2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

# X-MOS: A Heterogeneous Cross-LiDAR Generalization Framework for Moving Object Segmentation

Minjae Lee[1], Ilhwan Ha[1], Sang-Min Choi[1], Gun-Woo Kim[1,*], and Suwon Lee[1,*]

***Abstract*— Moving object segmentation (MOS) is foundational for autonomous vehicle safety. However, the increasing diversity of LiDAR sensors creates a significant domain shift problem, causing models trained on one sensor to perform poorly when deployed on another. A naive approach of training on combined data from heterogeneous sensors leads to a biased model that favors high-density sensors while failing on sparse, low-resolution sensors. To address this issue, we propose X-MOS, a novel generalization framework based on multi-teacher knowledge distillation. X-MOS generates sensor-specific expert teacher models and employs a sensor-aware knowledge distillation strategy. This strategy uses the sensor type as privileged information to activate the most appropriate teacher at each training step, providing unambiguous learning signals to a single student model. Extensive experiments on the HeLiMOS dataset, which comprises four different LiDAR sensors, demonstrate the effectiveness of our framework. X-MOS mitigates training bias and achieves an overall test mIoU of 0.717, outperforming both naive training and the best individual expert teacher. Notably, it more than doubles the performance on the most challenging low-channel sensor. Furthermore, our model exhibits strong zero-shot generalization to unseen datasets with similar sensor types. This work provides a robust and scalable methodology for achieving cross-sensor generalization, which is foundational for more practical and adaptable perception systems in autonomous driving.**

## I. Introduction

Autonomous driving has rapidly advanced in recent years, bringing significant convenience to human society. A central challenge in its commercialization lies in ensuring a safe driving environment. To achieve this, autonomous vehicles must accurately perceive and respond to moving objects in their surroundings using various sensors, such as cameras, radar, and LiDAR. Moving object segmentation (MOS), which segments moving objects from 3D point clouds, is instrumental in understanding dynamic environments.

Numerous preceding studies, including LMNet [1], 4DMOS [2], MapMOS [3], RVMOS [4], and StreamMOS [5] developed and validated their models on SemanticKITTI [6], [7] dataset and achieved excellent performance. For additional experiments, these studies also evaluated the performance on other datasets, such as nuScenes [8], Apollo [9], and the Sipailou Campus [10] datasets.

However, the recent LiDAR sensor market has been diversifying with the release of products featuring various channel counts, fields of view, and scan patterns to meet the demands of cost-effectiveness and specific applications. This diversity led to a domain-shift problem between datasets constructed using conventional high-quality sensors and those from new sensors because deep learning models are sensitive to such domain shifts and tend to exhibit performance degradation [11], [12]. HeLiMOS [13] experimentally demonstrated that the differences in LiDAR sensor specifications were a major factor hindering the generalization performance of the models. Specifically, a model trained on SemanticKITTI achieved an intersection over union (IoU) below 0.06 on Velodyne's 16-channel sensor, clearly indicating a severe domain gap.

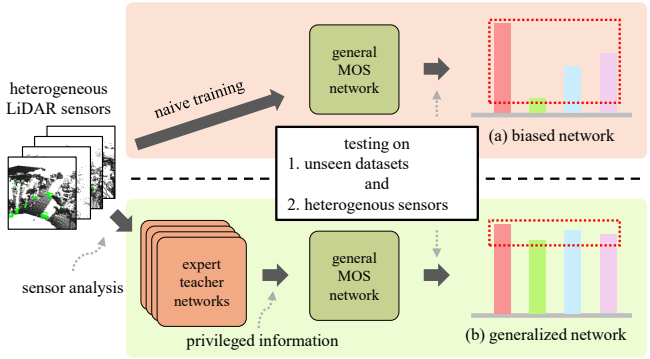



Fig. 1: Comparison of the training methods between a naive approach and the proposed X-MOS framework. A model trained naively on heterogeneous-sensor data becomes biased toward a specific sensor. In contrast, the proposed framework uses multiple expert teachers and privileged information to guide the MOS network to better generalization.

This performance degradation stems from differences in the physical specifications related to the sensor resolution, such as the number of channels, field of view (FoV), and scan patterns, which directly affect the density and distribution of point clouds. This causes a model trained on existing high-quality datasets to encounter unfamiliar input data, thereby causing a severe domain shift during the inference process. Therefore, for the practical dissemination of autonomous driving technology, it is essential to achieve robust general performance across these varied sensor specifications without training bias toward any particular sensor data.

A simple and naive approach (Fig. 1(a)) for addressing this issue involves using data from all available sensors for training. However, as discussed in Sec. IV, although this method contributes to an overall improvement in metrics, it introduces bias. From the perspective of sparse and dense sensors, numerical improvements tend to be skewed toward maximizing the performance of dense sensors, even while

*Corresponding authors: Gun-Woo Kim and Suwon Lee.
[1]The authors are with the Department of Computer Science and Engineering, Gyeongsang National University, Jinju-si 52828, South Korea.

This research was supported by the Regional Innovation System & Education (RISE) program through the RISE Center, Gyeongsangnam-do, funded by the Ministry of Education (MOE) and the Gyeongsangnam-do Provincial Government, Republic of Korea. (2025-RISE-16-001)

trading off the performance of sparse sensors.

Therefore, to overcome these limitations, we propose X-MOS, a framework that achieves a stable and well-rounded performance in heterogeneous LiDAR sensor environments (Fig. 1(b)) without bias toward any specific sensor. The main contributions of this study are as follows:

- **Multi-teacher learning**: We generated sensor-specific expert teacher models and trained a student model that avoided the training bias and performance tradeoffs inherent in naive single-model training.
- **Sensor-aware knowledge distillation**: We leverage the sensor type as privileged information during training to guide a single student model, enabling it to effectively absorb knowledge from multiple expert teachers without confusion.
- **Cross-sensor and cross-dataset generalization**: Through extensive experiments, we verify that our framework prevents models from overfitting to a specific sensor or dataset, thereby demonstrating practical robustness in real-world heterogeneous hardware environments.

## II. Related Work

### A. General Approaches in Moving Object Segmentation

Research on MOS has been conducted through various approaches that can be broadly categorized into projection-based and non-projection-based methods. Studies on LMNet, RVMOS, and MF-MOS [14] projected 3D point clouds onto range view (RV) images. They proposed various methods to accumulate these images over time and calculated their differences to segment the moving objects. StreamMOS highlighted the limitations of using only single RV images and proposed a multi-view technique that incorporates bird's-eye view (BEV) images alongside RV images.

However, these projection-based ideas are designed with prior consideration of the channel count (i.e., vertical resolution) of the spinning (omni-directional) LiDAR and the height resolution of the RV image. This directly leads to degradation of the image representation when projecting from a low-channel-spinning LiDAR, as multiple blank horizontal lines appear in the RV image. Furthermore, beyond the issue of vertical resolution, utilizing LiDARs with different scan patterns (e.g., Livox LiDAR) results in projected images with different patterns, unless the point cloud is extremely dense. This makes existing models highly susceptible to performance degradation caused by domain shifts.

Furthermore, when using sensors with different fields of view, creating an RV image similar to a 360-degree spinning LiDAR results in large empty regions, which presents a limitation that necessitates the redefinition of the RV image projection method for each sensor.

In contrast, non-projection-based MOS approaches, such as 4DMOS and MapMOS, stack the 3D point cloud along the time axis to use 4D spatiotemporal data as the input. This does not require different input processing depending on the sensor, unlike the projection-based methods, the approach faces several challenges. The unique scan patterns and resolutions of different sensors lead to distinct spatial point distributions, and, consequently, varied voxel representations. Therefore, a model must learn generalizable spatiotemporal patterns that are invariant to these sensor-specific artifacts, rather than overfitting the data structure of a particular sensor.

### B. Generalization in LiDAR Perception Tasks

In the MOS field addressed in this study, generalization approaches for heterogeneous sensors have not yet been actively researched. However, broader LiDAR-based technologies, such as 3D object detection [15], LiDAR semantic segmentation (LSS) [16], and simultaneous localization and mapping (SLAM) [17], efforts to resolve issues related to sensor-resolution differences and specification changes have been relatively persistent.

For instance, Wei et al. [18] identified a significant domain gap problem between high- and low-resolution sensors for 3D object detection. To avoid having to train a model from scratch using low-resolution data, they proposed a curriculum-learning approach. This method starts training with high-resolution data and applies progressive downsampling based on the vertical and horizontal angles, $\theta$ and $\phi$. They showed that by guiding the model to first learn the essential features of objects from information-rich data and then gradually adapt to sparser data, a stable generalization performance can be achieved.

The "Complete & Label" [19] in the LSS domain presented another novel methodology for domain generalization. As the name suggests, this study introduced a step to first complete sparse and incomplete point clouds, which vary depending on the sensor type at the voxel level. By generating this standardized, dense input data, the subsequent segmentation network learns domain-agnostic representations that are independent of the sensor resolution or scan pattern.

Although this allows us to achieve high segmentation performance, it incurs significant computational costs during inference and imputation. Meanwhile, KISS-SLAM [20] is a case of achieving robustness by reducing dependence on deep learning. Following the "Keep It Small and Simple" design philosophy, it minimizes the parameters of the deep learning model and rather adopts an approach based on classical geometric principles. Geometric principles can be universally applied regardless of the statistical properties of the sensor, such as the point density or distribution. Therefore, it achieves excellent generalization performance without being constrained by the input sensor or requiring a separate domain. adaptation process.

### C. Generalization Methods based on Knowledge Distillation

The concept of knowledge distillation begins with model compression as proposed by Hinton [21]. However, it has recently evolved into a method for resolving data distribution gaps between domains, such as different LiDAR datasets, or for enhancing the cooperative performance between modalities, such as LiDAR-to-camera [22]–[25]. Therefore, numerous studies have demonstrated that knowledge distillation is

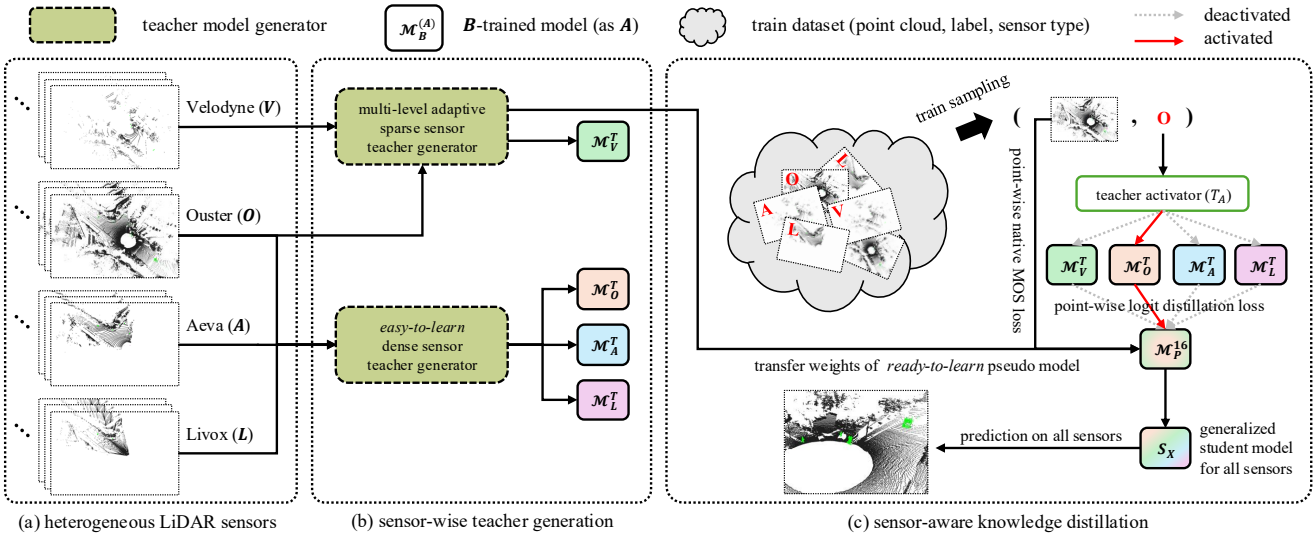



Fig. 2: Overview of the proposed X-MOS framework. (a) Given co-located sequences from heterogeneous LiDAR sensors, (b) expert teacher models are trained for each sensor type, tailored to its characteristics. (c) For an arbitrary sensor input, a specific teacher is selected using the sensor type as privileged information. The student model then absorbs knowledge from the teacher to become the final generalized model.

highly effective for domain adaptation and sensor fusion, significantly enhancing model performance.

Among these advancements, the multi-teacher knowledge distillation (MTKD) paradigm has emerged, which moves beyond the conventional single-teacher-student structure by utilizing domain-expert teachers [26], [27]. MTKD helps the student model secure robust generalization performance preventing it from being biased toward a specific domain, by integrating knowledge from multiple teacher models specialized for different data or tasks.

Furthermore, studies based on privileged information, which leverage additional information available only during the training process and not during inference, have focused on enhancing the efficiency of knowledge distillation [28]–[30]. Zhao et al. [30] defined this as the 'learning using privileged information (LUPI)' paradigm and proposed a knowledge distillation method that maximizes learning efficiency.

In a heterogeneous sensor environment, the sensor type and its specifications can be used as powerful privileged information. Therefore, an approach that effectively utilizes the expertise of each teacher in a multi-teacher environment based on privileged information remains an important research direction for the unresolved heterogeneous sensor generalization issue.

In summary, while promising generalization techniques such as knowledge distillation and privileged information have been explored in other fields such as LSS, most existing MOS studies have not directly addressed the issue of sensor heterogeneity, rather focusing on improving single-model performance. To overcome these limitations, a generalization framework based on knowledge distillation that leverages the sensor identity as privileged information naturally emerges.

## III. Methodologies

### A. Heterogeneous LiDAR Sensor Analysis

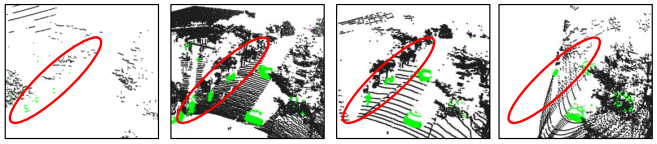



Fig. 3: Qualitative analysis of heterogeneous LiDAR sensors. From left: Velodyne (V), Ouster (O), Aeva (A), and Livox (L). Moving points are highlighted in green.

The HeLiMOS dataset serves as the foundation of this study by providing data for the same location scanned by four different heterogeneous LiDAR sensors. Figs. 3(a) to (d) show the point clouds collected by the Velodyne VLP-16 (16-channel), Ouster OS2-128 (128-channel), Aeva Aeries II, and Livox Avia sensors. These are defined as V, O, A, and L in Fig. 2. Each sensor has different physical specifications, leading to significant variations in the data characteristics. When qualitatively comparing the difficulty of the MOS in terms of density, the O, A, and L sensors shown in Figs. 3(b), (c), and (d) represent the shapes of moving pedestrians or cars with relative clarity (see green points).

By contrast, as shown in Fig. 3(a), low-resolution V sensor data exhibits extreme point sparsity for dynamic pedestrians, which worsens at greater distances. This range-dependent sparsity makes recognition difficult for both human observers and deep learning models [6]. While V and O differ significantly in vertical resolution (128 vs. 16 channels), A and L are limited by a front-only FoV. Training these diverse sensors together with a general MOS network degrades generalization and biases the model toward easier sensors. Therefore, we utilize dedicated expert teachers for each sensor to ensure stable and reliable learning.

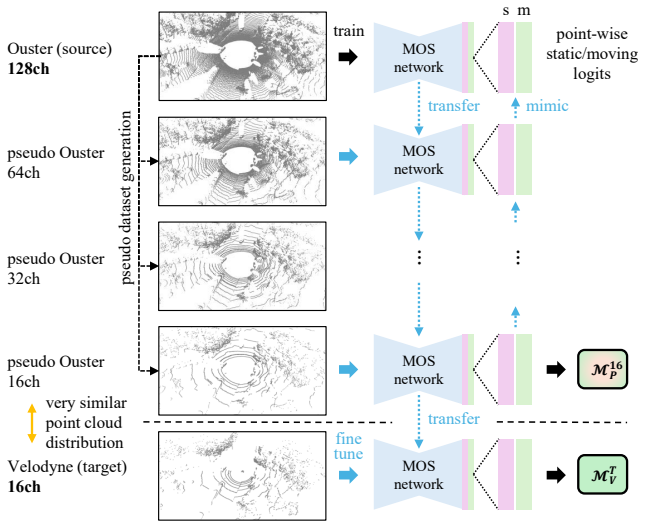



Fig. 4: The process of the multi-level adaptive teacher generator. To facilitate multi-stage adaptation from a high-resolution source sensor to a low-resolution target sensor, data with pseudo-resolutions are generated. This process guides the model for the lower-resolution sensor to mimic the output of the model for the higher-resolution sensor.

### B. Multi-Teacher Generation

The proposed framework generates expert teacher models for each sensor by considering the data density and corresponding learning difficulty. For the relatively easy-to-learn sensors O, A, and L, which have higher data densities, teacher models were created using an *Easy-to-Learn Teacher Generator*. For the significantly more challenging sensor V, which has extremely low density, an expert teacher was built using a *Multi-Level Adaptive Teacher Generator*.

We divided this training strategy into two because applying a single complex approach to all sensors would be inefficient. High-density sensors can achieve expert status through simple supervised learning, whereas low-density sensors require a progressive curriculum to ensure stable learning in a sparse data environment.

*1) Easy-to-Learn Teacher Generator:* Our proposed framework is designed to be universally applicable to any arbitrary 3D MOS network, without dependence on a specific architecture. Expert teachers with high density sensors O, A, and L ($M_O^T, M_A^T, M_L^T$) are generated by training the original MOS network architecture directly on each sensor's respective dataset without modification.

*2) Multi-Level Adaptive Teacher Generator:* The expert teacher for low-density sensor V also utilizes the original MOS network, but it incorporates a progressive curriculum into the training process. This method aims to enable the model to learn effectively from a sparse dataset, such as V's by using gradual adaptation to more difficult tasks. The entire process is depicted in detail in Fig. 4.

First, a model ($M_O^{128}$) was pre-trained using the easy-to-learn method from Sec. III-B.1 on the dataset from the high-density sensor O, which has a scan pattern most similar to V, except for the vertical resolution. Subsequently, rather than drastically reducing the density from 128 channels to 16, we sequentially generate pseudo low-density data ($P_{low}$) by progressively reducing the number of channels (e.g., $128 \rightarrow 64 \rightarrow 32 \rightarrow 16$). This process is defined by the channel down operator $\Psi$ as follows:

$$P_{low} = \Psi(P_{high}; K, M, p) \quad (1)$$

Here, the operator $\Psi$ represents the process of converting a high-resolution point cloud ($P_{high}$) into a low-resolution one ($P_{low}$). Specifically, the operator first calculates the elevation angle ($\theta$) for every point in the input cloud.

Inspired by [18], the operator performs K-Means clustering [31] at these angles to approximate the original beam structure with $K$ distinct clusters, each representing a vertical beam. This is because, rather than simply dropping points randomly, this approach mimics the physical structure of a real low-channel LiDAR sensor that emits laser beams at discrete, physically determined elevation angles.

From these $K$ clusters, the operator selects the target number of beams, $M$, by choosing clusters at uniform intervals based on their sorted centroids. Finally, to match the density of a native low-resolution sensor, the sensor probabilistically samples the points belonging to these selected beams with a keep probability of $p$, generating the final low-resolution point cloud ($P_{low}$).

Next, as shown by the first downward light blue dashed arrows in Fig. 4, the weights from the pre-trained 128-channel model ($M_O^{128}$) were transferred to a model intended for 64-channel training. In this step, the 64-channel model receives the pseudo 64-channel data ($P_{low}$) as input, whereas the 128-channel model receives the original high-density data ($P_{high}$). Knowledge distillation was then performed to guide the low-channel model to mimic the output of the high-channel model by defining the loss function as follows:

$$L_{mimic} = \lambda_1 L_{MOS}(y_{low}, \hat{y}_{low}) + \lambda_2 L_{KD}(\hat{y}_{high}, \hat{y}_{low}) \quad (2)$$

where $y_{low}$ is the ground truth label for the pseudo 64-channel data $P_{low}$, while $\hat{y}_{low}$ and $\hat{y}_{high}$ are the point-wise logits predicted by the student (64-channel) and teacher (128-channel) models, respectively. $L_{MOS}$ denotes a loss function of native MOS model and $L_{KD}$ represents the knowledge distillation loss that encourages the student's output to mimic that of the teacher.

The hyperparameters $\lambda_1$ and $\lambda_2$ are weighting factors that balance these two terms. The reason for defining the loss function in this simple form ensures the generality of the framework. As most MOS architectures output point-wise logits in their final layers, this method, which does not depend on complex internal feature maps, can be applied to any MOS model without modification.

By repeating this process for 16 channels, a model trained on pseudo 16-channel data ($M_P^{16}$) is generated. This model's weights are used to initialize the student model effectively and powerfully (see Sec. III-C and Sec. IV-E). Finally, after completing the training down to 16 channels with pseudo data, the model is fine-tuned on the real 16-channel V sensor dataset to create the final V expert teacher ($M_V^T$).

TABLE I: LiDAR performance comparison. Results from training, validating, and testing on the HeLiMOS dataset with the 4DMOS network without any special techniques (native method). The *Configuration Name* is an alias for the train-validation target pair. The best performance for each single sensor is marked in bold. The best model is chosen based on the highest mIoU (Val) score.

| #Case | Configuration Name | Train Target | Validation Target | mIoU (Val) | Omni-Directional LiDARs (Test) | | | Solid-State LiDARs (Test) | | | Test mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Ouster (O) | Velodyne (V) | Omni-mIoU | Aeva (A) | Livox (L) | Solid-mIoU | |
| 001 | Ouster | O | O | 0.786 | **0.706** | 0.086 | 0.396 | 0.570 | 0.523 | 0.547 | 0.471 |
| 002 | Velodyne | V | V | 0.508 | 0.321 | **0.215** | 0.268 | 0.215 | 0.392 | 0.304 | 0.286 |
| 003 | Aeva | A | A | 0.813 | 0.686 | 0.313 | 0.500 | **0.776** | 0.655 | 0.716 | 0.608 |
| 004 | Livox | L | L | 0.853 | 0.560 | 0.148 | 0.354 | 0.615 | **0.690** | 0.653 | 0.503 |
| 005 | Omni | O+V | O+V | 0.675 | 0.736 | 0.225 | 0.481 | 0.621 | 0.578 | 0.600 | 0.540 |
| 006 | Solid | A+L | A+L | 0.844 | 0.737 | 0.288 | 0.513 | 0.814 | 0.736 | 0.775 | 0.644 |
| 007 | All | O+V+A+L | O+V+A+L | 0.831 | 0.776 | 0.386 | 0.581 | 0.795 | 0.723 | 0.759 | 0.670 |

TABLE II: Ablation study. The *Configuration Name* follows the naming convention from Table I. MLA denotes the teacher training method using multi-level adaptation. Our proposed framework (Case 010) shows the best performance on both the validation set and the test sets for individual sensors.

| #Case | Configuration Name | Method | mIoU (Val) | Omni-Directional LiDARs (Test) | | | Solid-State LiDARs (Test) | | | Test mIoU |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Ouster (O) | Velodyne (V) | Omni-mIoU | Aeva (A) | Livox (L) | Solid-mIoU | |
| 008 | Velodyne | MLA Teacher Generation | 0.644 | 0.683 | **0.499** | 0.591 | 0.716 | 0.638 | 0.677 | 0.634 |
| 009 | All | Random Init Weights + Sensor-Aware | 0.825 | 0.810 | 0.396 | 0.603 | 0.816 | 0.733 | 0.775 | 0.689 |
| 010 | All (ours) | **Weight Transfer + Sensor-Aware** | **0.831** | **0.799** | **0.486** | **0.643** | **0.824** | **0.757** | **0.791** | **0.717** |
| 011 | All | Weight Transfer + Attention | 0.815 | 0.789 | 0.245 | 0.517 | 0.778 | 0.733 | 0.756 | 0.636 |
| 012 | All | Weight Transfer + Average | 0.810 | 0.790 | 0.263 | 0.527 | 0.798 | 0.733 | 0.766 | 0.646 |

### C. Sensor-Aware Knowledge Distillation

With the expert teachers ($M_O^T$, $M_V^T$, $M_A^T$, and $M_L^T$) generated via the process described in Sec. III-B and the initialized student model ($M_P^{16}$), we performed the multi-teacher knowledge distillation. The student model was trained by randomly sampling sensor data from the entire training dataset, as shown in Fig. 2(c). At this stage, the student model receives both the input data and 'privileged information', indicating the sensor ($i$) from which the data originate.

A 'teacher activator ($T_A$)' uses this privileged information to select the optimal expert teacher ($T_A(i)$) for the current training step. The student model then underwent both supervised learning with the ground truth ($y_i$) and knowledge distillation to mimic the predictions of an activated expert teacher. The final loss function is defined as follows:

$$L_{total} = \lambda_1 L_{MOS}(y_i, \hat{y}_i) + \lambda_2 L_{KD}(T_A(i), \hat{y}_i) \quad \text{where } i \in \{V, O, A, L\} \tag{3}$$

Through this process, the final student model ($S_X$), which has progressively learned from all expert teachers, achieves a robust cross-sensor generalization performance without bias toward any particular sensor.

## IV. Experiments

### A. Experimental Setup

To validate the generalization capabilities of the framework across diverse sensor types, we utilized the HeLiMOS dataset collected using heterogeneous LiDAR sensors. Following established practices in MOS research, we adopted the IoU as our primary evaluation metric, defined as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{4}$$

where, in the context of MOS, a 'positive' corresponds to a 'moving', and TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

To further assess the generalization performance across different datasets, we evaluated our model on the Apollo-Southbay, Sipailou Campus, and SemanticKITTI datasets. Performance was measured directly on the respective validation and test sets without any additional training. Regarding the dataset composition, Apollo-Southbay consists of five test sequences with no separate validation sets. The Sipailou Campus dataset included one validation sequence (seq. 06) and two test sequences (seq. 00 and 07). The SemanticKITTI comprises a validation sequence (seq. 08). and 11 test sequences (seq. 11-21). For datasets with multiple sequences, the reported performance was the average IoU score for each sequence.

### B. Implementation Details

We selected 4DMOS, a representative non-projection-based model, as our baseline to demonstrate the framework's generalizability. All experiments were conducted using an NVIDIA RTX 4090 GPU. While our multi-stage distillation process increases the total training time by approximately five to six times compared to standard baseline training, the number of parameters in the final student model remains identical. While we adopted the general training hyperparameters from 4DMOS, the key parameters for our channel down operator $\Psi$ (Eq. 1) were configured specifically.

For the multi-level teacher generation, the number of target beams $M$ was sequentially set to $[64, 32, 16]$, with the corresponding number of source clusters $K$ set to $[128, 64, 32]$. We applied an intra-beam keep probability of $p = 0.5$; this

TABLE III: Zero-shot generalization performance on unseen datasets. The Apollo-Southbay dataset consists only of test sequences. * denotes the IoU score reported in the original paper when the model was *trained* with the 4DMOS architecture.

| Test Dataset | Test Sensor | IoU(Val) | IoU(Test) |
|---|---|---|---|
| Apollo-Southbay | Velodyne VLS-128 | – | 0.779 |
| Sipailou Campus | Avia Livox | 0.978 | 0.948 |
| SemanticKITTI | Velodyne HDL-64E | 0.432 (*0.719) | 0.569 (*0.652) |

probabilistic reduction in point density is analogous to the stepwise halving of the channels. For our loss functions, we set $\lambda_1 = 0.3$ for $L_{\text{MOS}}$, $\lambda_2 = 0.7$ for $L_{\text{KD}}$, and temperature parameter $\tau = 3$. These hyperparameters were chosen by analyzing the relative signal strength between native MOS logit and distillation loss, ensuring the $L_{\text{KD}}$ term provides a dominant guiding force to maximize the training effect.

### C. Sensor-Ignoring Setup

Before proposing the sensor generalization methodology, we conducted preliminary experiments to establish a performance baseline for individual and combined sensor configurations. Tab. I compares the mIoU performances of the models trained on each sensor were compared independently with those trained on specific combinations.

The individual sensor training results from Cases 001-004 show that while sensors O, A, and L achieved high mIoU scores, sensor V exhibited poor performance at 0.215, suggesting an inherent difficulty in learning from its data. Furthermore, although Cases 005 and 006, in which the sensors were grouped by scanning type, showed an overall performance improvement, the gain was not substantial.

This highlights the limitations of a naive data integration strategy based on simple physical characteristics. The model trained on all the sensors (Case 007) yielded the highest overall mIoU (Val) of 0.831. However, a per-sensor analysis revealed that although the performance of sensor V improved to 0.386, it remained markedly lower than that of the other sensors. Conversely, sensors O, A, and L outperformed their individually trained counterparts.

This demonstrates that a high aggregate score can mask the underperformance of a specific sensor. These results demonstrate that naive data integration induces a training bias, where the optimization of the model is skewed toward specific sensors (O, A, and L), while hindering the performance improvement for more challenging sensors (V). Therefore, addressing this issue to ensure robust performance across all sensors requires more advanced methodology capable of reflecting the unique characteristics of each sensor.

### D. Effectiveness of Sensor-Aware Teacher Generation

Before training the student model, we generated an expert teacher for each sensor, as described in Sec. III-B. The effectiveness of the proposed multilevel adaptive teacher generator for the most challenging sensor, V, is clearly demonstrated in Case 008 in Tab. II. Compared to Case 002 in Tab. I, where the V sensor was trained in isolation, the test mIoU for the V sensor surged more than two-fold, from 0.215 to 0.499.

This indicates that a progressive learning curriculum for sparse 16-channel data is highly effective. Interestingly, this V-sensor teacher model also achieved an enhanced performance on other sensors (O, A, and L), nearly matching the respective expert teachers (Case 001, Case 003, and Case 004). This can be analyzed as a positive side effect, where learning in a difficult (sparse) environment enhances the general feature extraction capabilities of the models, which are transferred s to other dense environments.

### E. Ablation Study of Core Components

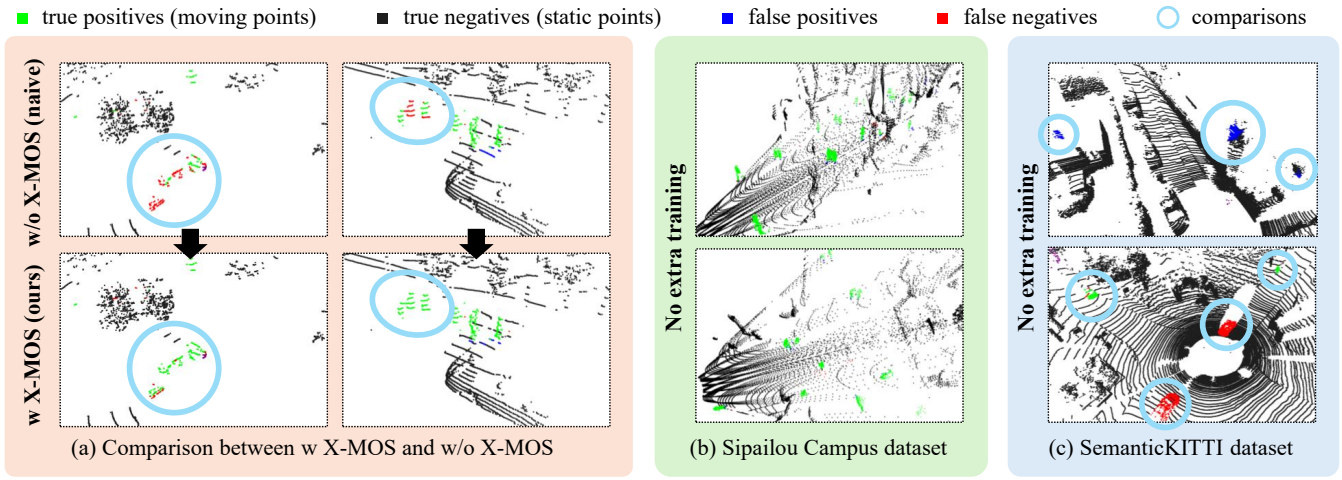
Next, we conducted an ablation study to verify the core components of student model training. To ascertain the importance of the pretrained weights ($M_P^{16}$), that is, the importance of the ready-to-learn model's weights shown in Fig. 2, we compared Case 010 (ours), which uses weight transfer, with Case 009, which was trained from randomly initialized weights (see Tab. II). The test mIoU increased from 0.689 to 0.717, indicating that the weight transfer was a significant contributor to the final performance.

This goes beyond providing a good starting point, which means that the student model begins with a strong prior on handling the most difficult sensor data distribution. This prior knowledge is a stable foundation that regularizes the entire training process, enabling students to absorb knowledge from other teachers more efficiently.

The result for Case 010 highlights the powerful synergy between weight transfer and sensor-aware knowledge distillation. This model achieved the highest performance among all the configurations, with a total test mIoU of 0.717. This score not only surpasses the naive 'all-in' training of Case 007 (0.670) but also exceeds the performance of the best individual expert teacher. The per-sensor test scores of O (0.799), V(0.486), A(0.824), and L(0.757) indicated that the students performed on par with or better than each expert teacher. This proves that X-MOS is a well-rounded generalization model that achieves high performance across all sensors without bias.

Finally, to validate the efficacy of the core component of our framework, the 'teacher activator ($T_A$)', we compared it with alternative teacher aggregation methods: an attention-based method (Case 011) and a teacher-logit averaging method (Case 012). Both alternatives performed significantly worse than the final model (0.717) and the randomly initialized model (0.689). This suggests that indiscriminate blending of knowledge from teachers with disparate specializations (e.g., sparse versus dense) provide conflicting and ambiguous learning signals that hinder student learning.

By contrast, our sensor-aware approach uses the sensor type as clear privileged information to activate the single most appropriate teacher at each step, ensuring that the student receives unambiguous guidance without interference. This ablation study clearly demonstrated that each proposed




(a) Comparison between w X-MOS and w/o X-MOS (b) Sipailou Campus dataset (c) SemanticKITTI dataset

Fig. 5: Qualitative comparison of results. (a) Comparison before and after applying the X-MOS framework. The cyan circles highlight the significant increase in true positives. (b) Excellent segmentation performance on the Sipailou Campus dataset without any fine-tuning. (c) Inference results on the SemanticKITTI dataset without fine-tuning, showing a considerable number of false positives and false negatives.

component is essential for achieving the final state-of-the-art (SOTA) performance.

### *F. Zero-Shot Generalization*

To validate the generalization capability of the proposed framework, we conducted zero-shot experiments by accurately evaluating the model on the validation and test sets of new unseen datasets without any additional training. The results are presented in Tab. III. The experiments utilized the Apollo-Southbay, Sipailou Campus, and SemanticKITTI datasets, each of which was composed of sensors with different specifications (e.g., beam (channel) count and scan pattern) and was collected in various environments and locations.

The experimental results show an excellent generalization performance on the Apollo-Southbay test sequences, achieving an IoU of 0.779 without any fine-tuning. Furthermore, the model demonstrated an overwhelmingly strong performance on the Sipailou Campus dataset, with an IoU of 0.978 for the validation set and 0.948 for the test set. This can be analyzed because the model effectively learns the features of the 128-channel omnidirectional sensor (Ouster) and the solid-state sensor (Livox) included in the HeLiMOS dataset during the sensor-aware knowledge distillation process. Consequently, it achieved a high generalization performance even on unseen datasets composed of similar sensors, specifically the 128-channel Velodyne sensor in Apollo-Southbay and the Livox sensor in the Sipailou Campus datasets.

By contrast, on the SemanticKITTI dataset, which is composed of data from a Velodyne HDL-64E sensor, the model achieved relatively modest IoUs of 0.432 and 0.569 for the validation and test sequences, respectively. The lower performance of this dataset was attributed to the model encountering an unfamiliar number of channels (64 beams) during inference, which was not experienced during the sensor-aware knowledge distillation stage, thus posing a challenge to generalization.

### *G. Qualitative Results*

Fig. 5 presents a qualitative analysis of the proposed framework. In Fig. 5(a), we compare the results obtained with (bottom row) and without (top row) the X-MOS framework. As highlighted by the cyan circles, our method successfully converted a significant number of FN into TP, which indicates an improved recall.

However, a limitation is observed in the vehicle example on the left: although the result is improved, the framework does not perfectly segment the entire object at the instance level. Figs. 5(b) and (c) show the zero-shot cross-sensor generalization performance of our model, which was trained solely on the HeLiMOS dataset and then evaluated on unseen datasets without any fine-tuning.

For the Sipailou Campus dataset, the model achieved a high IoU value of 0.978 (Tab. III). This strong quantitative result is visually supported by Fig. 5(b), which shows that most points were correctly classified as TP (green), with very few FN or FP. This demonstrates that the proposed model robustly generalizes to new scenes and sensor types.

Conversely, Fig. 5(c) shows the limitation in evaluating the SemanticKITTI dataset. Unlike in the previous case, there were noticeable numbers of FN and FP. We attributed this performance degradation to the use of a 64-channel LiDAR sensor, and the configuration of the model was not encountered during training, as mentioned in Sec. IV-F.

Furthermore, it should be noted that due to the inherent limitations of the 4DMOS architecture used as a baseline, there is a tendency for a certain level of FPs and FNs to occur on the SemanticKITTI dataset, even with direct training (see * mark in Tab. III). In summary, our qualitative analysis suggests that X-MOS operates robustly in novel environments with familiar sensor configurations. However, its performance is less stable when encountering sensors with characteristics such as the number of channels that were not represented in the training data.

## V. Conclusion

This study proposes X-MOS, a framework that addresses the generalization in heterogeneous LiDAR environments using sensor-specific expert teachers and sensor types as privileged information. Our framework mitigates the training bias of naive data integration and achieves a balanced generalization. The student model matches or exceeds the performance of individual expert teachers, particularly for challenging low-specification sensors. Although it generalizes well to similar environments, its performance remains modest for sensors with novel specifications.

Future work will integrate a few-shot learning [32] approach to facilitate rapid adaptation to unknown sensors with minimal data, while simultaneously conducting extensive experiments on diverse model architectures to further analyze the broad robustness of the proposed framework.

## References


[1] X. Chen, S. Li, B. Mersch, L. Wiesmann, J. Gall, J. Behley, and C. Stachniss, "Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6529–6536, 2021.

[2] B. Mersch, X. Chen, I. Vizzo, L. Nunes, J. Behley, and C. Stachniss, "Receding moving object segmentation in 3d lidar data using sparse 4d convolutions," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7503–7510, 2022.

[3] B. Mersch, T. Guadagnino, X. Chen, I. Vizzo, J. Behley, and C. Stachniss, "Building volumetric beliefs for dynamic environments exploiting map-based moving object segmentation," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 5180–5187, 2023.

[4] J. Kim, J. Woo, and S. Im, "Rvmos: Range-view moving object segmentation leveraged by semantic and motion features," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8044–8051, 2022.

[5] Z. Li, Y. Cui, J. Zhong, and Z. Fang, "Streammos: Streaming moving object segmentation with multi-view perception and dual-span memory," *IEEE Robotics and Automation Letters*, 2024.

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

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[8] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[9] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song, "L3-net: Towards learning based lidar localization for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6389–6398.

[10] B. Zhou, J. Xie, Y. Pan, J. Wu, and C. Lu, "Motionbev: Attention-aware online lidar moving object segmentation with bird's eye view based appearance and motion features," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8074–8081, 2023.

[11] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," *Advances in neural information processing systems*, vol. 31, 2018.

[12] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

[13] H. Lim, S. Jang, B. Mersch, J. Behley, H. Myung, and C. Stachniss, "Helimos: A dataset for moving object segmentation in 3d point clouds from heterogeneous lidar sensors," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 14 087–14 094.

[14] J. Cheng, K. Zeng, Z. Huang, X. Tang, J. Wu, C. Zhang, X. Chen, and R. Fan, "Mf-mos: A motion-focused model for moving object segmentation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 499–12 505.

[15] Y. Wu, Y. Wang, S. Zhang, and H. Ogai, "Deep 3d object detection networks using lidar data: A review," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1152–1171, 2020.

[16] D. P. Singh and M. Yadav, "Deep learning-based semantic segmentation of three-dimensional point cloud: a comprehensive review," *International Journal of Remote Sensing*, vol. 45, no. 2, pp. 532–586, 2024.

[17] M. U. Khan, S. A. A. Zaidi, A. Ishtiaq, S. U. R. Bukhari, S. Samer, and A. Farman, "A comparative survey of lidar-slam and lidar based sensor technologies," in *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. IEEE, 2021, pp. 1–8.

[18] Y. Wei, Z. Wei, Y. Rao, J. Li, J. Zhou, and J. Lu, "Lidar distillation: Bridging the beam-induced domain gap for 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 179–195.

[19] L. Yi, B. Gong, and T. Funkhouser, "Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 363–15 373.

[20] T. Guadagnino, B. Mersch, S. Gupta, I. Vizzo, G. Grisetti, and C. Stachniss, "Kiss-slam: A simple, robust, and accurate 3d lidar slam system with enhanced generalization capabilities," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 5363–5370.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] M. Li, Y. Zhang, Y. Xie, Z. Gao, C. Li, Z. Zhang, and Y. Qu, "Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3829–3837.

[23] J. Li, M. Lu, J. Liu, Y. Guo, Y. Du, L. Du, and S. Zhang, "Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2489–2498, 2023.

[24] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, "Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8637–8646.

[25] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-voxel knowledge distillation for lidar semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8479–8488.

[26] Z. Li, H. Liang, H. Wang, M. Zhao, J. Wang, and X. Zheng, "Mkd-cooper: Cooperative 3d object detection for autonomous driving via multi-teacher knowledge distillation," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1490–1500, 2023.

[27] A. Amirkhani, A. Khosravian, M. Masih-Tehrani, and H. Kashiani, "Robust semantic segmentation with multi-teacher knowledge distillation," *IEEE Access*, vol. 9, pp. 119 049–119 066, 2021.

[28] M. H. Aslam, M. O. Zeeshan, M. Pedersoli, A. L. Koerich, S. Bacon, and E. Granger, "Privileged knowledge distillation for dimensional emotion recognition in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3338–3347.

[29] Q. Li, W. Xia, L. Yin, J. Jin, and Y. Yu, "Privileged knowledge state distillation for reinforcement learning-based educational path recommendation," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 1621–1630.

[30] P. Zhao, L. Xie, J. Wang, Y. Zhang, and Q. Tian, "Progressive privileged knowledge distillation for online action detection," *Pattern Recognition*, vol. 129, p. 108741, 2022.

[31] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020.

[32] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.