Continual Learning and Unknown Object Discovery in 3D Scenes via Self-Distillation

Mohamed El Amine Boudjoghra[1], Jean Lahoud[1], Hisham Cholakkal[1], Rao Muhammad Anwer[1,2], Salman Khan[1,3], and Fahad Shahbaz Khan[1,4]

[1] Mohamed bin Zayed University of Artificial Intelligence
[2] Aalto University
[3] Australian National University
[4] Linköping University

Abstract. Open-world 3D instance segmentation is a recently introduced problem with diverse applications, notably in continually learning embodied agents. This task involves segmenting unknown instances and learning new instances when their labels are introduced. However, prior research in the open-world domain has traditionally addressed the two sub-problems, namely continual learning and unknown object identification, separately. This approach has resulted in limited performance on unknown instances and cannot effectively mitigate catastrophic forgetting. Additionally, these methods bypass the utilization of the information stored in the previous version of the continual learning model, instead relying on a dedicated memory to store historical data samples, which inevitably leads to an expansion of the memory budget. In this paper, we argue that continual learning and unknown object identification are desired to be tackled in conjunction. To this end, we propose a new exemplar-free approach for 3D continual learning and unknown object discovery through continual self-distillation. Our approach, named OpenDistill3D, leverages the pseudo-labels generated by the model from the preceding task to improve the unknown predictions during training while simultaneously mitigating catastrophic forgetting. By integrating these pseudo-labels into the continual learning process, we achieve enhanced performance in handling unknown objects. We validate the efficacy of the proposed approach via comprehensive experiments on various splits of the ScanNet200 dataset, showcasing superior performance in continual learning and unknown object retrieval compared to the state-of-the-art. Code and model are available at github.com/aminebdj/OpenDistill3D.

1 Introduction

While common recognition and detection methods rely on a fixed set of object labels, this setting does not properly represent real-world scenarios. A better representation is achieved in the open-world setting [2,8,11,20,27,29], where object labels are progressively presented to a given model. In the open-world setting, the target is twofold: (1) The model is required to recognize both known and

unknown objects, and (2) the model should maintain knowledge of previously known objects when a new set of labels is available. This open-world formulation has been recently explored for 3D instance segmentation [3], which predicts masks for known and unknown 3D object instances, along with their labels.

The open-world performance is measured by the ability of a model to retain knowledge of old classes and effectively recognize unknown objects, which necessitates a holistic solution that tackles this interconnected nature.

Moreover, the difference between the quality of known labels and pseudo-labels of unknown objects still poses a challenge in the open-world. This requires additional measures during training to better distinguish between known and unknown objects.

A common approach to open-world 3D instance segmentation [3] maintains the ability to recognize previously encountered classes by finetuning the model with exemplars. The exemplars are saved samples of previously labeled data, allowing the model to have access to the previous set of labels. Nevertheless, exemplars do not fully represent the previous data, which results in models being prone to forgetting the knowledge related to the old classes after training on new labels. Moreover, since recognizing unknown objects requires a good understanding of objectness and a good differentiation between known and unknown object classes, a degraded model performance in recognizing old classes affects its ability to recognize unknown objects. We show a sample scene in Figure 2, where 3D-OWIS [3] fails to predict previous classes and unknown objects.

In this work, we propose an open-world 3D instance segmentation method that employs self-distillation to preserve the knowledge of both previously known and unknown objects, as illustrated in Figure 2. Instead of relying on saved exemplars, our model exploits previously learned representation for known and unknown object recognition. Moreover, we introduce a specialized loss function designed to manage data uncertainty in the open world and leverage the objectness priors of pseudo-labels to achieve optimal open-world performance for both known and unknown objects.

Our key contributions are threefold:

- We propose a novel distillation approach for joint 3D continual learning and discovery of unknown objects.
- We introduce a loss function that emphasizes learning from high-quality unknown pseudo-labels.
- When compared to previous open-world 3D instance segmentation methods, our experiments show that our proposed method results in a major improvement in segmenting both known and unknown objects. Our approach achieves an absolute gain of as high as 25.5% in terms of unknown recall while also enhancing the performance on previous classes by up to 12.1% in terms of mean average precision (mAP), compared to the best reported results in literature [3].

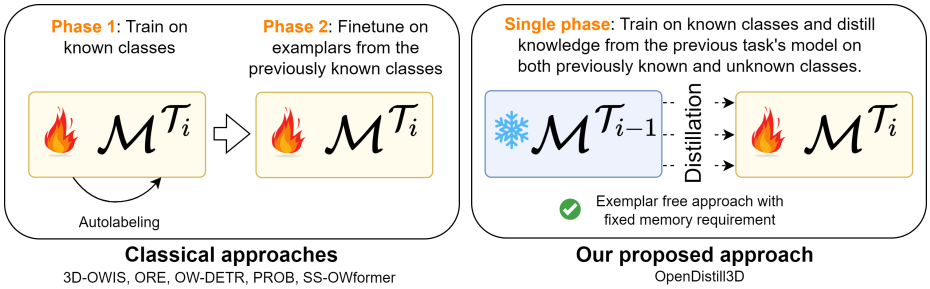

Fig. 1: (**Left**) The depicted pipeline illustrates the classical approaches employed in the open-world task 3D-OWIS [3], ORE [11], OW-DETR [8], PROB [29], and SS-OWformer [16]. These methods entail a dual-phase process: the initial phase focuses on learning new classes introduced in the current Incremental Task while simultaneously learning new objects via Auto-Labeling. Subsequently, the second phase aims to alleviate catastrophic forgetting by using exemplar replay. (**Right**) The illustrated pipeline showcases our novel exemplar free approach for open-world tasks, addressing both catastrophic forgetting and the identification of unknown objects through self-distillation from the model of the preceding task.

2 Related Work

2.1 3D instance segmentation

The challenge of instance segmentation in 3D scenes has been explored in various studies, with different approaches to address it. Some methods [9, 23] opt for a top-down strategy, generating proposals from one stream and combining them with local features from a parallel stream to predict final masks for each instance. While others [10,17,25,28] follow a bottom-up approach which involves predicting instances through semantic segmentation followed by instance prediction in the same stream. Unlike other approaches that adopt either a top-down or bottom-up strategy, SoftGroup [25] introduces a two-stage architecture, which combines both strategies. The state-of-the-art 3D instance segmentation model Mask3D [21] uses a hybrid CNN-Transformer architecture to generate a class-mask prediction for all instances in the scene, generated from refined queries. Another work [1] builds on top of Mask3D [21] and shows the potential of spatial supervision in improving the performance of 3D instance segmentation models.

We argue that all previously proposed methods exhibit poor performance when it comes to distinguishing between background and known classes. For instance, a trained Mask3D [21] pipeline generates class agnostic masks for all objects that share some similarity with one of the known classes, which results in many false positive predictions from the background. In our paper, we demonstrate that the model's performance on known instances can be significantly improved through self-distillation. This involves leveraging the knowledge of unknown objects from a pre-trained teacher model with an identical architecture,

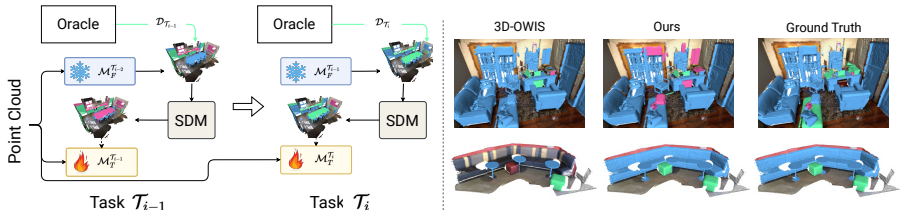

Fig. 2: Proposed method for open-world 3D instance segmentation. (**Left**) We propose a novel Self-Distillation Module (SDM), which takes the prediction of the teacher model $\mathcal{M}_F^{\mathcal{T}_{i-1}}$ from the previous task $\mathcal{T}_{i-1}$, to generate high-quality pseudo-labels and distill the knowledge on the previously known (■) and unknown (■) classes to the student model $\mathcal{M}_T^{\mathcal{T}_i}$ being trained in the current Task $\mathcal{T}_i$, the instances in green (■) represent the currently known classes. (**Right**) We show a qualitative comparison between our method and 3D-OWIS [3] Task $\mathcal{T}_2$. Our self-distillation approach helps the model better recognize unknowns (■) as shown in the scene on top, and also better preserve the knowledge of the previously seen classes (■) as demonstrated in the scene at the bottom.

to supervise negative classes from the background (unknowns), in addition to the known classes.

2.2 Continual learning

Continual learning or lifelong learning, involves training models in phases that include various subsets of the label space. Recent approaches to incremental learning fall into two primary categories: Knowledge Distillation (KD) aims to preserve knowledge from a previous model version. Some methods rely on logit alignment, namely [12], while others distill feature-level information [7]. Exemplar Replay (ER) is another method that is widely used in continual learning [8, 11, 13, 26], where a reservoir of samples from earlier training rounds is stored and replayed in subsequent phases to recall past knowledge.

In the realm of incremental object detection (IOD), the application of incremental learning is particularly challenging. Unlike incremental image classification, IOD deals with images containing multiple objects, including both old and new types, with only the new types annotated in any given training phase. KD and ER have been previously applied to object detection, with [22] using KD on the output of Faster R-CNN [18]. Alternatively, ORE [11] suggests storing a set of exemplars and fine-tuning the model on these exemplars after each incremental step. A recent method [14] makes use of both KD and ER to further improve the performance. In contrast to 2D continual learning, 3D continual learning remains mostly unexplored. A recent method for 3D instance segmentation [3] uses a bank of exemplars from the previous iteration to fine-tune the model in the current iteration, to retain the knowledge of the previously seen classes. Storing exemplars requires additional memory, thus leading to a limited number of exemplars that can be stored per class. Fine-tuning a model with a

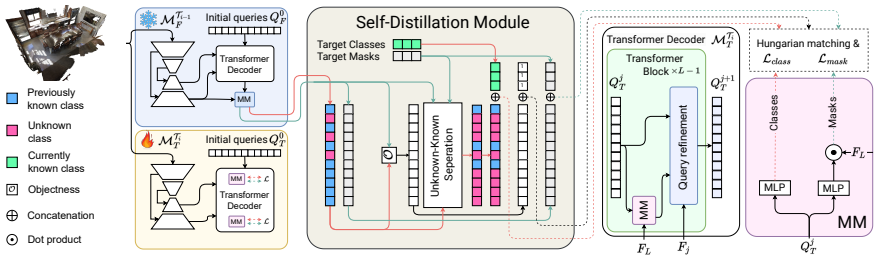

Fig. 3: Proposed known-unknown self-distillation pipeline. From left to right, the input point cloud passes through the frozen model $\mathcal{M}_F^{\mathcal{T}_{i-1}}$ to generate mask (■) and class (■) proposals with the same number of initial queries (one query generates one proposal). These proposals are then used to estimate an objectness score following an objectness estimation function $\mathcal{O}$ which outputs a score for every proposal ranging from 0 to 1. The Known-Unknown Separation component takes the generated proposals to correct the proposals that are likely to belong to the unknown objects (■) and misclassified as one of the previously known classes (■). Updated proposals are then concatenated with the labels of the classes known in the current task $\mathcal{T}_i$ (■), to output the final targets that are used to supervise the mask-class prediction heads in the Mask Module (MM) in all levels of the Transformer Decoder of the student model $\mathcal{M}_T^{\mathcal{T}_i}$.

limited number of exemplars results in a lower performance on the previously known classes. As an alternative, we propose to use the saved model from the previous task to generate weak labels for the previously known classes. This requires a fixed amount of storage for the model, instead of an increasing memory requirement at every iteration for exemplars for the previously seen classes.

2.3 Open-world tasks

In the domain of open-world object recognition, the concept was initially introduced in [2]. Shifting gears to open-world object detection, numerous studies [8, 11, 29] have delved into this field. In ORE [11], a strategy involving the generation of pseudo-labels for unknowns was implemented, enabling contrastive clustering during training to enhance the separation between unknown and known classes. To address incremental learning, exemplar replay was adopted to prevent catastrophic forgetting of old classes.

Taking a similar approach to ORE [11], OW-DETR [8] employs a transformer-based model for open-world object detection and introduces an alternative method for generating pseudo-labels for unknowns. This method involves a novel objectness estimation, incorporating a foreground objectness branch to distinguish between background and foreground. In the realm of outdoor 3D Lidar point cloud semantic segmentation, [4] puts forward a model predicting old, novel, and unknown objects through three separate classification heads. In Open-World 3D instance segmentation, 3D-OWIS [3] proposes an architecture that addresses the problem of unknown objects instances segmentation through internal unknown

object pseudo-labels generation, where the model uses the generated class agnostic proposals to provide unknown pseudo-labels, while the continual learning problem is addressed via exemplar replay with fine-tuning on a small set of previously seen classes, similar to ORE [11].

A critical limitation observed in preceding studies on the open-world task is their tendency to exclusively address one of the two sub-tasks associated with the problem. Previous methods train the model in a specific task while using the same model for pseudo-label generation; due to catastrophic forgetting, the quality of pseudo-labels is very low at the early stages of training. Consequently, this approach restricts the model's capacity to learn distinctive features for the unknown object as it forgets the previously seen classes in the process of learning novel ones.

3 Background

3.1 3D instance segmentation

A typical approach to 3D instance segmentation uses a sparse 3D convolutional network encoder and a transformer decoder [21]. We employ a similar pipeline, where the input point cloud is first voxelized, then voxel features are extracted with a 3D U-Net-shaped CNN. A transformer decoder takes as input a set of queries and refines them through a series of self-attention and cross-attention with the voxel features at multiple scales. Each refined query is then used to generate an object binary mask and a corresponding semantic label.

We define the set of the final refined queries from the transformer decoder as $\mathcal{Q} = \{q_i \in \mathbb{R}^d \mid i \in (1, 2, ..., n_{\mathcal{Q}})\}$ where d is the dimension of one single query, and $n_{\mathcal{Q}}$ is the total number of queries. A query q is used to generate a mask f_{mask} and a class prediction f_{class} after remapping it using an MLP. Let $f_{mask}(q) = \text{Sigmoid}\left(\text{MLP}_{mask}(q) \cdot F_L\right)$ and $f_{class}(q) = \text{Softmax}\left(\text{MLP}_{class}(q)\right)$, where N is the number of input voxels, $F_L \in \mathbb{R}^{d \times N}$ are the per-voxel features extracted from the high-resolution layer of the CNN backbone, $\text{MLP}_{mask} : \mathbb{R}^d \mapsto \mathbb{R}^{d_F}$ is a Multi-Layer Perceptron that maps the query to the dimension of the feature map, and $\text{MLP}_{class} : \mathbb{R}^d \mapsto \mathbb{R}^{C+1}$ is a Multi-Layer Perceptron that maps from the query's dimension to the total number of classes (knowns + background).

We define an arbitrary objectness function $\mathcal{O} : (m \in \mathbb{R}^N, p \in [0, 1]) \mapsto [0, 1]$, that maps a certain heatmap prediction $m = f_{mask}(q)$ and a class probability $p = \max(f_{class}(q))$ to an objectness score.

3.2 Open-world 3D instance segmentation

Unlike closed-set 3D instance segmentation, open-world 3D instance segmentation predicts a class *unknown* for unlabeled classes as well, and incrementally learns new classes when accessing new label, while preserving the knowledge of the previously seen classes without using all of their labels.

One of the major limitations in previous works in the task of open-world object detection/segmentation is the uni-tasking of the open-world problems.

For instance, OW-DETR [8], ORE [11], PROB [29], SS-OWformer [16], and 3D-OWIS [3] target the sub-task of unknown identification without relying on the continual learning sub-task. In our work, we construct a framework that jointly solves the two sub-tasks of the open-world, continual learning and unknown identification, which performs remarkably better than the approaches in the literature. We also propose a novel reduction technique for cross-entropy using the objectness of samples as weights, to achieve an improved known-unknown performance trade-off. In this work, we exploit the weights of the model from the previous task, without any exemplars, to achieve competitive performance in alleviating catastrophic forgetting and improving unknown object instance segmentation.

4 Methodology

Given a 3D point cloud, with each point represented by a 3D coordinate and RGB colors, the target is to segment the point cloud into K binary masks along with their corresponding semantic labels. Our proposed method starts by adapting a closed-set 3D instance segmentation method into the open-world setting. We then propose a self-distillation approach to improve the open-world performance. Additionally, we propose a cross-entropy loss tailored for the open-world task. Our pipeline is shown in Figure 3.

4.1 Self-Distillation Module (SDM)

The model trained in task $\mathcal{T}_{i-1}$ is capable of predicting pseudo-labels for both the previously known and unknown objects. However, with the introduction of new labels in the current task, $\mathcal{T}_i$, the unknowns change, resulting in an overlap between the previous task's model predictions for unknowns with the current task's known classes. This problem is addressed in the literature by selecting class-agnostic mask predictions with low intersection over union with the class-agnostic mask of currently known instances. However, we conjecture that these techniques exhibit sub-optimal performance on unknowns in intermediate tasks due to the inadequate distinction between unknowns and previously known classes.

Sperating previously known from the unknown objects: Since 3D-OWIS relies on IoU between ground truth classes and predictions to generate pseudo labels, it results in wrong proposals of previously known classes to be labeled as unknown since the labels of the previously known classes are not available in the training set at a given incremental task. We propose an unknown-known pseudo-labels separation technique that uses overlap between predictions of the teacher model and confidence for separation (more detail in Suppl. Material).

Our adopted instance segmentation architecture Mask3D [21] refines a set of queries in the transformer block to predict instance masks and classes. The frozen model from the previous task $\mathcal{M}_F^{\mathcal{T}_{i-1}}$ refines $n_{\mathcal{Q}_F}$ to generate $n_{\mathcal{Q}_F}$ labels

and masks for objects in the input scene, where the final refined queries from the frozen teacher model are denoted $\mathcal{Q}_F = \mathcal{Q}_F^L$ (see Figure 3) and L is the number of transformer blocks. Even though some of these predictions are for unknown objects, additional pseudo-labels are required to achieve better performance on the unknowns when training the model $\mathcal{M}_T^{\mathcal{T}_i}$ in task $\mathcal{T}_i$. We also highlight that these $n_{\mathcal{Q}_F}$ predictions contain many unknown objects that were misclassified as one of the previously known classes and require a separation technique to effectively detect and differentiate them from the real previously known classes. The Unknown-Known Separation (UKS) component takes the predictions from the model from the previous Task $\mathcal{T}_{i-1}$, and target masks of the known classes in Task $\mathcal{T}_i$ to generate better pseudo-labels for the previously known classes and unknown objects. The pseudo-labels for the unknown, denoted $\mathcal{U}$, are selected as all predictions from $\mathcal{M}_F^{\mathcal{T}_{i-1}}$ that do not overlap with the currently known instances and with the highly confident pseudo-labels for the previously known classes from $\mathcal{M}_F^{\mathcal{T}_{i-1}}$, while the pseudo-labels for the previously known classes, denoted $\mathcal{P}$, are all the remaining predictions from the model $\mathcal{M}_F^{\mathcal{T}_{i-1}}$ from the previous Task $\mathcal{T}_{i-1}$ (more detail in Suppl. Material).

Training targets: We define the targets $\tilde{T}_{\mathcal{T}_i} = (\tilde{M}, \tilde{Y}, \tilde{S})$ that are used to train $\mathcal{M}_T^{\mathcal{T}_i}$ as

$$\tilde{T}_{\mathcal{T}_i} = (\{m\}, \{y\}, \{s\} \;\mid\; \forall (m, y, s) \in \mathcal{P} \oplus \mathcal{U} \oplus \mathrm{GT}_{\mathcal{T}_i})$$

Where $\tilde{M}$ is the set of masks, $\tilde{Y}$ is the set of labels, and $\tilde{S}$ is the set of objectness scores predicted by the teacher model. $\mathrm{GT}_{\mathcal{T}_i}$ are the ground truth labels for the currently known classes from the incremental Task $\mathcal{T}_i$.

4.2 Open world cross-entropy loss

Existing methods designed for open-world tasks, such as [3,8,11], involve training models with pseudo-labels for unknown objects, assuming an equivalence in quality between these pseudo-labels and ground-truth labels. We contend that this simplistic approach is ill-suited for the complexities of an open-world setting, given that the per-class count of pseudo-labels for the previously seen classes is higher than that of the unknown objects, and some pseudo-labels are of lower quality (lower objectness score). In response to the challenges posed by the open-world scenario, we introduce a novel loss function that prioritizes pseudo-labels characterized by high confidence and low frequency over those with low confidence and high frequency.

Weight adjustment with score prior We start by defining the final refined queries $\mathcal{Q}_T = \mathcal{Q}_T^L$ generated by the last transformer block $L-1$ of the trainable student model (see Figure 3). Given a set of target pseudo-labels $\tilde{T}_{\mathcal{T}_i} = (\tilde{M}, \tilde{Y}, \tilde{S})$ generated by the Self-Distillation Module 4.1. We define the class predictions from the set of queries $\mathcal{Q}_T$, extracted from the trainable model $\mathcal{M}_T^{\mathcal{T}_{i+1}}$ as $\hat{Y}_T = \{y = f_{class}(q) \mid \forall q \in \mathcal{Q}_T\}$, and the mask predictions as $\hat{M}_T =$

$\{m = f_{mask}(q) \mid \forall q \in \mathcal{Q}_T\}$. Afterwards we perform Hungarian matching between $(\hat{M}_T, \hat{Y}_T)$ and $(\tilde{M}_T, \tilde{Y}_T)$ to obtain two permutation matrices π_p and π_t to match prediction pairs to target pairs, where t stands for target and p for prediction. After the matching, we obtain a triplet of label-prediction-score sets denoted $(\overline{Y}_t^{\mathcal{T}_i}, \overline{Y}_p^{\mathcal{T}_{i+1}}, \overline{S}_t^{\mathcal{T}_i})$, where $\overline{Y}_t^{\mathcal{T}_i} = \tilde{Y}[\pi_t]$, $\overline{Y}_p^{\mathcal{T}_{i+1}} = \hat{Y}_T[\pi_p]$, and $\overline{S}_t^{\mathcal{T}_i} = \tilde{S}[\pi_t]$. Finally, we train the model $\mathcal{M}_T^{\mathcal{T}_{i+1}}$ for the task $\mathcal{T}_{i+1}$ by optimizing the following objective loss function.

$$\begin{cases} \mathcal{L}_{owce} = \dfrac{1}{\eta} \displaystyle\sum_{(y_p, y_t, s_t) \in (\overline{Y}_p^{\mathcal{T}_{i+1}}, \overline{Y}_t^{\mathcal{T}_i}, \overline{S}_t^{\mathcal{T}_i})} W_I(s_t, y_t) \cdot l(y_p, y_t) \\ \\ W_I(s_t, y_t) = \begin{cases} 1 - 0.5e^{-\alpha(s_t - 0.5)}, & s_t > 0.5, \;\; y_t = \mathbf{0} \\ s_t, & s_t > 0.5, \;\; y_t \neq \mathbf{0} \\ 0, & \text{Otherwise} \end{cases} \end{cases}$$

where $l(y_p, y_t)$ is the cross-entropy loss between the prediction and the label, $\mathbf{0}$ is the label for the unknown object, and α is set to 15 for all experiments. η is a normalization factor defined as follows

$$\eta = \sum_{(y_t, s_t) \in (\overline{Y}_t^{\mathcal{T}_i}, \overline{S}_t^{\mathcal{T}_i})} W_I(s_t, y_t)$$

Due to the small number of pseudo-labels for the unknown object, compared to the ones from the previously seen classes, we upscale the weight for the unknown pseudo-labels before computing the loss. This showed improvement in the performance of the known classes, as the model can better differentiate them from the unknowns.

5 Experiments

We evaluate our method on the ScanNet200 dataset [5, 19] which contains reconstructed point clouds of indoor scenes with labels for 200 object classes. We experiment with the three open-world splits proposed in [3]. For the evaluation metrics, we report the mean average precision (mAP) for previously known and current classes, unknown recall (U-Recall), wilderness impact (WI) [6], and absolute open set error (A-OSE) [15]. Furthermore, to test the generalizability of our model to new geometries, we train it on ScanNet200 and use it with OpenMask3D [24] as a 3D proposal network and report the mAP on the Replica Dataset compared to the original proposal network Mask3D.

5.1 Implementation details

Open-World training: In Task $\mathcal{T}_1$, we initialize the teacher model $\mathcal{M}_F^{\mathcal{T}_0}$ with a closed-set model trained for 300 epochs on the known classes only.

Table 1: State-of-the-art comparison. We show the performance of our method compared to 3D-OWIS. Our model outperforms 3D-OWIS in all cases on the previously known classes and unknown classes as reflected by the higher mAP and U-Recall.

Task IDs ($\rightarrow$)	**Task** $\mathcal{T}_1$					**Task** $\mathcal{T}_2$						**Task** $\mathcal{T}_3$		
	WI	A-OSE	U-Recall	mAP (↑)		WI	A-OSE	U-Recall	mAP (↑)			mAP (↑)		
	(↓)	(↓)	(↑)	Current known	All	(↓)	(↓)	(↑)	Previously known	Current known	All	Previously known	Current known	All
Split A														
Upper Bound	0.057	290	62.82	39.51	39.38	0.000	60	42.51	39.02	20.91	31.53	29.31	17.69	25.88
Mask3D [21]	-	-	-	39.12	39.12	-	-	-	38.30	20.57	29.15	28.61	18.33	25.58
3D-OWIS [3]	**0.397**	607	34.75	**40.20**	**39.70**	**0.007**	**126**	27.03	29.40	**16.40**	22.70	20.20	**15.20**	18.70
OURS	0.841	**564**	**41.27**	39.29	38.81	0.352	199	**48.82**	**36.21**	15.61	**25.65**	**22.84**	14.91	**20.50**
			+6.52					+21.79	+6.81			+2.64		
Split B														
Upper Bound	1.000	791	72.08	29.00	29.19	0.400	400	73.16	24.60	23.86	24.48	25.44	29.19	26.83
Mask3D [21]	-	-	-	23.48	23.48	-	-	-	21.81	18.91	20.37	24.20	29.22	26.06
3D-OWIS [3]	3.684	1780	24.79	23.60	23.30	**0.755**	**581**	24.21	18.70	17.30	17.90	18.70	24.60	20.90
OURS	**0.873**	**1243**	**30.11**	**26.03**	**25.71**	0.781	758	**33.78**	**24.64**	**17.85**	**21.53**	**19.89**	**26.00**	**22.16**
			+5.32					+9.57	+5.94			+1.19		
Split C														
Upper Bound	0.583	853	70.80	25.90	26.10	0.500	469	71.61	25.20	26.45	25.91	28.01	27.37	27.80
Mask3D [21]	-	-	-	20.82	21.15	-	-	-	22.67	26.67	24.13	25.41	25.21	25.35
3D-OWIS [3]	**0.419**	**1294**	14.34	18.00	17.60	**0.152**	**303**	15.80	13.90	22.20	17.80	17.80	17.70	17.80
OURS	0.876	1403	**29.31**	**25.22**	**24.90**	1.371	849	**41.31**	**26.02**	**22.66**	**24.24**	**23.14**	**20.49**	**22.28**
			+14.97					+25.51	+12.12			+5.34		

Table 2: Per class results on Replica dataset. We report classes with zero $mAP50$ by one of the models. We show that OpenMask3D with our model generalizes to new classes. We trained OpenDistill3D on ScanNet200 labels for the same number of epochs as Mask3D, in addition to pseudo labels for unknown classes generated by the SDM. As a teacher model, we used a pre-trained Mask3D on ScanNet200.These results demonstrate that training models with unknown-known self-distillation can enhance the ability of OpenVocabulary models to adapt to new geometries.

class	clock	tissue-paper	tablet	basket	blanket	pillar	sculpture	Mean
OpenMask3d (w/ **Mask3D**)	0.00	0.00	0.00	0.00	0.00	**15.0**	**27.3**	15.90
OpenMask3D (w/ **OpenDistill3D**)	**31.2**	**27.2**	**26.4**	**25.9**	**07.5**	00.0	0.00	**18.40**

The student model $\mathcal{M}_T^{\mathcal{T}_1}$ is trained for a total number of 600 epochs, on the labels for the known classes and the pseudo-labels generated by the self-distillation module. In the subsequent task, the teacher and the student models are initialized with thebest checkpoint from the model of the previous Task $\mathcal{M}_T^{\mathcal{T}_1}$. We further train the student model for 400 epochs on novel classes from Task $\mathcal{T}_2$, and pseudo-labels for the previously known and unknown objects, generated from the self-distillation module. The training scheme continues similarly to Task $\mathcal{T}_2$ for all subsequent Tasks.

Table 3: Generalizability test in an Open-Vocabulary setting of OpenMask3D on Replica dataset with three proposal methods

Training classes	Proposal method	mAP (%)↑	$mAP50$ (%)↑	$mAP25$ (%)↑
Scannet200	Mask3D	09.50	15.90	22.90
	3D-OWIS	09.70	15.60	**24.10**
	Ours	**11.00**	**18.40**	23.30

Table 4: Ablation results. We show in the table below the effect of progressively adding our contributions. In task $\mathcal{T}_1$, we use the same checkpoint for the model with and without Unknown-Known Separation (UKS) in rows one and two, since there are no previously known classes. In Task $\mathcal{T}_2$ we show that UKS improves over the naive distillation on both unknowns and knowns. In the third row, we showcase the effectiveness of using W_I in achieving a good balance in performance between the knowns and unknowns reflected by the higher mAP.

Split C													
Row ID (↓)	**Task IDs** (→)		**Task $\mathcal{T}_1$**					**Task $\mathcal{T}_2$**					
			WI	A-OSE	U-Recall	mAP (↑)		WI	A-OSE	U-Recall	mAP (↑)		
	UKS	W_I	(↓)	(↓)	(↑)	Current known	All	(↓)	(↓)	(↑)	Previously known	Current known	All
1	×	×	1.015	1391	29	23.12	22.83	1.853	778	36.57	22.47	21.46	21.83
2	✓	×	1.015	1391	29	23.12	22.83	2.211	875	39.53	24.04	22.26	23.30
3	✓	✓	0.876	1403	29.31	25.22	24.90	1.371	849	41.31	26.02	22.66	24.24

Novel geometry aware class agnostic model for open vocabulary task We initialize the teacher model with a pre-trained model on the ScanNet200 dataset for 600 epochs. The student model is randomly initialized and trained for 600 epochs by optimizing the class loss 4.2 and the mask loss of [21]. The training targets are the set of labels and pseudo-labels generated by SDM. The trained model is used as class agnostic mask proposal network for Open-Vocabulary 3D instance segmentation models.

5.2 Results

Open-world results: We show in Table 1 that using self-distillation from a teacher model improves the performance of the student model on the unknown objects, compared to training with pseudo-labels of the unknowns generated through autolabeling by the student model itself in a two-stage approach. Our findings illustrated through Figure 4 indicate a significant decline in pseudo-label generation when employing a two-stage solution like 3D-OWIS, which stems from the reliance on the model's inherent sense of objectness for generating proposals of unknown objects. As the model rapidly forgets previously known classes (As illustrated in the highlighted area in Figure 4), its capacity to generate proposals diminishes accordingly. The boost in open-world performance in our approach is a result of high-quality and consistent unknown pseudo-labels generated by the

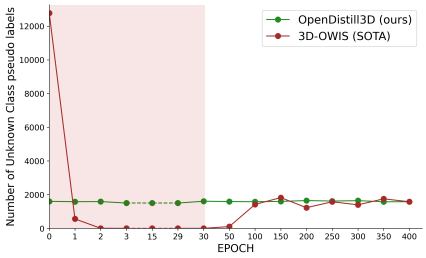

Fig. 4: Effect of catastrophic forgetting on unknown pseudo labels generation. The graph above illustrates the number of pseudo labels with a confidence threshold exceeding 0.8, generated via two distinct methods: self-distillation (OpenDistill) and two-phase training (3D-OWIS). We highlight the catastrophic forgetting area in red.

student model from the early stages of the model until the end of the training, as shown in Figure 4. Furthermore, since our self-distillation method relies on a pre-trained frozen model from the previous tasks, it generates a relatively higher number of pseudo-labels compared to 3D-OWIS, which occasionally results in performance degradation in terms of WI and A-OSE as shown in Table 1.

Incremental learning results: Our exemplar-free self-distillation method shows better performance on the previously known classes for all tasks compared to 3D-OWIS, as shown in Table 1. This performance improvement is due to the high number of pseudo-labels generated by the teacher model from the preceding Task. Which preserves a similar performance to the teacher model on the previously seen classes. Exemplar replay relies on a small number of samples per class for fine-tuning the model, once trained on the novel classes, this restriction on the number of exemplars results in a lower performance than our self-distillation method, which uses the previous model as a teacher to generate several pseudo-labels larger than the stored number of exemplars by 3D-OWIS.

Table 5: Similarity between knowns and unknowns in the CLIP space.

Top k unknowns	Split A	Split B	Split C
1	36.63	22.68	35.24
2	54.69	31.51	43.94
3	64.06	45.21	50.00

Closed-set results: In Table 1, our results highlight that training a model on unknowns (Upper Bound of our model), in comparison to Mask3D, which is trained exclusively on presently known classes, enhances the model's performance on the known classes. Due to the potential high similarity among known objects, the model may mistakenly predict various unknowns as knowns, leading to a decline in performance on the known classes. To give clear empirical proof of how some known classes might be highly similar to some of the unknown objects, we provide a measure of the similarity of the class names in the CLIP embedding spaces (see Table 5), where we provide Top k unknowns that their sum of similarity is higher than 0.9 e.g. a chair (known) is considered similar to armchair (unknown) and folded chair (unknown) for top 2 unknowns (details in Suppl. Material). We show that closed-set model performance might be improved with unknown object self-distillation.

Table 6: Closed-set 3D instance segmentation results. We show the performance of the models used as a class-agnostic mask proposal in OpenMask3D in a closed setting.

	mAP(↑)	mAP_{50}(↑)	mAP_{25} (↑)
CSC	-	25.24	-
LGround	-	26.09	-
Mask3D(w/o DBSCAN)	25.73	34.09	38.87
Ours(w/o DBSCAN)	**27.37**	**35.98**	**41.44**
Mask3D(w/ DBSCAN)	27.40	37.00	42.30
Ours(w/ DBSCAN)	**28.00**	**37.20**	**43.40**

The results demonstrate that self-distillation of unknowns can improve the model's performance on known classes, as evident in Tables 6 and 1. In Table 1, our Upper Bound model, trained on unknowns in a closed-set, outperforms Mask3D, which is trained solely on presently known classes, in Task $\mathcal{T}_1$ on currently known classes. Additionally, Table 6 illustrates that providing explicit supervision on unknowns through self-distillation enhances Mask3D's performance on the 200 classes of the ScanNet200 validation set.

5.3 Discussion and analysis

Ablation study: We show in Table 4 that using the weight function W_I with the loss improves the performance on the unknowns and the known classes in row 2 compared to row 1, as it offers a good balance between good and bad pseudo-labels. Also, the Unknown-Known separation improves the performance of the student model on the previously known classes and the unknown objects, compared to the naive self-distillation (when distilling the output of the teacher model) in row 1 in the table.

Table 7: The choice of score rescaling function W_I. We show the importance of rescaling the weights for a sample with a certain objectness score in the following table. The second row is when using the scores themselves as weights, while the third row is our choice of weighting function.

Task ID	**Task 1**				
Weight Function	WI	A-OSE	U-Recall	mAP (↑)	
	(↓)	(↓)	(↑)	Current known	All
Split A					
None	**0.543**	559	**43.47**	38.11	37.66
s	0.719	**480**	40.83	**39.59**	**39.09**
$W_I(s)$	0.841	564	41.27	39.29	38.81
Split B					
None	0.619	1179	30.35	24.68	24.39
s	**0.347**	**1104**	**31.24**	23.13	22.86
$W_I(s)$	0.873	1243	30.11	**26.03**	**25.71**
Split C					
None	1.015	1391	29.00	23.12	22.83
s	**0.838**	**1356**	28.19	22.36	22.08
$W_I(s)$	0.876	1403	**29.31**	**25.22**	**24.90**

Choice of the loss weight function: In Table 7, we show that using the objectness itself as a weight function improves the performance in Split A, but it negatively affects the performance in Split B and Split C. We hypothesize that this drop in performance in splits B and C is due to the low Unknown Pseudo-Labels to True labels Ratio (demonstrated in the suppl. material) for both splits, as the loss does not receive enough good pseudo-labels to allow the model to differentiate between known and unknown objects. We also show that choosing a weight function that scales up the score helps the model differentiate between unknowns and knowns, improving the known classes' performance on the three splits.

Open-Vocabulary results: Table 3 demonstrates that employing OpenDistill3D for 3D mask proposal generation enhances generalizability when evaluated on new datasets, surpassing Autolabeling utilized in 3D-OWIS or closed set training in Mask3D. Additionally, Table 2 illustrates that OpenDistill adapts well to new classes with varying geometries in the Replica dataset.

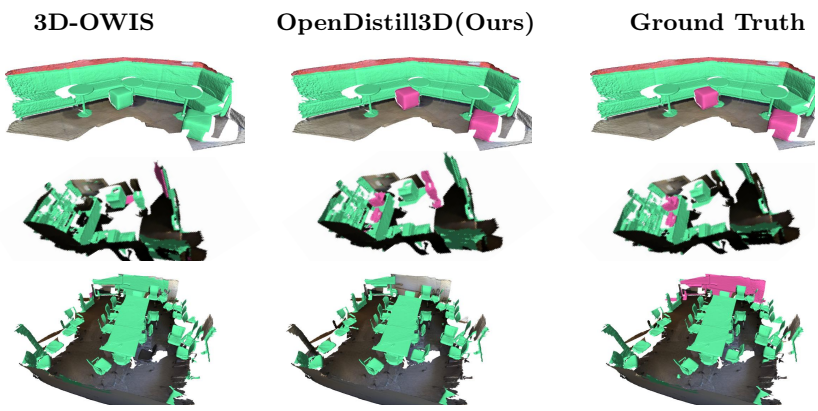

Fig. 5: Qualitative results. Our model exhibits qualitatively better results on the known classes (■) and unknown objects (■), compared to 3D-OWIS [3]. In the first scene, our model generates fewer false positives from the unknown classes, as it correctly classified the stool chair (unknown). In the second one, our model is capable of correctly segmenting background objects as unknowns. In the last one, we show that our model correctly segments the bottom-right chair (known) in the scene.

Qualitative results: We qualitatively show the superiority of our method in maintaining knowledge of the previously known classes in Figure 2, and improving the unknown object segmentation in Figure 5 compared to State Of The Art 3D-OWIS. In Figure 5, the background and unknown objects are correctly segmented due to the high-quality unknown object pseudo-labels used for training.

Limitations Our model is tested for indoor environments only and may exhibit sub-optimal performance in extremely sparse outdoor point cloud scenes since the performance of the student model depends on the quality of pseudo-labels generated by a teacher model. Additionally, our analysis of the pseudo-labels created by the SDM indicates that different pseudo-labels for unknown objects are generated when using augmented versions of the same scene. This observation suggests that incorporating various augmentations could enhance performance on the unknown objects.

6 Conclusion

We address the challenges of open-world 3D instance segmentation by introducing a novel approach with self-distillation that handles both continual learning and unknown object identification. Traditional methods treating these subproblems separately have shown limitations in handling unknown instances and mitigating catastrophic forgetting. The proposed method leverages self-distillation, utilizing pseudo-labels from the preceding task to enhance the model's performance in recognizing unknown objects during training while mitigating catastrophic forgetting.

Acknowledgement The computations were enabled by resources provided by NAISS at Alvis partially funded by Swedish Research Council through grant agreement no. 2022-06725, LUMI hosted by CSC (Finland) and LUMI consortium, and by Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the NSC.

References

1. Al Khatib, S., El Amine Boudjoghra, M., Lahoud, J., Khan, F.S.: 3d instance segmentation via enhanced spatial and semantic supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 541–550 (2023)
2. Bendale, A., Boult, T.: Towards open world recognition. In: CVPR (2015)
3. Boudjoghra, M.E.A., Al Khatib, S.K., Lahoud, J., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.: 3d indoor instance segmentation in an open-world. In: NeurIPS (2023)
4. Cen, J., Yun, P., Zhang, S., Cai, J., Luan, D., Wang, M.Y., Liu, M., Tang, M.: Open-world semantic segmentation for lidar point clouds (2022)
5. 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. pp. 5828–5839 (2017)
6. Dhamija, A., Gunther, M., Ventura, J., Boult, T.: The overlooked elephant of object detection: Open set. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1021–1030 (2020)
7. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 86–102. Springer (2020)
8. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: CVPR (2022)
9. 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. pp. 4421–4430 (2019)
10. Jiang, H., Yan, F., Cai, J., Zheng, J., Xiao, J.: End-to-end 3d point cloud instance segmentation without detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12796–12805 (2020)
11. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: CVPR (2023)
12. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
13. Liu, X., Wu, C., Menta, M., Herranz, L., Raducanu, B., Bagdanov, A.D., Jui, S., de Weijer, J.v.: Generative feature replay for class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 226–227 (2020)
14. Liu, Y., Schiele, B., Vedaldi, A., Rupprecht, C.: Continual detection transformer for incremental object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23799–23808 (2023)
15. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3243–3249. IEEE (2018)

16. Mullappilly, S.S., Gehlot, A.S., Anwer, R.M., Khan, F.S., Cholakkal, H.: Semi-supervised open-world object detection (2024)
17. 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. pp. 13550–13559 (2023)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2016)
19. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: European Conference on Computer Vision. pp. 125–141. Springer (2022)
20. Saito, K., Hu, P., Darrell, T., Saenko, K.: Learning to detect every thing in an open world. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 268–284. Springer (2022)
21. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d for 3d semantic instance segmentation. arXiv preprint arXiv:2210.03105 (2022)
22. Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: Proceedings of the IEEE international conference on computer vision. pp. 3400–3409 (2017)
23. 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. pp. 551–560 (2023)
24. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. NeurIPS (2023)
25. 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. pp. 2708–2717 (2022)
26. Wang, K., Liu, X., Bagdanov, A.D., Herranz, L., Jui, S., van de Weijer, J.: Incremental meta-learning via episodic replay distillation for few-shot image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3729–3739 (2022)
27. Wang, W., Feiszli, M., Wang, H., Tran, D.: Unidentified video objects: A benchmark for dense, open-world segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10776–10785 (2021)
28. 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. pp. 2569–2578 (2018)
29. Zohar, O., Wang, K.C., Yeung, S.: Prob: Probabilistic objectness for open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11444–11453 (2023)