.0	67.5	80.8	27.9	75.7	46.5	51.7	59.6	55.9	60.0	**100.0**	65.4	76.7	67.6	99.4	56.0
SoftGroup [4]	76.1	**100.0**	80.8	84.5	**71.6**	86.2	24.3	82.4	65.5	62.0	73.4	69.9	79.1	98.1	71.6	84.4	76.9	**100.0**	59.4
SoftGroup++ [5]	76.9	**100.0**	80.3	**93.7**	68.4	86.5	21.3	87.0	66.4	57.1	75.8	70.2	**80.7**	**100.0**	65.3	**90.2**	79.2	**100.0**	62.6
MDGroup (ours)	**78.1**	**100.0**	82.1	90.2	71.4	**88.0**	27.1	**88.4**	**67.6**	**63.9**	**77.3**	**71.2**	80.5	**100.0**	66.7	88.3	**79.6**	**100.0**	**63.5**

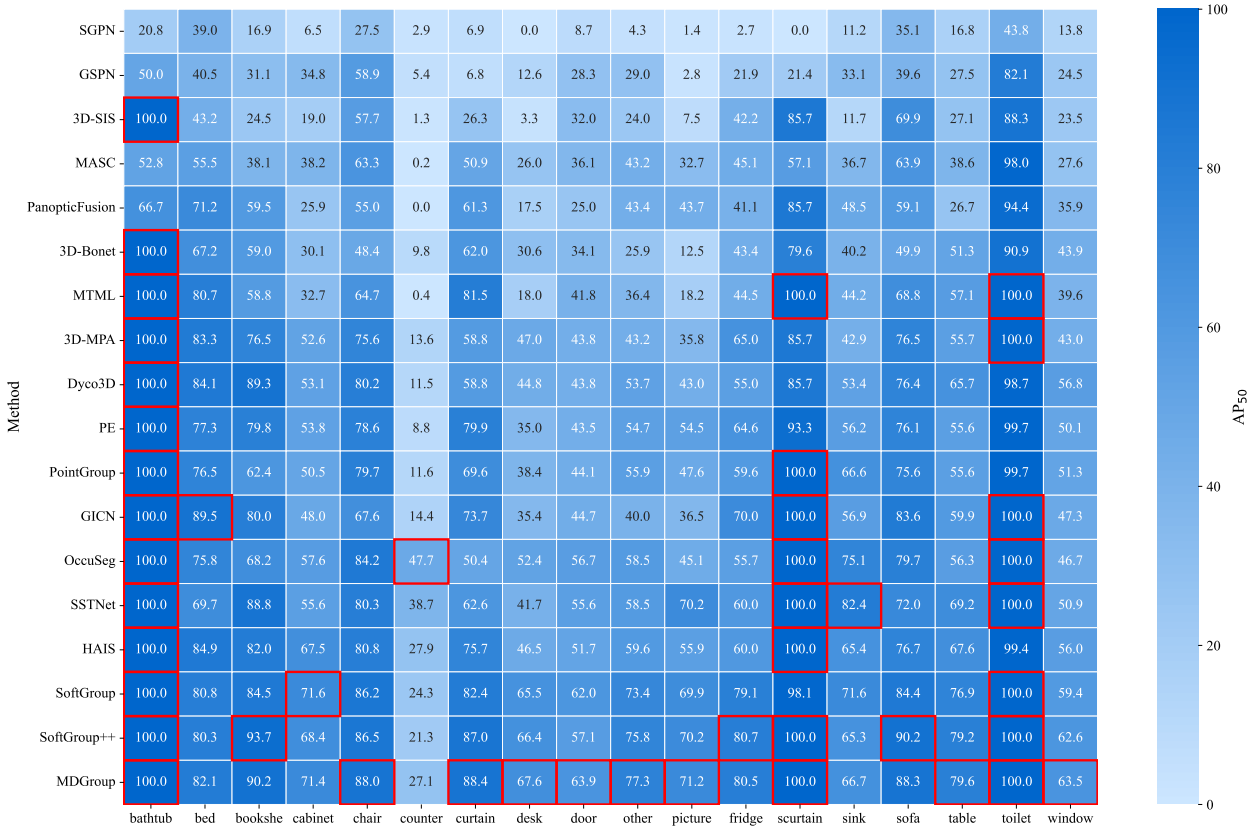

Figure 4. Per-class AP_{50} heatmap for 3D instance segmentation on the ScanNet v2 hidden test set.

Beyond per-class analysis, this section introduces a novel structured evaluation approach by grouping the 18 categories into four functional clusters based on semantic and usage characteristics. These clusters are defined as Large Furniture, Kitchen and Bathroom, Storage and Decoration, and Building Components. Figure 5 presents the comparative results under this structured analysis.

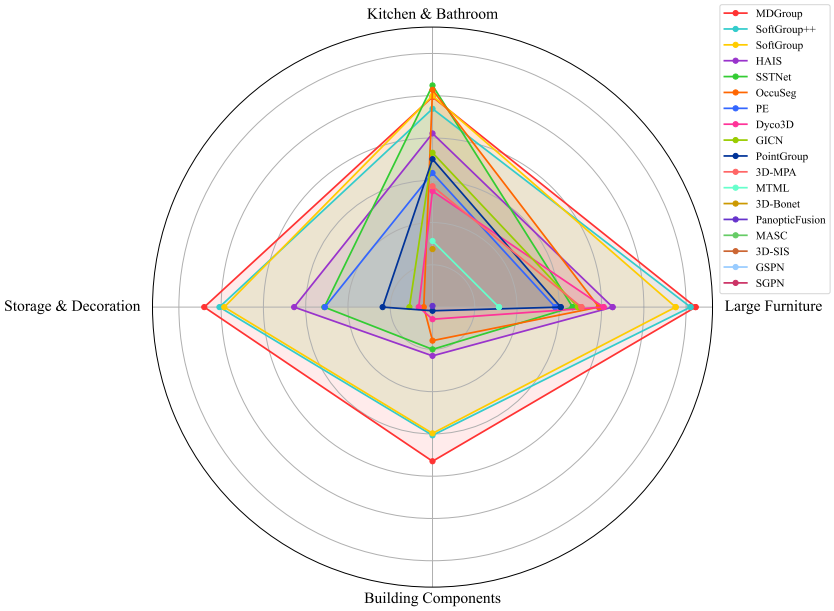

Figure 5. Structured analysis of AP_{50}-based 3D instance segmentation results on the ScanNet v2 hidden test set.

This approach elevates the evaluation of point cloud instance segmentation from fragmented per-class accuracy to structured semantic understanding. It aligns more closely with human perception of indoor scenes and helps reveal systemic algorithmic limitations. This perspective supports targeted optimization, facilitates downstream application adaptation, simplifies data management, and enhances cross-scene generalization analysis. One of the core goals of point cloud instance segmentation is to enable machines to understand the semantic structure of indoor environments. Human cognition naturally follows a hierarchical logic from broad categories to fine-grained classes. Structured evaluation allows for rapid identification of weaknesses in functional feature extraction and reduces bias caused by occasional errors in small categories. It also helps balance data distribution and alleviates the issue of sparse samples in rare classes. While the distribution of fine-grained categories varies significantly across scenes, the functional logic of broader categories remains relatively stable. Therefore, structured analysis better reflects the generalization capability of segmentation algorithms.

4.7. Comparison with Related Work on STPLS3D Dataset

This section reports the instance segmentation results of the proposed MDGroup method on the STPLS3D dataset. The detailed results are presented in Table 10, where the best performance for each metric is highlighted in bold.

Table 10. 3D instance segmentation results on the STPLS3D dataset.

Method	AP	AP_{50}	AP_{25}
PointGroup [1]	23.3	38.5	48.6
HAIS [3]	35.1	46.7	52.8
SoftGroup [4]	47.3	63.1	71.4
SoftGroup++ [5]	46.5	62.9	71.8
MDGroup (ours)	**52.8**	**68.5**	**75.6**

STPLS3D is a synthetic outdoor scene dataset whose data generation process closely resembles that of photogrammetric point clouds acquired from aerial surveys. Compared to ScanNet v2 and S3DIS, STPLS3D features higher sparsity and more pronounced variations in object sizes, posing greater challenges to the robustness of segmentation algorithms.

MDGroup achieves 52.8% AP, 68.5% AP_{50}, and 75.6% AP_{25}. These results represent improvements of 11.6%, 8.6%, and 5.9% over SoftGroup, and 13.5%, 8.9%, and 5.3% over SoftGroup++, respectively. The experimental findings demonstrate that the proposed method not only excels on indoor datasets but also shows significant advantages in outdoor environments, further highlighting its effectiveness and generalizability.

4.8. Comparison with Related Work on SemanticKITTI Dataset

To further validate the generalizability of our architecture and its potential for dynamic scene understanding, we extend our evaluation to the challenging large-scale autonomous driving dataset, SemanticKITTI. This dataset is particularly suitable for this purpose as it provides sequential LiDAR scans with consistent instance annotations across time, enabling a rigorous assessment of a model's ability to leverage temporal signal correlation. This experiment aims to demonstrate that our core architecture can be effectively extended to handle spatio-temporal data through the integration of the proposed TAG module.

For this evaluation, the LSTQ metric is of particular importance as it directly quantifies the critical aspects of both segmentation and tracking quality, making it the most significant indicator for the efficacy of our module. The mAQ and mIoU metrics serve as valuable

secondary indicators, allowing for a more granular analysis of instance fidelity and semantic context respectively.

A direct evaluation of our baseline MDGroup model on this benchmark is not feasible, as its single-frame processing nature produces instance identifiers that are independent and inconsistent between consecutive frames. To establish a fair and meaningful comparison, we construct a robust baseline, denoted as MDGroup (w/o TAG), by augmenting the frame-by-frame predictions of the original model with a standard IoU-based greedy matching algorithm for post-hoc instance association. This ensures that the only variable between the baseline and our temporally-aware model, MDGroup (w/ TAG), is the presence of the feature-level temporal fusion mechanism.

The quantitative results are presented in Table 11. Our full model, MDGroup (w/ TAG), demonstrates outstanding performance, achieving the highest LSTQ of 64.7 and a leading mAQ of 68.1 among the compared methods. This result underscores the strength of our core instance segmentation paradigm in complex outdoor environments. More importantly, the ablation study clearly reveals the significant contribution of the TAG module. The integration of temporal information yields a substantial improvement of 2.2 points in LSTQ over the baseline. This demonstrates that the TAG module effectively enhances temporal stability and instance tracking continuity. Furthermore, performance gains are also observed in mAQ and mIoU, which indicates that leveraging historical context at the feature level not only improves tracking but also refines the precision of instance masks and strengthens the overall semantic understanding of the scene. These findings confirm that our proposed architecture provides a solid foundation that is readily extensible to dynamic scenarios, successfully bridging the gap between static and sequential point cloud processing.

Table 11. Quantitative results for 4D Panoptic Segmentation on the SemanticKITTI validation set.

Method	LSTQ	mAQ	mIoU
Panoptic4D [53]	56.9	56.4	57.4
Contrastive Instance Association [63]	63.1	65.7	60.6
4D-StOP [64]	63.9	**69.5**	58.8
Mask4D [65]	64.3	66.4	62.2
MDGroup (w/o TAG)	62.5	67.0	62.3
MDGroup (w/ TAG)	**64.7**	68.1	**62.9**

4.9. Robustness to Real-World Sensor Artifacts

To validate the practical applicability of our method we extend our evaluation to datasets that contain inherent imperfections from real optical imaging links. Standard benchmarks often consist of clean or reconstructed point clouds which may not fully represent the challenges encountered in real-world applications. We therefore test MDGroup on LiDAR-Net and OCID two datasets characterized by sensor-specific noise and complex scene structures. These experiments are crucial for demonstrating the model's ability to maintain the integrity of geometric and boundary signals under realistic degradation.

We selected LiDAR-Net and OCID because they introduce two distinct and complementary types of real-world challenges. LiDAR-Net is captured using a mobile LiDAR scanner and features raw point clouds with authentic sensor artifacts. These include non-uniform scanning densities leading to holes and lines as well as explicitly annotated reflection noise and motion-induced ghosts. This dataset directly tests the model's resilience to sensor-level noise that corrupts local geometric information. In contrast the OCID dataset is captured with consumer-grade RGB-D sensors which have a different noise profile. Its primary challenge lies in extreme object clutter including physically touching

and stacked instances. This setup creates severe boundary ambiguity and tests the model's ability to segment instances when low-level geometric separation cues are absent. Success on both datasets would demonstrate broad robustness across different sensor modalities and challenging physical configurations.

Table 12 presents the instance segmentation results on the LiDAR-Net dataset. Our MDGroup significantly outperforms previous methods. It achieves an mPrec of 46.8 percent and an mRec of 29.2 percent representing substantial improvements of 7.0 and 4.1 percentage points over the strongest baseline. The notable increase in mPrec is largely attributable to our DRNet. Its high-resolution branch preserves fine-grained geometric details which is critical for maintaining precision in the presence of non-uniform point distributions and sensor noise. Furthermore the significant gain in mRec highlights the effectiveness of our boundary-aware mechanisms. The four-branch prediction and the BAWA module effectively handle ambiguous regions caused by reflection noise and ghosts enabling more accurate instance grouping and reducing missed detections.

Table 12. Instance segmentation performance on the LiDAR-Net dataset.

Method	mPrec	mRec
ASIS [59]	37.3	25.1
3D-BoNet [12]	39.8	24.6
MDGroup (ours)	**46.8**	**29.2**

The results on the OCID dataset shown in Table 13 further confirm the superiority of our approach in highly cluttered scenes. MDGroup consistently surpasses all baseline methods across both the ARID20 and YCB10 subsets. Notably it outperforms LCCP a strong geometry-based method that relies on local convexity. The performance gains are most pronounced in mRec and consequently mIoU. For instance on the ARID20 subset without background labels MDGroup improves mRec by 7 percentage points over LCCP. This indicates that MDGroup effectively mitigates the under-segmentation errors common to geometry-based methods when faced with touching or stacked objects. This success is driven by the model's learned semantic and directional priors which enable it to infer instance boundaries even without clear geometric separation.

Table 13. Quantitative instance segmentation results on the ARID20 and YCB10 subsets of the OCID dataset. Performance is reported w/ and w/o considering the background label.

Subset	Method	w/ Background Label				w/o Background Label			
		mPrec	mRec	F1-Score	mIoU	mPrec	mRec	F1-Score	mIoU
ARID20	GCUT [66]	0.67	0.56	0.56	0.44	0.66	0.54	0.54	0.42
	SCUT [67]	0.74	0.71	0.69	0.60	0.72	0.70	0.68	0.59
	V4R [68]	0.76	0.77	0.73	0.68	0.75	0.76	0.72	0.66
	LCCP [69]	0.93	0.83	0.85	0.79	0.92	0.82	0.84	0.78
	MDGroup (ours)	**0.95**	**0.90**	**0.92**	**0.86**	**0.94**	**0.89**	**0.91**	**0.84**
YCB10	GCUT [66]	0.62	0.54	0.51	0.39	0.59	0.55	0.51	0.38
	SCUT [67]	0.73	0.71	0.69	0.61	0.70	0.69	0.67	0.58
	V4R [68]	0.77	0.79	0.76	0.72	0.75	0.77	0.74	0.69
	LCCP [69]	0.96	0.88	0.90	0.86	0.96	0.87	0.89	0.85
	MDGroup (ours)	**0.97**	**0.92**	**0.94**	**0.90**	**0.97**	**0.91**	**0.94**	**0.89**

In summary, the comprehensive evaluations on LiDAR-Net and OCID validate the robustness of MDGroup. The results show that our architectural designs effectively counteract signal degradation from both sensor-level artifacts and complex scene-level occlusions. This demonstrated resilience to real-world imperfections underscores the practical value and advanced capabilities of our proposed method for real-world robotic applications.

5. Conclusions

This work revisits the key bottlenecks of grouping-based point cloud instance segmentation from three perspectives: backbone representation, cross-scale feature fusion, and aggregation priors. A set of interconnected solutions is proposed to address these challenges.

The designed DRNet enables decoupled learning of coarse-grained global semantic context and fine-grained local geometric structure through four parallel branches for semantic, offset, boundary, and direction prediction. By combining sparse convolution with a lightweight alignment strategy, DRNet achieves a balance between detail preservation and computational efficiency. HAMF injects multi-scale features into the detail stream without introducing additional high-resolution branches. It extracts scene-level global semantic features via CAPS and class-level local structural features via CADV. These features are fused with high-resolution geometric details to generate multi-granularity representations, allowing the model to maintain stable discriminative power even under long-range dependencies. BAWA explicitly introduces boundary and directional priors to compensate for the limitations of guided-based aggregation in tightly connected regions. It leverages a point shift refinement mechanism to suppress cross-instance propagation and significantly improves boundary prediction quality in complex scenarios. These three components complement each other in both structural design and optimization strategy. Together, they deliver substantial performance gains over a strong baseline, particularly in the segmentation of small objects and the separation of adjacent instances.

Empirical results show that the proposed method not only improves standard metrics but also demonstrates consistent advantages in boundary clarity and instance consistency. By integrating the Temporal Adaptive Gating module and resolution-aware feature extraction, the model successfully extends its capabilities to dynamic scenes and adapts to the physical constraints of different optical sensors. More importantly, the quantified real-time metrics, including inference time and FPS and lightweight parameter scale, guarantee its practical applicability to real-time data stream processing of optical sensors in core application scenarios such as autonomous driving and robotic navigation while maintaining memory efficiency and inference latency control. Nevertheless, it is necessary to discuss the trade-offs inherent in the proposed design compared to other approaches. The integration of the dual-resolution backbone and the multi-branch prediction heads inevitably increases the computational complexity and memory footprint compared to single-path baselines. While the inference speed remains practical on high-performance GPUs, deployment on resource-constrained edge devices may require further optimization, such as model pruning or knowledge distillation. Furthermore, although the adaptive mechanisms reduce the dependency on manual tuning, the model's sensitivity to the initial voxel resolution requires careful alignment with sensor specifications. Beyond these computational and architectural constraints, in extremely sparse scenes or those with high annotation noise, the reliability of boundary and direction supervision may still affect performance. Additionally, dynamic balancing of multi-task weights and memory scheduling in large-scale scenes remain areas for further optimization.

Future research may explore several directions. These include developing self-supervised or weakly supervised boundary and direction cues to reduce reliance on precise annotations, introducing uncertainty-aware multi-task weighting mechanisms for more robust training dynamics, and designing end-to-end differentiable aggregation strategies to replace heuristic clustering for improved consistency. The method can also be extended to panoptic segmentation and cross-domain adaptation to enhance generalization and robustness in open-world 3D environments.

In summary, MDGroup introduces a unified paradigm of decoupling, alignment, and constraint, offering a systematic solution that balances accuracy and efficiency in

point cloud instance segmentation. It validates the necessity and practical value of explicitly modeling temporal correlations and dual-awareness of class and boundary within a grouping-based framework. The proposed approach effectively bridges the gap between static spatial processing and dynamic real-world sensor data, providing reusable structural components and methodologies for future research.

Author Contributions: Conceptualization, R.H.; methodology, R.H.; software, R.H. and W.S.; validation, R.H.; formal analysis, R.H. and W.S.; data curation, R.H.; writing—original draft preparation, R.H.; writing—review and editing, R.H. and W.S.; visualization, R.H.; supervision, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The codes that support the findings of this study are available in https://github.com/MapleAttach/MDGroup, accessed on 15 January 2026. The datasets used in this study were obtained from the publicly accessible websites as follows. ScanNet v2: https://github.com/ScanNet/ScanNet, accessed on 15 February 2026. S3DIS: Official download via https://docs.google.com/forms/d/e/1FAIpQLScDimvNMCGhy_rmBA2gHfDu3naktRm6A8BPwAWWDv-Uhm6Shw/viewform, accessed on 15 January 2026, alternative accessible mirror: https://cvg-data.inf.ethz.ch/s3dis, accessed on 15 February 2026. STPLS3D: https://github.com/meidachen/STPLS3D, accessed on 15 February 2026. SemanticKITTI: https://semantic-kitti.org/index.html, accessed on 15 February 2026. LiDAR-Net: http://lidar-net.njumeta.com, accessed on 15 February 2026. OCID: https://www.acin.tuwien.ac.at/vision-for-robotics/software-tools/object-clutter-indoor-dataset, accessed on 15 February 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDGroup	Multi-grained Dual-aware Grouping algorithm
DRNet	Dual-Resolution 3D U-Net
HAMF	Hierarchical Adaptive Multi-grained Feature fusion framework
BAWA	Boundary-Aware Weighted Aggregation mechanism
TAG	Temporal Adaptive Gating
CADV	Class-Aware Dynamic Voxelization
CAPS	Class-Aware Pyramid Scaling
3D-RPN	3D Region Proposal Network
3D-RoI	3D Region-of-Interest
BEV	bird's-eye view
VDRAE	Variational Denoising Recursive Autoencoder
SubMConv3d	Submanifold Sparse Convolution
FPN	feature pyramid network
BCE	Binary Cross-Entropy
MLP	Multi-Layer Perceptron
k-NN	k-Nearest Neighbors
BFS	Breadth-First Search
AP	Average Precision
mAP	mean Average Precision
IoU	Intersection over Union
mPrec	mean Precision
mRec	mean Recall
FPS	Frames Per Second
LSTQ	Lidar Segmentation and Tracking Quality
mAQ	mean Association Quality
mIoU	mean Intersection over Union

References

1. Jiang, L.; Zhao, H.; Shi, S.; Liu, S.; Fu, C.W.; Jia, J. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020*; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 4867–4876.
2. Liang, Z.; Li, Z.; Xu, S.; Tan, M.; Jia, K. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 2783–2792.
3. Chen, S.; Fang, J.; Zhang, Q.; Liu, W.; Wang, X. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 15467–15476.
4. Vu, T.; Kim, K.; Luu, T.M.; Nguyen, T.; Yoo, C.D. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022*; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 2708–2717.
5. Vu, T.; Kim, K.; Nguyen, T.; Luu, T.M.; Kim, J.; Yoo, C.D. Scalable softgroup for 3d instance segmentation on point clouds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 1981–1995. [CrossRef] [PubMed]
6. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 2961–2969.
7. Hou, J.; Dai, A.; Nießner, M. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 4421–4430.
8. Yi, L.; Zhao, W.; Wang, H.; Sung, M.; Guibas, L.J. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 3947–3956.
9. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 652–660.
10. Liu, S.H.; Yu, S.Y.; Wu, S.C.; Chen, H.T.; Liu, T.L. Learning gaussian instance segmentation in point clouds. *arXiv* **2020**, arXiv:2007.09860. [CrossRef]
11. Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; Nießner, M. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020*; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 9031–9040.
12. Yang, B.; Wang, J.; Clark, R.; Hu, Q.; Wang, S.; Markham, A.; Trigoni, N. Learning object bounding boxes for 3d instance segmentation on point clouds. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 605. [CrossRef]
13. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114. [CrossRef]
14. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
15. Kolodiazhnyi, M.; Vorontsova, A.; Konushin, A.; Rukhovich, D. Top-down beats bottom-up in 3d instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2024*; IEEE Computer Society: Los Alamitos, CA, USA, 2024; pp. 3566–3574.
16. Zhang, F.; Guan, C.; Fang, J.; Bai, S.; Yang, R.; Torr, P.H.; Prisacariu, V. Instance segmentation of lidar point clouds. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–15 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 9448–9455.
17. Shi, Y.; Chang, A.X.; Wu, Z.; Savva, M.; Xu, K. Hierarchy denoising recursive autoencoders for 3D scene layout prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 1771–1780.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010. [CrossRef]
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
21. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 1290–1299.

22. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
23. Liu, Z.; Zhang, Z.; Cao, Y.; Hu, H.; Tong, X. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 2949–2958.
24. Misra, I.; Girdhar, R.; Joulin, A. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 2906–2917.
25. Pan, X.; Xia, Z.; Song, S.; Li, L.E.; Huang, G. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 7463–7472.
26. Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; Jia, J. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 8500–8509.
27. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 16259–16268.
28. Park, C.; Jeong, Y.; Cho, M.; Park, J. Fast point transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 16949–16958.
29. He, T.; Shen, C.; Van Den Hengel, A. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 354–363.
30. Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; Leibe, B. Mask3d: Mask transformer for 3d semantic instance segmentation. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 29 May–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 8216–8223.
31. Sun, J.; Qing, C.; Tan, J.; Xu, X. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023*; AAAI Press: Palo Alto, CA, USA, 2023; Volume 37, pp. 2393–2401.
32. Lai, X.; Yuan, Y.; Chu, R.; Chen, Y.; Hu, H.; Jia, J. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 3693–3703.
33. Lu, J.; Deng, J.; Wang, C.; He, J.; Zhang, T. Query refinement transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 18516–18526.
34. Jia, X.; De Brabandere, B.; Tuytelaars, T.; Gool, L.V. Dynamic filter networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 667–675. [CrossRef]
35. He, T.; Yin, W.; Shen, C.; Van den Hengel, A. Pointinst3d: Segmenting 3d instances by points. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 286–302.
36. Wu, Y.; Shi, M.; Du, S.; Lu, H.; Cao, Z.; Zhong, W. 3d instances as 1d kernels. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 235–252.
37. Ngo, T.D.; Hua, B.S.; Nguyen, K. Isbnet: A 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 13550–13559.
38. Wang, W.; Yu, R.; Huang, Q.; Neumann, U. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018*; IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 2569–2578.
39. Pham, Q.H.; Nguyen, T.; Hua, B.S.; Roig, G.; Yeung, S.K. JSIS3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 8827–8836.
40. Lahoud, J.; Ghanem, B.; Pollefeys, M.; Oswald, M.R. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 9256–9266.
41. Han, L.; Zheng, T.; Xu, L.; Fang, L. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020*; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 2940–2949.

42. Zhang, B.; Wonka, P. Point cloud instance segmentation using probabilistic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 8883–8892.
43. Zhao, W.; Yan, Y.; Yang, C.; Ye, J.; Yang, X.; Huang, K. Divide and conquer: 3d point cloud instance segmentation with point-wise binarization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 562–571.
44. Yuan, Y.; Tian, Y. Semantic segmentation algorithm of underwater image based on improved DeepLab v3+. In *Proceedings of the International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2022), Guangzhou, China, 24 December 2022*; SPIE: Bellingham, WA, USA, 2023; Volume 12604, pp. 862–867.
45. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
46. Graham, B.; Engelcke, M.; Van Der Maaten, L. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018*; IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 9224–9232.
47. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 5828–5839.
48. Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016*; IEEE Computer Society: Los Alamitos, CA, USA, 2016; pp. 1534–1543.
49. Chen, M.; Hu, Q.; Yu, Z.; Thomas, H.; Feng, A.; Hou, Y.; McCullough, K.; Ren, F.; Soibelman, L. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv* **2022**, arXiv:2203.09065.
50. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 9297–9307.
51. Guo, Y.; Li, Y.; Ren, D.; Zhang, X.; Li, J.; Pu, L.; Ma, C.; Zhan, X.; Guo, J.; Wei, M.; et al. Lidar-net: A real-scanned 3d point cloud dataset for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024*; IEEE Computer Society: Los Alamitos, CA, USA, 2024; pp. 21989–21999.
52. Suchi, M.; Patten, T.; Fischinger, D.; Vincze, M. EasyLabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 6678–6684.
53. Aygun, M.; Osep, A.; Weber, M.; Maximov, M.; Stachniss, C.; Behley, J.; Leal-Taixé, L. 4d panoptic lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 5527–5537.
54. Loghmani, M.R.; Caputo, B.; Vincze, M. Recognizing objects in-the-wild: Where do we stand? In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 2170–2177.
55. Calli, B.; Singh, A.; Bruce, J.; Walsman, A.; Konolige, K.; Srinivasa, S.; Abbeel, P.; Dollar, A.M. Yale-CMU-Berkeley dataset for robotic manipulation research. *Int. J. Robot. Res.* **2017**, *36*, 261–268. [CrossRef]
56. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In *Proceedings of the Neural Information Processing Systems (NIPS) Workshop on Autodiff, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
57. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
58. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
59. Wang, X.; Liu, S.; Shen, X.; Shen, C.; Jia, J. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 4096–4105.
60. Sun, W.; Rebain, D.; Liao, R.; Tankovich, V.; Yazdani, S.; Yi, K.M.; Tagliasacchi, A. Neuralbf: Neural bilateral filtering for top-down instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 551–560.
61. Liu, C.; Furukawa, Y. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv* **2019**, arXiv:1902.04478. [CrossRef]

62. Narita, G.; Seno, T.; Ishikawa, T.; Kaji, Y. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 4205–4212.
63. Marcuzzi, R.; Nunes, L.; Wiesmann, L.; Vizzo, I.; Behley, J.; Stachniss, C. Contrastive instance association for 4d panoptic segmentation using sequences of 3d lidar scans. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1550–1557. [CrossRef]
64. Kreuzberg, L.; Zulfikar, I.E.; Mahadevan, S.; Engelmann, F.; Leibe, B. 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 537–553.
65. Marcuzzi, R.; Nunes, L.; Wiesmann, L.; Marks, E.; Behley, J.; Stachniss, C. Mask4d: End-to-end mask-based 4d panoptic segmentation for lidar sequences. *IEEE Robot. Autom. Lett.* **2023**, *8*, 7487–7494. [CrossRef]
66. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
67. Pham, T.T.; Do, T.T.; Sünderhauf, N.; Reid, I. Scenecut: Joint geometric and object segmentation for indoor scenes. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 3213–3220.
68. Potapova, E.; Richtsfeld, A.; Zillich, M.; Vincze, M. Incremental attention-driven object segmentation. In *Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots, Madrid, Spain, 18–20 November 2014*; IEEE: Piscataway, NJ, USA, 2014; pp. 252–258.
69. Christoph Stein, S.; Schoeler, M.; Papon, J.; Worgotter, F. Object partitioning using local convexity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014*; IEEE Computer Society: Los Alamitos, CA, USA, 2014; pp. 304–311.