xMHashSeg: Cross-modal Hash Learning for Training-free Unsupervised LiDAR Semantic Segmentation

Jialong Zhang[1], Yachao Zhang[2*], Yao Wu[2,3], Jiangming Shi[1], Fangyong Wang[4], Yanyun Qu[1,2*]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Institute of Artificial Intelligence, Xiamen University, Xiamen, China
[2]School of Informatics, Xiamen University, Xiamen, China
[3]Fuzhou University, Fuzhou, China
[4]Hanjiang National Laboratory
jialongzhang@stu.xmu.edu.cn, yachaozhang@stu.xmu.edu.cn, yyqu@xmu.edu.cn

Abstract

3D semantic segmentation serves as a fundamental component in many applications, such as autonomous driving and medical image analysis. Although recent methods have advanced the field, adapting these methods to new environments or object categories without extensive retraining remains a significant challenge. To address this, we introduce xMHashSeg, a novel training-free cross-modal LiDAR semantic segmentation framework. xMHashSeg leverages foundation models and non-parametric network to extract features from 2D images and 3D point clouds, subsequently integrating these features through hash learning. Specifically, We develop point-SANN, a novel self-adaption non-parametric network that can extract robust 3D features from raw point clouds, while 2D features are directly extracted through the foundation model DINOv2. To reconcile inconsistencies across different modals, we introduce a Hash Code Learning Module that projects all information into a common hash space, learning a consistent hash code that enhances feature integration. Additionally, depth maps are utilized as an intermediary form between 2D and 3D data to facilitate convergence during hash code learning. Our experimental results on various multi-modality datasets demonstrate that xMHashSeg outperforms zero-shot learning approaches and achieve performance close to that of unsupervised domain adaptation and test-time adaptation methods, without requiring any annotations or additional training.

Code — https://github.com/Kznnd/xMHashSeg

Introduction

3D understanding (Qi et al. 2017; Wu et al. 2019; Zhang et al. 2024) has emerged as a pivotal research domain, garnering significant attention from both academic and industrial sectors. The advancements in this field have substantially propelled innovations across multiple sectors including autonomous driving, robotic navigation, virtual reality, and medical image analysis. With the continuous evolution of deep learning algorithms and remarkable improvements in computational power, the accuracy and efficiency of 3D semantic segmentation have witnessed unprecedented enhancements, showcasing its tremendous potential in addressing complex real-world problems.

Despite these advancements, a significant challenge remains in the application of 3D semantic segmentation: the time-consuming and resource-intensive process of learning new models for different scenarios. Traditionally, adapting segmentation models to new environments or object categories necessitates extensive retraining. In recent years, the advent of training-free technologies has offered a promising alternative. These approaches leverage pre-trained foundational models to perform inference without the need for additional training, significantly reducing the associated costs and making it feasible to apply 3D semantic segmentation in a broader range of open-vocabulary settings.

In this paper, we introduce a novel training-free multi-modal LiDAR semantic segmentation framework named xMHashSeg. xMHashSeg focuses on leveraging foundation models and non-parametric network to extract multi-modal (or multi-view) features and optimizing a unified hash code for effective feature fusion, achieving segmentation without training data and annotations. Specifically, we address two pivotal questions that guide our approach to achieving accurate LiDAR semantic segmentation. Intuitively, integrating multiple modalities information can complement each other and provide richer information. However, as no joint training with labels, the feature space and representation among different modality features are different, and directly adding or concatenating them for feature fusion can result in feature imbalance. So **how to effectively integrate multi-modal information?** To solve this problem, this paper introduce HCLM to map multi-view features into a common hash space and learn a unified hash representation, promoting efficient fusion of multi-modal information. This naturally raises another question, **which modal information should be used?** Firstly, RGB images provide rich color and texture details, while point clouds offer precise 3D geometric structures. These two modalities are usually easily obtainable with current technology and provide us with rich information. However, there are significant modal gap between them, which are not conducive to the iterative convergence of the unified hash code. Therefore, we further in-

*Correspondence author

troduce depth map modality which is like an intermediary form between 2D images and 3D point clouds, functioning exceptionally well as a bridge to integrate information from both. It significantly enhances convergence during the iterative optimization of hash code, which is validated through the ablation experiment below.

Based on the above analysis, without loss of generality, we employ DINOv2 (Oquab et al. 2024) to extract 2D features from images. To complement this, we develop a novel non-parametric point cloud network called Point-SANN (Self-Adaptive Non-Parametric Network) that directly extracts 3D features from raw point clouds, which can be used in different scenarios with uneven point density. However, when integrating features of different views, a critical challenge arises from the inconsistency in feature representation. To overcome this challenge, We adopt the concept of Binary Multi-View Clustering (BMVC) (Zhang et al. 2018) to construct a Hash Code Learning Module (HCLM). HCLM projects all view information into a common hash space, where a unified hash representation is learned. This unified learning process harmonizes the variations among different view representations and ensures that complementary information from various modalities is fully exploited. Furthermore, considering the modal gap between 2D and 3D data, we introduce depth map view, which not only enriches the available information but also serves as an intermediate state that bridges 2D images and 3D point clouds, thereby facilitating the convergence of these three views towards a unified center during the learning of unique hash codes. After that, the learned hash code can be used for accurate segmentation.

In summary, the main contributions of this paper are depict as follows:

- We propose xMHashSeg, a novel training-free framework for multi-modal LiDAR semantic segmentation in open-world scenarios, which replace traditional learning methods that require retraining when encountering new scenes or categories.
- We develop a self-adaption non-parametric network Point-SANN to extract 3D features from raw LiDAR point clouds in open-scene with uneven point density, which fills the gap left by the current absence of foundation models in the 3D domain.
- To address the challenge of inconsistent feature scales derived from different data views, we introduce Hash Code Learning Module to project multiple views into the same hash space and learn a unified hash representation. This technique ensures that features from diverse views are harmoniously integrated, allowing for complementary information to be fully utilized.
- Through comprehensive experiments, the proposed framework achieves appreciable results in 3D semantic segmentation tasks, even without using any labels or engaging in additional training processes.

Related Work

LiDAR semantic segmentation

LiDAR semantic segmentation has emerged as a pivotal task within the realm of computer vision, essential for applications ranging from autonomous driving to robotics navigation. Initially, the field was dominated by fully supervised learning approaches, which rely heavily on large datasets with precise annotations for training deep neural networks (Alonso et al. 2020). However, the acquisition and annotation of voluminous point cloud data are both time-consuming and costly endeavors. To mitigate the dependency on exhaustive labeling, semi-supervised learning strategies (Zhang et al. 2021, 2025) have been introduced. By leveraging a small set of labeled data alongside a larger pool of unlabeled data, these techniques aim to enhance model generalization while reducing the need for extensive manual annotation (Mei et al. 2019; Kong et al. 2023).

Unsupervised Domain Adaptation (UDA) (Wu et al. 2024a) represents a significant advancement by addressing the issue of domain shift between training and testing environments. In scenarios where annotated data from one domain is insufficient or unavailable for another, UDA facilitates the transfer of knowledge across domains. Techniques such as adversarial training (Peng et al. 2021; Yuan et al. 2023) have shown promise in adapting models trained on source data to perform well on target scans without requiring additional labeled data. Building on this concept, Test-Time Adaptation (TTA) (Shin et al. 2022; Cao et al. 2024) enables models to adapt dynamically during inference. TTA exploits the incoming data stream at test time to refine model parameters or predictions, thereby enhancing accuracy under changing conditions.

In addition, Zero-Shot Learning (ZSL) methods presents an innovative approach that bypasses the necessity for any labeled examples of unseen classes during training. ZSL frameworks (Yang et al. 2023; Lu et al. 2023) typically incorporate auxiliary information, such as attributes or textual descriptions, to bridge the gap between seen and unseen categories. However, these methods are all based on learning and time-consuming.

Foundation models

Foundation models have become a cornerstone in the field of artificial intelligence, representing large-scale pre-trained models that are designed to perform a wide array of tasks with minimal fine-tuning. These models, such as Stable Diffusion (Rombach et al. 2022), DINOv2 (Oquab et al. 2024), and CLIP (Radford et al. 2021), have been developed across various modalities including text, images, and videos. In terms of application, foundation models have demonstrated remarkable versatility when applied to downstream tasks such as classification and segmentation. These models often require only minor adjustments or even none at all. For example, many of works (Ali and Khan 2023; Qian and Hu 2024) have successfully employed CLIP for zero-shot image classification, relying on the model's pre-existing knowledge base to infer class labels. Similarly, work done on DINOv2 (Veasey and Amini 2024) shows that it can achieve state-of-the-art results on image classification benchmarks with minimal fine-tuning, highlighting the robustness of its learned representations. Segmentation tasks have also benefited from the advent of foundation models. SAM (Segment Anything Model) (Kirillov et al. 2023), which builds upon

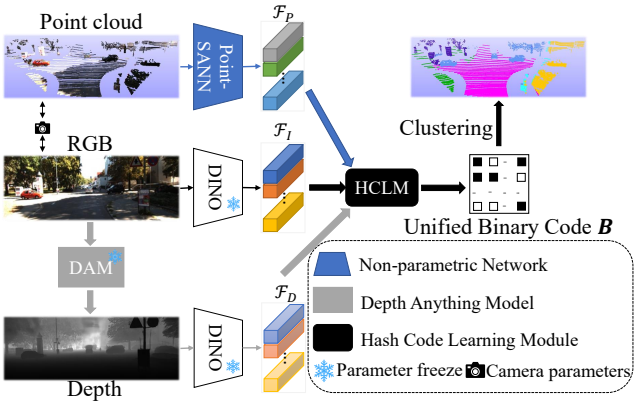

Figure 1: Overview framework. It employs DINOv2 to extract both 2D and depth features. Depth estimation is achieved through DAM, while point features are acquired using the developed Point-SANN. These views are subsequently aligned by learning a unified hash representation.

the principles of foundation modeling, offers a unified approach to image segmentation that requires no task-specific training. It can generate high-quality segmentation masks for any object in an image, guided merely by user-defined prompts. In summary, foundation models represent significant milestones in AI research, offering powerful tools for a multitude of applications.

Method

Problem Definition

Training-free (TF) refers to methods that leverage pre-trained foundation models or other non-learning techniques to perform downstream tasks without updating the model's parameters. The cross-modal TF LiDAR semantic segmentation task aims to derive 3D semantic segmentation results Y^P from a set of paired 2D images and 3D point clouds $\{X^I, X^P\}$ by leveraging foundation models and non-parametric network. Unlike UDA, TTA and ZSL methods, our approach does not require labeled data or a data-specific trained model, making it highly versatile and efficient for various applications.

Overview

Based on the two Q&A mentioned in the introduction, this paper constructs a framework as shown in Fig. 1. 2D images X^I are first processed through DAM to generate the corresponding depth map X^D, which is displayed in grayscale format. Since X^D is fundamentally a 2D image, it is fed into DINOv2 alongside X^I to extract their respective features F^I, F^D. For 3D feature extraction, we utilize an self-adaptive non-parametric network that builds upon the Point-NN (Zhang et al. 2023), called Point-SANN. It demonstrates robustness to variations in point cloud density, making it particularly well-suited for handling LiDAR point clouds with uneven density and effectively applicable to any scenario. Point-SANN processes the 3D point cloud to obtain detailed point-wise features F^P. Subsequently all the modality features (F^I, F^D, F^P) are input HCLM to learn a unified hash code and perform clustering to achieve semantic segmentation. Further details are provided below.

Multi-modal Information Extraction

For 2D images, our method directly employs DINOv2 to obtain pixel-wise features. Since DINOv2 divides images into multiple patches based on patch size to extract features, we upsample these features to the original image size using bicubic interpolation to obtain F^I.

For depth maps, to reduce the initial amount of information required, we utilize DAM to predict them. It generates relative depth map using just a single RGB image as input, without relying on additional sensors or prior scene information. Once the depth maps have been generated, we perform feature extraction using a methodology analogous to that employed for 2D images. This process allows us to derive features F^D.

For 3D point clouds, we develop Point-SANN, an improvement over Point-NN (Zhang et al. 2023), to extract features from LiDAR point clouds. Point-NN is a non-parametric network that leverages geometric priors and attention mechanisms to perform tasks such as classification and segmentation in a unsupervised manner. However, Point-NN does not address the issue of uneven point cloud density. When there are significant differences in density between different regions of the LiDAR point cloud, the receptive field used to capture local features can vary greatly, leading to substantial discrepancies in feature representation. Furthermore, Point-NN relies solely on positional encoding during initialization, which lacks semantic information and is insufficient for modeling long-range dependencies or handling non-sequential data structures, potentially introducing bias when applied to unordered inputs. To overcome these limitations, we introduce a self-adaption mechanism to capture features without sensitive to density variations and a richer embedding strategy that incorporate both positional and semantic information.

The framework of Point-SANN and the differences from Point-NN is shown in Fig. 2. We uses Fast Point Feature Histograms (FPFH) (Szalai-Gindl and Varga 2024) concatenate with point coordinates as the initial embedding method:

$$E = concat(FPFH(X^P), X^P), \tag{1}$$

$$FPFH(p) = S(p,q) + \frac{1}{|N(p)|}\sum_{q \in N(p)} w_q \cdot S(q,p), \tag{2}$$

where $p \in X^P$ and $N(p)$ represent the k-neighborhood of p. $S(p,q) = [\alpha, \phi, \theta]$ are the angle calculated between the vectors of p, q, and their estimated normals, representing the pose-invariant relationship between two points. Similarly, $S(q,p)$ is defined in the same manner. The weight w_q is the reciprocal of the Euclidean distance between p and q. In this way, E demonstrates enhanced robustness to noise and variations in data without being overly sensitive to small perturbations or sampling differences.

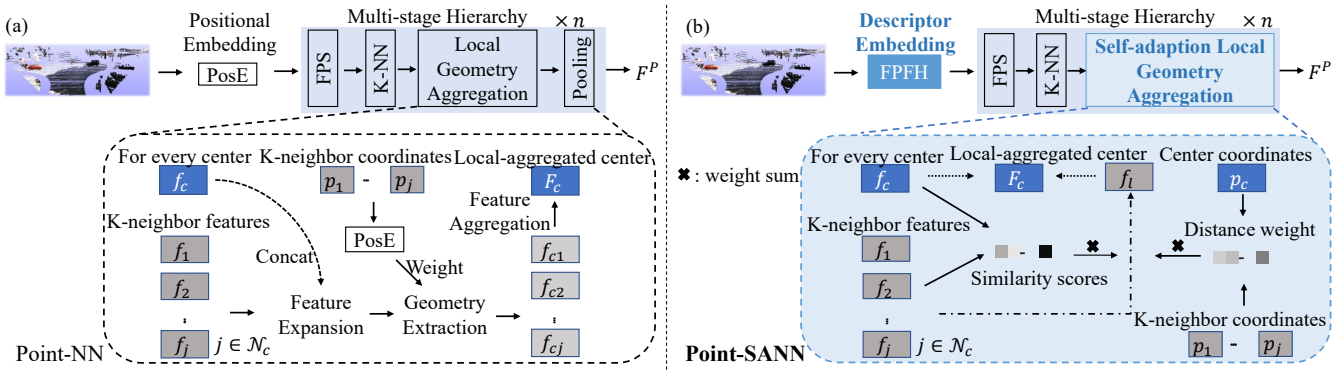

Figure 2: **The framework of (a) Point-NN and (b) Point-SANN.** The primary distinctions are as follows: (1) **Embedding Method.** In Point-SANN, we modify the embedding method by incorporating FPFH descriptors, which provide richer semantic information; (2) **Geometric Aggregation.** The incorporation of feature similarity scores and distance-weighted local features enhances the discriminative power for different geometric objects and improves robustness to variations in point cloud density.

Self-Adaption Local Geometry Aggregation. To further enhance the local features. In this procedure, E will be input into a multi-stage hierarchy to obtain point cloud features F^P, which mainly includes three modules: Farthest Point Sampling (FPS), K-Nearest Neighbors (KNN), and Self-adpation Local Geometric Aggregation. For each stage, we denote the input point cloud and feature from previous stage as $\{p_i, f_i\}_{i=1}^{M}$, where i is the point index and M denotes point number. FPS is adopted to downsample the point number from M to $\frac{M}{2}$ by selecting a subset of points from the point cloud in a way that ensures even distribution and maximum coverage:

$$\{p_c, f_c\}_{c=1}^{\frac{M}{2}} = FPS(\{p_i, f_i\}_{i=1}^{M}), \tag{3}$$

Then, k-NN is used to grouping k spatial neighbors for each center c:

$$\mathcal{N}_c = KNN(p_c, \{p_i\}), \tag{4}$$

where $\mathcal{N}_c$ represents the indices of k nearest neighbors.

Then we introduce two processes to incorporate local information, thereby enhancing the feature representation. Normally, if neighboring points and the center point do not originate from the same object, their feature descriptions E_i always exhibit differences. Regarding this prior, we tend to preserve similar feature to improve local feature expression:

$$W_i^s = cos(f_c, f_i),\ i \in \mathcal{N}_c, \tag{5}$$

where W^s denotes the similarity score between the center point and each neighbor point, and cos represents the cosine similarity operation. In addition, considering the issue of uneven distribution of point cloud density, we also take into account the distance between points within local region:

$$W_i^d = exp(-ED(p_c, p_i)^2/(2 * \sigma^2)), i \in \mathcal{N}_c, \tag{6}$$

where ED represents the Euclidean distance between center point p_c with neighbor points p_i and σ is a hyperparameter. Then the local features of the center point can be obtained by weighting the features of neighboring points:

$$f_l = \sum_{i \in \mathcal{N}_c} W_i^s \times (W_i^d \times f_i), \tag{7}$$

where $\times$ represent matrix multiplication. Finally, integrating the two features f_l with f_c to obtain aggregated features F_c:

$$F_c = (1 - \lambda) f_c + \lambda f_l, \tag{8}$$

where $\lambda \in [0, 1]$ is a balance factor and we set $\lambda = 0.25$ in our experiments.

Feature Decoding. Starting from the last stage of encoding, upsample the point cloud features step by step. Interpolate the aggregated center point features and assign features to the points before downsampling:

$$F_i = \sum_{c \in \mathcal{N}_i} W_c^d F_c,$$
$$W_c^d = \frac{1}{ED(p_i, p_c)},\ \mathcal{N}_i = KNN(p_i, \{p_c\}), \tag{9}$$

the features output from the final stage is F^P.

Hash Code Learning Module

To integrate multi-modal data, we project each feature into a shared space using a randomly initialized anchor for uniform dimensions. These features are then encoded into a hash space using a binary hash encoder. A unified hash code is generated by aggregating the weights of each view, as shown in Fig. 3. This approach finds the optimal clustering center and learns a generalized representation, which can be summarized by two optimization functions:

Collaborative Discrete Representation Learning (CDRL). For data point x_s^v from the v-th view, define the binary hash function as:

$$h_s^v(f(x_s^v), U^v) = sgn(U^v f(x_s^v)), \tag{10}$$

where $sgn(\cdot)$ is an element-wise sign operator, $f(x_s^v) \in \mathbb{R}^m$ indicates a nonlinear embedding constructed from a randomly initialized anchor, s denote the s-th point, and $U^v \in \mathbb{R}^{l \times m}$ is the mapping matrix for the v-th view. Here l is the length of the hash code and m is the number of anchor sam-

ples. The objective of CDRL is formulated as:

$$\begin{aligned}
&\min_{U^v,b_s,a} \sum_{v=1}^{M}(a^v)^r \\
&(\sum_{s=1}^{n}||b_s - h_s^v||_F^2 + \beta||U^v||_F^2 - \gamma\sum_{s=1}^{n} var(h_s^v)) \\
&s.t.\ \sum_v a^v = 1;\ a^v > 0;\ b_s \in \{-1,1\}^{l\times 1}, \qquad (11)
\end{aligned}$$

where b_s is the collaborative binary code for the s-th point, $a = [a^1, ..., a^V] \in \mathbb{R}^V$ is a nonnegative normalized weighting vector balancing the significance of different views, $r > 1$ is a scalar controlling the weights, and β, γ are both nonnegative constant. $||\cdot||_F$ denotes Frobenius norm. The first term ensures the learning of a unified binary code across different views, while the second and third terms contribute to achieving stable solutions and balanced bit distributions, respectively.

Binary Clustering Structure Learning (BCSL). To maintain consistent cluster structures among different views, the following equation is used to construct the BCSL scheme:

$$\begin{aligned}
&\min_{C,g_s} ||b_s - Cg_s||_F^2 \\
&s.t.\ C^T\mathbf{1} = 0;\ C \in \{-1,1\}^{l\times c}; \\
&\quad g_s \in \{0,1\}^c;\ \sum_i g_{is} = 1, \qquad (12)
\end{aligned}$$

where C and g_s are the clustering centroids and indicator vector, respectively. $\mathbf{1}$ represents the all-ones vector.

Overall Objective Function. Combining both parts, the overall objective function is formulated as:

$$\begin{aligned}
&\min F(U^v, B, C, G, a) \\
&= \sum_{v=1}^{M}(a^v)^r(||B - U^v f(x^v)||_F^2 + \beta||U^v||_F^2 \\
&- \frac{\gamma}{n}tr((U^v f(x^v))(U^v f(x^v))^T)) + \epsilon||B - CG||_F^2 \\
&s.t.\ C^T\mathbf{1} = 0; \sum_v a^v = 1; a^v > 0; B \in \{-1,1\}^{l\times n}; \\
&\quad C \in \{-1,1\}^{l\times c}; G \in \{0,1\}^{c\times n}; \sum_i g_{is} = 1, \qquad (13)
\end{aligned}$$

where $B = [b_1, ..., b_n], G = [g_1, ..., g_n]$, and ϵ is the regularization parameter. n is the number of points.

The optimization process involves iteratively updating several key components while keeping others fixed. For more details, please refer to BMVC (Zhang et al. 2018). Different from its original purpose, our method performs fine-grained point-wise operation within a single scene, rather than simple object-level clustering. After these processes, multi-modality features with different expressions are projected into a same hash representation B, fully integrating the information of each view.

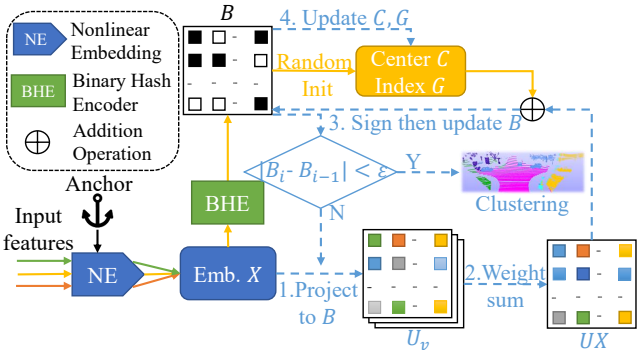

Figure 3: **The process of HCLM.** The initial variables are constructed via the non-blue solid line process. The Anchor initializes all view information into a common space, which is then transformed into the hash space. The binary code B and cluster center C are iteratively refined through the blue dashed line process.

Experiments

Datasets

In this study, we utilize two classical publicly available outdoor datasets for evaluation: nuScenes (Caesar et al. 2020) and SemanticKITTI (SKITTI) (Behley et al. 2019). Both datasets provide synchronized RGB and 3D point cloud pairs, along with calibration parameters that facilitate seamless 2D to 3D projection. We compare our proposed method against existing multi-modal UDA, TTA, and ZSL methods. This comparison helps validate the effectiveness of our approach in handling complex, real-world data without requiring extensive labeled datasets or high computational costs.

When compare with UDA and TTA methods, our evaluation covers four primary configurations: **nuScenes** with 6 classes, and **SKITTI** with 6, 9, 10 classes. For the UDA experiments, we evaluate on setups **SKITTI-nuScenes**, **VKITTI-SKITTI** and **A2D2-SKITTI** with 6, 6, 10 classes, respectively. VirtualKITTI (VKITTI) (Gaidon et al. 2016) and A2D2 (Geyer et al. 2020) are datasets frequently employed in autonomous driving research. In contrast, the TTA scenarios closely mirror those of UDA but substitute the **VKITTI-SKITTI** setup with **Synthia-SKITTI**, which includes 9 classes. Synthia (Ros et al. 2016) is another widely adopted outdoor dataset. To maintain fairness, we re-evaluated the TTA method on the test set, rather than the

Method [Type]	nuScenes mIoU	SKITTI mIoU
3DGenZ(Michele et al. 2021)[ZSL]	3.2	6.5
TGP(Chen et al. 2023)[ZSL]	14.1	13.3
SMKM(Lu et al. 2023)[ZSL]	**22.4**	**23.9**
xMHashSeg (ours)[TF]	**24.4**	**24.2**

Table 1: Quantitative results (mIoU, %) compared to all settings of **ZSL**. Bold indicates the highest value in each type.

Method	Type	Train	nuScenes(6 class) [source] mIoU	SKITTI(6/9´ class) [source] mIoU	SKITTI(10 class) [source] mIoU
xMUDA(Jaritz et al. 2020)	UDA	√	[SKITTI] ***64.0**	[VKITTI] 48.2	[A2D2] 44.0
CLIP2UDA(Wu et al. 2024b)	UDA	√	[SKITTI] *58.9	[VKITTI] 60.4	[A2D2] 50.0
LMSAM(Peng et al. 2024)	UDA	√	-	[VKITTI] **64.9**	[A2D2] **52.1**
FtD(Wu et al. 2025)	UDA	√	[SKITTI] *55.3	[VKITTI] 52.6	[A2D2] 51.1
MM-TTA(Shin et al. 2022)	TTA	√	[SKITTI] ***55.0**	[Synthia] *37.5´	[A2D2] *51.1
Latte(Cao et al. 2024)	TTA	√	[SKITTI] *54.5	[Synthia] ***39.6´**	[A2D2] ***53.4**
xMHashSeg (ours)	TF	×	[N/A] **40.9**	[N/A] **32.7(24.4́)**	[N/A] **23.5**

Table 2: Quantitative results (mIoU, %) compared to all settings of **UDA and TTA**. ´´ ' indicates different category setting in TTA. '*' indicates the reproduced results. Bold indicates the highest value in each type.

entire dataset as reported in the original publication.

For ZSL methods, we compare them in two scenarios: **nuScenes** and **SKITTI**, each involving another 6 classes which are unseen classes in ZSL methods.

Implementation Details

For DINOv2, we utilize the 'ViT-B' pre-trained model as the 2D feature extractor. On the other hand, DAM employs the 'ViT-L' model trained on 'VKITTI', which is particularly well-suited for outdoor scenes.

When evaluating performance, we use the Hungarian algorithm to optimally match the clustering results with the ground truth labels before calculating the mIoU value.

The number of iterations for hash code generation is set to 300, with an early stopping mechanism that triggers when the binary code B no longer changes. Under the UDA or TTA class setting, the anchor dimension m is configured to be 500, and the hash code length L is set to 32. In the comparison setting with ZSL, we set the anchor dimension to 30 and the hash code length L to 8.

All experiments were performed on a system equipped with a single NVIDIA RTX 3090 GPU. The average inference efficiency of our method is about 1.1 seconds per scene under nuScenes dataset and about 1.5 seconds per scene under SKITTI dataset, outperforming Latte's inference efficiency of 1.2 seconds and 2.7 seconds.

Results and Analysis

According to our understanding, the order of conditional sufficiency between UDA, TTA, ZSL and TF method is roughly: UDA $\approx$ TTA $>$ ZSL $>$ TF. This is because UDA or TTA have fully supervised data on other domain, while ZSL has fully supervised data on other classes in the same domain. In contrast, our TF method xMHashSeg performs segmentation completely unsupervised and without additional training, which is obviously the most demanding condition. By comparing with these methods, we aim to verify that our approach can achieve competitive results under more challenging conditions.

We compare the mean Intersection over Union (mIoU) of our TF method with multi-modal UDA, TTA and ZSL methods in outdoor scenes to demonstrate the efficiency and scene generalization ability of our method. The comparison results are shown in Tab. 1 and Tab. 2.

Comparison with ZSL methods As shown in Tab. 1, it demonstrate that our method outperforms ZSL methods in unseen classes. The mIoU values of our method are 24.4% and 24.2% in nuScenes and SKITTI respectively, which are 2 and 0.3 points higher than the SOTA multi-modal ZSL method. This comparison is especially pertinent because the ZSL method has the closest conditions to the TF method but is more advantageous, the improvement in performance here further demonstrates the effectiveness of our method.

Comparison with UDA and TTA methods From Tab. 2, it can be seen that although our method does not perform as well as UDA and TTA methods, it also demonstrates that our method has the potential to achieve similar performance under more limited conditions.

Overall, our method eliminates the need for labor-intensive labeling and high computational training. It operates without requiring any pre-existing dataset, enabling direct inference on arbitrary scene images and making it highly effective across a wide range of real-world scenarios.

Fig. 4 visualizes the qualitative semantic segmentation results of first two settings with 6 classes. It indicate that our method achieves promising overall segmentation accuracy, effectively delineating major object boundaries and regions. However, finer details such as small objects and intricate edges are not captured as accurately. Future work will focus on improving the model's ability to handle fine details.

Ablation Study

In this section, we analyze the effectiveness of each component in our method through ablation study. In order to improve ablation efficiency, we split the Night scenes (a subset) from nuScenes datasets and use it for ablation study, using the same class settings as xMUDA.

Effectiveness of Point-SANN To assess the effectiveness of the improved Point-SANN, we performed an ablation study where all other variables were kept constant, and the original Point-NN was used to extract point cloud features. The segmentation performance was subsequently evaluated. The results, presented in the first row of Tab. 3, show that using the original Point-NN resulted in a mIoU score of 32.1%.

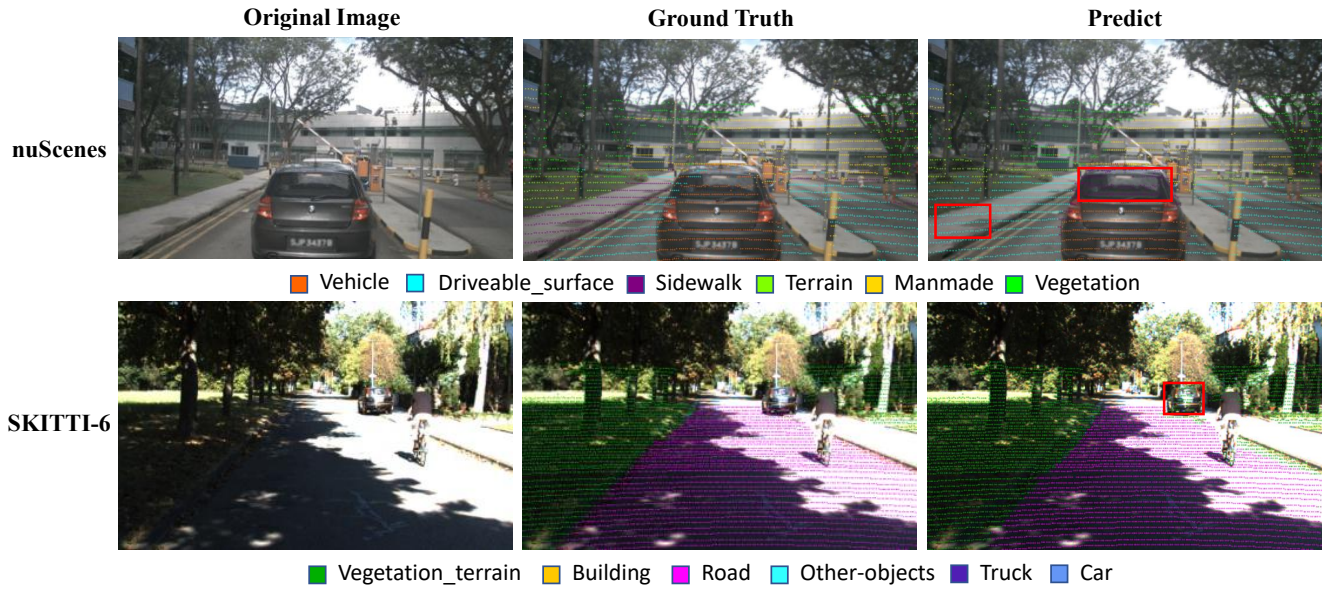

Figure 4: Qualitative results of our method. The incorrectly segmented area is highlighted by a red rectangle.

Model	Point-SANN	Depth	Hash Learning	mIoU (%)
#1		√	√	32.1
#2	√		√	36.9
#3	√	√		33.9
#4	√	√	√	**41.7**

Table 3: Ablation study on the effectiveness of important components of our methods.

This score is 9.6 points lower than the mIoU achieved with the improved Point-SANN, clearly indicating the enhanced performance of the updated model.

Role of Depth modality To evaluate the contribution of depth information to our framework, we removed it and assessed performance using only two views: 2D images and 3D point clouds. The results, presented in the second row of Tab. 3, show a mIoU value of 36.9%. This is approximately 4.8 points lower than the 41.7% achieved when depth information is included. These findings highlight the critical role of depth maps in bridging the gap between 2D images and 3D point clouds. Even though depth maps are estimated values generated through our DAM, they significantly enhance the overall segmentation performance. This underscores the importance of incorporating depth information for achieving more accurate multi-modal feature integration.

Effectiveness of Hash Learning To verify the effectiveness of learning a unified hash expression for clustering, we compared our method with k-means. Specifically, we first normalized the three features separately to ensure consistent scales before concatenating them. We then evaluated their segmentation performance using k-means clustering. The results, presented in the third row of Tab. 3, show an mIoU value of 33.9% for the k-means approach. This is 7.8 points lower than the mIoU achieved by learning a hash code for clustering. This underscores the significance of aligned feature representation in achieving superior performance.

λ	0	**0.25**	0.5	0.75	1
mIoU (%)	37.1	**41.7**	38.7	37.8	37.6

Table 4: Parameter sensitivity analysis of balance factor λ.

Parameter Sensitivity Analysis To evaluate the sensitivity of our method to hyper-parameters, we conducted a sensitivity analysis on the Night scene. Specifically, we analyzed the impact of the balance factor λ during feature aggregation in Point-SANN. The results, presented in Tab. 4, demonstrate that the optimal performance is achieved when $\lambda = 0.25$. This optimal setting was chosen for all configurations in our experiments.

Conclusion

In conclusion, this paper introduces xMHashSeg, a novel training-free framework for unsupervised LiDAR semantic segmentation using cross-modal hash learning. xMHashSeg leverages foundation models for 2D feature extraction and captures robust 3D features from raw point clouds via developed Point-SANN. It reconciles feature scale inconsistencies across views through HCLM and enhances hash code convergence by introducing depth maps as a bridge. Experimental results on multi-modality outdoor datasets show that xMHashSeg outperforms ZSL methods and approaches UDA and TTA performance without requiring annotations or additional training. Its strong generalization and adaptability make it a promising solution for real-world applications with unlabeled data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62176224 and Grant 62306165; in part by the Science and Technology on Sonar Laboratory under Grant 2024-JCJQ-LB-32/07; in part by the Fundamental Research Funds for the Central Universities under Grant 20720250031.

References

Ali, M.; and Khan, S. 2023. Clip-decoder: Zeroshot multi-label classification using multimodal clip aligned representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4675–4679.

Alonso, I.; Riazuelo, L.; Montesano, L.; and Murillo, A. C. 2020. 3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation. *IEEE Robotics and Automation Letters*, 5(4): 5432–5439.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cao, H.; Xu, Y.; Yang, J.; Yin, P.; Ji, X.; Yuan, S.; and Xie, L. 2024. Reliable spatial-temporal voxels for multi-modal test-time adaptation. In *European Conference on Computer Vision*, 232–249.

Chen, R.; Zhu, X.; Chen, N.; Li, W.; Ma, Y.; Yang, R.; and Wang, W. 2023. Bridging language and geometric primitives for zero-shot point cloud segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5380–5388.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4340–4349.

Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A. S.; Hauswald, L.; Pham, V. H.; Mühlegg, M.; Dorn, S.; et al. 2020. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*.

Jaritz, M.; Vu, T.-H.; Charette, R. d.; Wirbel, E.; and Pérez, P. 2020. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12605–12614.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kong, L.; Ren, J.; Pan, L.; and Liu, Z. 2023. Lasermix for semi-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21705–21715.

Lu, Y.; Jiang, Q.; Chen, R.; Hou, Y.; Zhu, X.; and Ma, Y. 2023. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21674–21684.

Mei, J.; Gao, B.; Xu, D.; Yao, W.; Zhao, X.; and Zhao, H. 2019. Semantic segmentation of 3D LiDAR data in dynamic scene using semi-supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(6): 2496–2509.

Michele, B.; Boulch, A.; Puy, G.; Bucher, M.; and Marlet, R. 2021. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision*, 992–1002. IEEE.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 1–32.

Peng, D.; Lei, Y.; Li, W.; Zhang, P.; and Guo, Y. 2021. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7108–7117.

Peng, X.; Chen, R.; Qiao, F.; Kong, L.; Liu, Y.; Sun, Y.; Wang, T.; Zhu, X.; and Ma, Y. 2024. Learning to adapt sam for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, 54–71.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qian, Q.; and Hu, J. 2024. Online zero-shot classification with clip. In *European Conference on Computer Vision*, 462–477.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16928–16937.

Szalai-Gindl, J. M.; and Varga, D. 2024. FPFH Revisited: Histogram Resolutions, Improved Features, and Novel Representation. *IEEE Access*, 12: 67325–67354.

Veasey, B. P.; and Amini, A. A. 2024. Parameter-Efficient Fine-Tuning of DINOV2 Vision Transformers for Lung Nodule Classification. In *2024 IEEE International Symposium on Biomedical Imaging*, 1–5. IEEE.

Wu, B.; Zhou, X.; Zhao, S.; Yue, X.; and Keutzer, K. 2019. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 international conference on robotics and automation*, 4376–4382. IEEE.

Wu, Y.; Xing, M.; Zhang, Y.; Luo, X.; Xie, Y.; and Qu, Y. 2024a. Unidseg: Unified cross-domain 3d semantic segmentation via visual foundation models prior. *Advances in Neural Information Processing Systems*, 37: 101223–101249.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; Peng, K.; and Qu, Y. 2025. Fusion-then-Distillation: Toward Cross-modal Positive Distillation for Domain Adaptive 3D Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–17.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024b. Clip2uda: Making frozen clip reward unsupervised domain adaptation in 3d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8662–8671.

Yang, Y.; Hayat, M.; Jin, Z.; Zhu, H.; and Lei, Y. 2023. Zero-shot point cloud segmentation by semantic-visual aware synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11586–11596.

Yuan, Z.; Cheng, M.; Zeng, W.; Su, Y.; Liu, W.; Yu, S.; and Wang, C. 2023. Prototype-guided multitask adversarial network for cross-domain LiDAR point clouds semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhang, R.; Wang, L.; Wang, Y.; Gao, P.; Li, H.; and Shi, J. 2023. Starting from non-parametric networks for 3d point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5344–5353.

Zhang, Y.; Hu, R.; Li, R.; Qu, Y.; Xie, Y.; and Li, X. 2024. Cross-modal match for language conditioned 3d object grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7359–7367.

Zhang, Y.; Lan, Y.; Xie, Y.; Li, C.; and Qu, Y. 2025. Cross-cloud consistency for weakly supervised point cloud semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15520–15528.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2018. Binary multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1774–1782.