2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

OW3Det: Toward Open-World 3D Object Detection for Autonomous Driving

Wenfei Hu[1] and Weikai Lin[1] and Hongyu Fang[1] and Yi Wang[2] and Dingsheng Luo[1,*]

***Abstract*— Despite their success in LIDAR object detection, modern detectors are vulnerable to uncommon instances and corner cases (e.g., a runaway tire) since they are closed-set and static. Networks under the closed-set setup only predict labels of seen classes, while static models suffer from catastrophic forgetting when gradually learning novel concepts. This motivates us to formulate the open-world 3D object detection task for autonomous driving, which aims to 1) tackle the closed-set issue by identifying unseen instances as unknown and 2) incrementally learn novel classes without forgetting previously obtained knowledge. To achieve the open-world objectives, we propose Open-World 3D Detector (OW3Det), the first framework for open-world 3D object detection. The OW3Det comprises a base detector, a self-supervised unknown identifier, and a knowledge-distillation-restricted incremental learner. Although knowledge distillation facilitates preserving memories, imposing penalties on areas containing unknown objects hinders the incremental learning process. We mitigate this hindrance by employing unknown-driven pivotal mask, which eliminates unnecessary restrictions on regions overlapping with novel instances. Abundant experiments and visualizations demonstrate that the proposed OW3Det attains state-of-the-art performance.**

I. Introduction

A reliable 3D detector is essential for the perception system of autonomous vehicles. Recent advancements [1], [2], [3], [4] in LIDAR point cloud detection methods have made significant strides on benchmarks such as KITTI [5] and nuScenes [6]. Despite this progress, real-world driving scenarios still present challenges due to the presence of unknown (unlabeled) classes. These unknown instances pose a safety risk as most conventional 3D detectors are closed-set and static. Figure1 illustrates a comparison of the ground-truth and detection results of several settings. The closed-set setup makes models detect all objects as categories encountered during training, assigning unseen instances with the labels of seen classes. Meanwhile, the static paradigm restricts detectors to limited driving scenes, as finetuning a static model to new classes carries the risk of forgetting the initially learned classes.

To address this issue, [7] formalize open-world recognition by updating the image classifier to recognize new classes. [8] extend the concept of open-world recognition to object detection. They utilize contrastive clustering to enhance the energy-based unknown identifier. To prevent forgetting previously learned information, they employ example replay during incremental learning. Whereas these methods take a step further, significant differences exist between 2D and 3D tasks, making open-world 3D object detection not yet fully explored. Existing methods such as the examplar replay and knowledge distillation imposing penalties on areas containing unknown objects, thus hindering the learning process of open-world 3D object detection. These methods have yet to fully leverage the potential of combining open-set detection and incremental learning. Despite being unfamiliar, humans can easily identify unseen objects as individual instances and incrementally learn novel concepts. It is our innate curiosity about the unknown that fuels the desire to learn new concepts. Unknown instances can provide valuable insights for more accurate incremental learning, as the novel classes incorporated by the incremental learner are a subset of the unknown classes.

To bridge the aforementioned gap, we formulate the open-

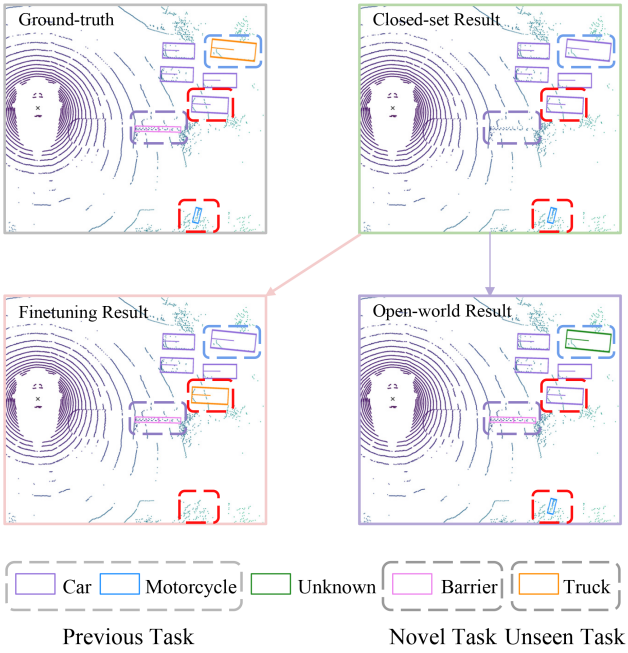

Fig. 1. Models under the closed-set setup fail to detect unknown objects (highlighted by the purple dotted rectangles) or assign unseen instances with the labels of old classes (highlighted by the blue dotted rectangles). Directly finetuning to novel classes leads to the forgetting of the initially learned tasks (highlighted by the red dotted rectangles). In contrast, the open-world model is capable of identifying unseen objects as unknown and incrementally learning these unseen classes.

[1] Wenfei Hu, Weikai Lin, Hongyu Fang and Dingsheng Luo are with the National Key Laboratory of General Artificial Intelligence, Key Laboratory of Machine Perception (MoE), School of Intelligence Science and Technology, Peking University and PKU-WUHAN Institute for Artificial Intelligence, Beijing, China.
[2] Yi Wang are with the Department of Automation, Tsinghua University, Beijing, 100084, China
* Dingsheng Luo is the corresponding author, email: dsluo@pku.edu.cn

world 3D object detection task for autonomous driving, which is composed of the open-set detection task and the incremental learning task. As shown in Figure1, the open-world model is capable of identifying unseen objects as unknown and incrementally learning these unseen classes. To achieve the open-world objectives, we propose the Open-World 3D Detector (OW3Det). We apply knowledge distillation (KD) restrictions on the network through logit KD [9] and label KD [10] to preserve memories. The logit KD extracts output responses from the previous model to provide guidance, while the label KD utilizes predictions from the original network for label assignments. To eliminate unnecessary restrictions during incremental learning, We propose the unknown-driven pivotal mask for knowledge distillation. Distinct from existing works, the knowledge distillation with the unknown-driven pivotal mask applies relatively weak supervision signals to pixels that overlap with unknown instances or belong to backgrounds, thereby removing unnecessary constraints. Meanwhile, regions with known objects are restricted by the distillation loss to alleviate forgetting. By utilizing the proposed mask, we further minimize misleading predictions to carry on accurate label assignments. To summarize, our contributions are as follows:

- To the best of our knowledge, our work is the first to formally define the open-world 3D object detection task for LIDAR point clouds, which is composed of the open-set detection task and the incremental learning task.
- We propose a novel framework, OW3Det, comprising a base detector, a self-supervised unknown identifier, and a knowledge-distillation-restricted incremental learner to address the challenges of open-world 3D object detection.
- The unknown-driven pivotal mask is employed to relieve unnecessary knowledge distillation restrictions.
- We construct evaluation protocols to measure the efficacy of open-world 3D detectors and demonstrate the proposed OW3Det achieves state-of-the-art performance.

II. Related Work

A. Closed-set 3D Object Detection

To capture features from sparse and irregular LIDAR point clouds, most existing approaches first used a 3D backbone to convert the point cloud into regular representations, such as regular voxels [11], or pillars [2]. Another stream of research [12] directly operated on raw point clouds without quantization before feature extraction. Since LiDAR and camera are a pair of complementary modalities, researchers began fusing multi-sensor information at the proposal-level [13] and point-level [14]. 3D detection heads were then employed to identify instances, including anchor-based [2], center-based [3], and transformer-based [4]. To reduce the computation overhead, [15] extended knowledge distillation to 3D object detection to build efficient 3D detectors by performing distillation on areas containing objects. This inspires us to further minimize distillation restrictions on regions containing unknown instances and focus on positions where exist known objects.

B. Open-world Recognition

Existing closed-set and static methods encountered significant obstacles when deploying in real-world scenarios. Therefore, endowing models with open-set and incremental learning capabilities has aroused significant interests. [16] formulated open-set recognition by balancing the performance of one-vs-rest classifiers and the open space risk. [17] preserved the original model by restricting the cross-entropy loss with knowledge distillation. [7] formalized open-world recognition by combining open-set and incremental tasks. [8] further extended open-world to object detection using the energy-based unknown classifier and the example replay. Based on vision transformers, a single-stage open-world object detector was proposed by [18] to better model the context and discover potential unknowns in an attention-driven manner. Despite these progress, open-world 3D object detection is still under-explored, which motivates us to propose the OW3Det to achieve open-world objectives.

III. OW3Det: Open-World 3D Detector

A. Problem Formulation

In this section, we formalize the definition of open-world 3D object detection. We denote by $\mathcal{K}^t = \{1, 2, \dots, C\}$ the set of C known object categories at time step t. To realistically simulate dynamic and diverse driving scenes, some unknown classes $\mathcal{U}^t = \{C+1, \dots\}$ might be encountered at test time. Let $\mathcal{P}$ be the input LIDAR point clouds with their associated ground-truth $\mathcal{G}^t$. Each point in the training sample $\mathcal{P}$ is represented by its 3D location and reflectance. The corresponding annotation $\mathcal{G}^t = \{\mathbf{g}_1, \dots, \mathbf{g}_N\}$ encodes ground-truth labels $\mathbf{g}_i = (y_i, \mathbf{q}_i, \mathbf{s}_i, \mathbf{e}_i)$ of N object instances, where y_i is the class label, $\mathbf{q}_i = (u_i, v_i, d_i)$ represents the 3D center location, the bounding box size is expressed by $\mathbf{s}_i = (w_i, l_i, h_i)$, and $\mathbf{e}_i = (\sin(r_i), \cos(r_i))$ denotes the yaw rotation along the z-axis.

In the open-world 3D object detection setting, a model M^t at time t is trained to identify objects belonging to previously encountered classes $\mathcal{K}^t$. Due to the potential security threat posed by certain unknown categories $\mathcal{N}^{t+1} = \{C+1, \dots, C+n\}$, a set of unknown instances $\mathbf{U}^t \subset \mathcal{U}^t$ are forwarded to the ground-truth oracle which can provide the associated annotations $\mathcal{G}^{t+1}$ of $\mathcal{N}^{t+1}$. It should be noted that instances of old classes are not annotated. The incremental learner is employed to gradually incorporate novel classes into the existing knowledge base without retraining from scratch. The known categories are expanded from $\mathcal{K}^t$ to $\mathcal{K}^{t+1} = \mathcal{K}^t \cup \mathcal{N}^{t+1}$. Simultaneously, the unknown identifier learns to assign instances belonging to the updated unseen classes $\mathcal{U}^{t+1} = \{C+n+1, \dots\}$ with the unknown label (denoted by 0). Detectors under the open-world settings continue the aforementioned cycle over their lifespan to adaptively update themselves without compromising the existing knowledge.

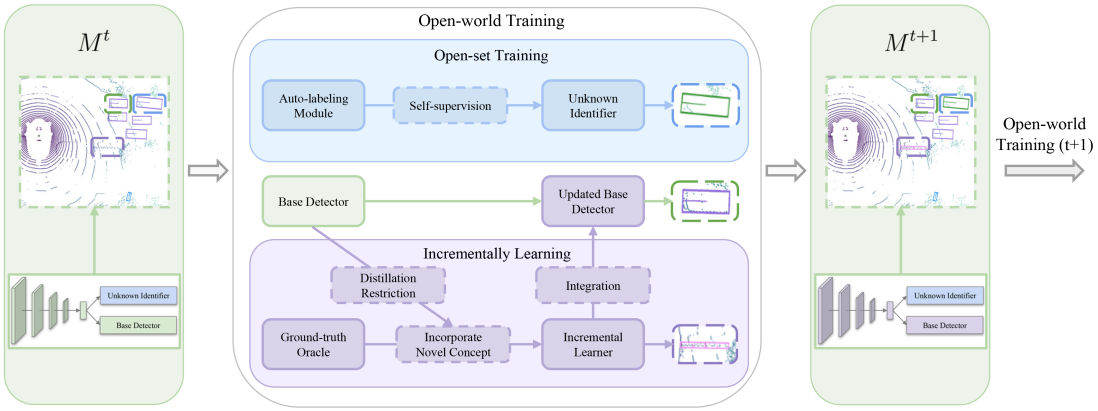

Fig. 2. The proposed OW3Det framework employs a static model M^t to generate preliminary detection results. To locate unseen instances, the auto-labeling module is introduced to train the unknown identifier in a self-supervised manner. The incremental learner gradually incorporates novel concepts into the existing knowledge base with restrictions from knowledge distillation. The open-world training enables the base detector to achieve developmental self-evolution by integrating the incremental learner.

B. Overall Architecture

An overview of our approach is illustrated in Figure2. A static model M^t is utilized to generate preliminary detection results. Due to the sparse and irregular nature of the lidar point cloud, we leverage 3D backbone [1] to extract the bird's eye view (BEV) level feature before conducting object detection utilizing detectors. The 3D backbone quantizes the point cloud into regular voxels. After inner-voxel feature extraction, the voxel-level feature is obtained and then flattened into the BEV-level feature. We adopt CenterPoint [3] as our base detector to identify instances belonging to known categories. By leveraging keypoint estimation to discover objects, CenterPoint simplifies 3D detection and accurately detects known objects in a single-stage manner.

However, there still exist numerous unseen objects in real-world driving scenes. To address these corner cases, we utilize the unknown identifier to discover unseen instances. This prevents the model from assigning known labels to unknown instances, thereby improving the reliability of the proposed OW3Det. The training procedure of the unknown identifier is conducted in a self-supervised manner to minimize the resource consumption caused by labeling numerous unknown objects. Meanwhile, we send unknown classes that pose safety hazards to ground-truth oracle to generate the corresponding annotations. Subsequently, the incremental learner is employed to learn these novel categories. We use knowledge distillation to alleviate the network forgetting of originally acquired knowledge. The unknown-driven pivotal mask is proposed to mitigate unnecessary restrictions. Positions containing objects detected by the base detector receive more enforcement, while regions containing unknown objects obtained from the unknown identifier are less restricted, enabling the incremental learner to incorporate novel classes.

C. Base Detctor

Given the input point cloud $\mathcal{P}$, the known category $\mathcal{K}^t$ and the ground-truth $\mathcal{G}^t$, we utilize $\mathcal{P}$ and $\mathcal{G}^t$ to train a CenterPoint detection head that can recognize categories belonging to $\mathcal{K}^t$ as the base detector. By taking BEV-level feature as input, the base detector locates objects via keypoint estimation. The predicted heatmap $\hat{H} \in [0,1]^{W\times H\times C}$ of size $W \times H$ is generated for C classes in $\mathcal{K}^t$, which is used to detect objects. Each local maximum (i.e., pixels whose heatmap score is greater than its eight neighbors) in $\hat{H}$ indicates an object center. The training target H of heatmap is obtained by drawing Gaussian kernels on the ground-truth object centers as follows:

$$H_{\mathbf{p},c} = \max_{i:y_i=c} \exp\left(-\frac{(\mathbf{p}-\mathbf{q}_i)^2}{2\sigma_i^2}\right), \tag{1}$$

in which $c \in \mathcal{K}^t$, $\mathbf{p}$ denotes the position at the target heatmap, and σ is the Gaussian radius controlling the size of the Gaussian peak. The use of Gaussian kernels, instead of simply setting the positions on the heatmap corresponding to the center of ground-truth objects to 1, provides denser supervision signals and accelerates convergence. We utilize focal loss [19] to calculate the training objective of the heatmap as follows:

$$\mathcal{L}_{hm}(\mathcal{K}^t) = \frac{-1}{N}\sum_{\mathbf{p},c}\begin{cases}(1-\hat{H}_{\mathbf{p},c})^{\alpha}\log(\hat{H}_{\mathbf{p},c}) \text{ if } H_{\mathbf{p},c}=1\\ (1-H_{\mathbf{p},c})^{\beta}(\hat{H}_{\mathbf{p},c})^{\alpha} \text{ otherwise}\\ \log(1-\hat{H}_{\mathbf{p},c})\end{cases}, \tag{2}$$

where N represents the number of ground-truth instances in $\mathcal{G}^t$, α and β are hyperparameters.

To retrieve bounding boxes, a regression map $\hat{R} = [\hat{S}, \hat{O}, \hat{E}]$ shared by all classes is simultaneously generated. This predicted regression map $\hat{R}$ is composed of three sub-maps, where $\hat{S}$ denotes the size prediction, $\hat{O}$ refers to the local offset, and $\hat{E}$ represents the rotation estimate. The loss of the regression map $\hat{R}$ can be expressed as follows:

$$\mathcal{L}_{reg}(\mathcal{G}^t) = \frac{1}{N}\sum_{i=1}^{N}(|\hat{S}_{\mathbf{q}_i} - \log(\mathbf{s}_i)| + |\hat{O}_{\mathbf{q}_i} - \mathbf{o}_i| + |\hat{E}_{\mathbf{q}_i} - \mathbf{e}_i|), \tag{3}$$

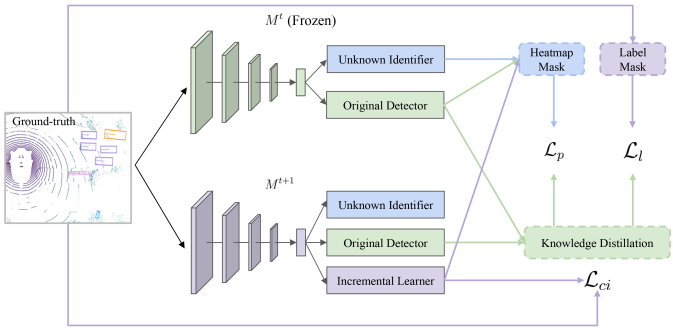

Fig. 3. Loss computation procedure of the incremental learner. We utilize knowledge distillation to preserve memories. The unknown-driven pivotal mask, including the heatmap mask and label mask, is employed to reduce unnecessary and misleading restrictions.

where $\mathbf{o}_i = \mathbf{q}_i - [\mathbf{q}_i]$ is introduced to recover the quantization error caused by the striding of the backbone. The training objective of CenterPoint at time step t is obtained by combining the heatmap loss and the regression loss as follows:

$$\mathcal{L}_c(\mathcal{K}^t, \mathcal{G}^t) = \mathcal{L}_{hm}(\mathcal{K}^t) + \mathcal{L}_{reg}(\mathcal{G}^t) \tag{4}$$

At time step t, the base detector is initialized according to the aforementioned procedure. At subsequent time steps $\tau > t$, the base detector is updated by integrating the incremental learner.

D. Unknown Identifier

Most existing 3D detection methods suffer from closed-set assumptions and label unknown instances as known objects. Confusing known objects with unknown instances during object detection leads to performance drops in downstream tasks like trajectory prediction and freespace detection, which can negatively impact the reliability of autonomous driving systems. To address this issue, we propose a self-supervised unknown identifier to discover unseen instances. However, annotating all unknown objects in the training data is infeasible due to the massive resource consumption. As a surrogate, we introduce a center-based auto-labeling module to provide supervision signals.

Specifically, the auto-labeling module decodes unknown object centers from the predicted heatmap $\hat{H}$. Pixels in $\hat{H}$ that have a high heatmap score and do not overlap with ground-truth objects are labeled as the object centers of candidate unknown objects. The properties of these unknown objects are retrieved from the corresponding position in the predicted regression map $\hat{R}$. We then select the top-K potential unknown objects sorted by heatmap score to create the pseudo-label set, denoted as $\mathcal{G}_s^\tau$. At time step $\tau \geq t$, the training objective of the self-supervised open-set training is calculated as follows:

$$\mathcal{L}_s = \mathcal{L}_c(\{0\}, \mathcal{G}_s^\tau), \tag{5}$$

where the set containing only the unknown category is denoted by $\{0\}$. Open-set training facilitates the OW3Det to learn clear class separation in the latent space, which is vital for discriminating unknown instances.

E. Incremental Learner

In real-world driving scenarios, traffic accidents can often occur as a result of uncommon instances and corner cases. To address these security risks, it is crucial to incorporate unknown categories into the existing knowledge base of autonomous vehicles without degrading the performance of previously trained models. However, this can be a resource-intensive process if done by retraining the models from scratch. To overcome this challenge, we propose the use of an incremental learner for adaptive knowledge acquisition. At time step $\tau \geq t+1$, we calculate the following loss to train the incremental learner:

$$\mathcal{L}_{ci} = \mathcal{L}_c(\mathcal{N}^\tau, \mathcal{G}^\tau). \tag{6}$$

As the labels of previous classes $\mathcal{K}^{\tau-1}$ are not provided during incremental learning, knowledge distillation is utilized to prevent the base detector from forgetting previously learned classes. Two techniques, logit KD and label KD, are employed for this purpose. With logit KD, the OW3Det preserves knowledge by mimicking the output heatmap and regression map predicted by the previous model. In label KD, predictions from the original network are used as pseudo-labels to provide supervision signals.

Motivated by the unbalanced ratio of informative foreground to redundant background in 3D scenes, a recent method [15] attempts to perform distillation according to heatmap score, which enables the model to focus on regions containing objects. However, utilizing the Gaussian kernel to construct the training objective for heatmaps leads to inflated heatmap scores for areas that are in proximity to known objects, even when those positions contain unknown objects. This limits the ability of the incremental learner to learn novel categories. While enforcing output-level imitation on positions containing known instances facilitates preserving memories, distillation on regions including unseen objects can be detrimental to the incremental learning process. In addition, the previous model may produce misleading results, which could also impose unnecessary constraints.

Distinct from previous methods, we propose the unknown-driven pivotal mask to alleviate the restrictions on areas overlapping with unseen instances during incremental learning. The Unknown-driven pivotal mask is composed of two constituent parts: a heatmap mask $\tilde{\mathbf{m}}$ responsible for logit KD and a label mask $\mathring{\mathbf{m}}$ responsible for label KD. In logit KD, we use $\tilde{\mathbf{m}}$ to control the penalties of knowledge distillation for each position in the heatmap, which can be calculated as follows:

$$\tilde{\mathbf{m}}_b = \hat{H}_b^{\tau-1}(1 - \hat{H}_u^{\tau-1})(1 - \max_{i \in \mathcal{N}^\tau} \hat{H}_i^{\tau-1}), \tag{7}$$

where $b \in \mathcal{K}^{\tau-1}$, $u \in \{0\}$, $\hat{H}_b^{\tau-1}$ denotes the heatmap predicted by the previous base detector, and $\hat{H}_u^{\tau-1}$ is obtained from the previous unknown identifier. Heatmap mask $\tilde{\mathbf{m}}$ indicates where to learn and what information should not

TABLE I

DATA SPLIT FOR THE PROPOSED EVALUATION PROTOCOL, WHERE CLASSES WITH ✓ REPRESENT BEING INTRODUCED AS PART OF THE CURRENT TASK. ABBREVIATIONS: CONSTRUCTION VEHICLE (C.V.), PEDESTRIAN (P.E.D.), MOTORCYCLE (MOTOR.), TRAFFIC CONE (T.C.).

Task		Car	Bus	Bicycle	P.E.D.	Truck	C.V.	Trailer	Barrier	Motor.	T.C.
Open-world	T^t	✓	✓	✓	✓						
	T^{t+1}					✓	✓	✓			
	T^{t+2}								✓	✓	✓
Incremental	5+5						✓	✓	✓	✓	✓
	7+3								✓	✓	✓
	9+1										✓

TABLE II

OPEN-WORLD 3D OBJECT DETECTION RESULTS ON THE PROPOSED EVALUATION PROTOCOL. THE METRIC MARKED WITH ↓ INDICATES THAT A LOWER VALUE REPRESENTS BETTER PERFORMANCE, WHEREAS THE METRICS WITH ↑ SHOULD HAVE HIGH VALUES FOR OPTIMAL PERFORMANCE. THE BEST RESULTS, EXCEPT FOR THE UPPER BOUND, ARE HIGHLIGHTED IN BOLD.

Task	T^t		T^{t+1}				T^{t+2}		
Method	mAOSE↓	mAP$_{cur}$ ↑	mAOSE↓	mAP$_{pre}$ ↑	mAP$_{cur}$ ↑	mAP$_{both}$ ↑	mAP$_{pre}$ ↑	mAP$_{cur}$ ↑	mAP$_{both}$ ↑
Upper Bound	11568.50	70.2	6247.25	70.2	37.8	56.3	56.3	66.7	59.4
Finetuning	15173.00	**70.7**	13067.50	0	**34.4**	14.7	0	**59.9**	17.9
OW-DETR	**12156.50**	70.3	8819.00	63.0	24.0	46.3	43.9	47.0	44.8
OW3Det			**8613.25**	**69.4**	33.4	**54.0**	**51.6**	58.9	**53.8**

be forgotten, which facilitates fine-grained enforcement of knowledge distillation. The loss function of logit KD can be expressed as follows:

$$\mathcal{L}_p = \sum_b \mathbb{E}[\tilde{\mathbf{m}}_b || \hat{H}_b^\tau - \hat{H}_b^{\tau-1}||_2] + \mathcal{L}_{reg}^{kd}(\hat{R}^\tau, \hat{R}^{\tau-1}), \quad (8)$$

where $\mathbb{E}$ represents the average operator, $\mathcal{L}_{reg}^{kd}$ denotes L1 regression loss, $\hat{R}^\tau$ and $\hat{R}^{\tau-1}$ are the current and previous regression maps, respectively. Only regions containing previously learned objects get strong penalties to mitigate forgetting, while positions with unknown instances and backgrounds are not supervised by the distillation loss.

We employ the label KD as a complementary approach to the logit KD. Object centers and properties are decoded from the previous base detector's heatmap $\hat{H}_b^{\tau-1}$ and regression map $\hat{R}^{\tau-1}$. We create a set of pseudo-labels, denoted as $\mathcal{G}_l^\tau$, using instances with heatmap scores greater than a threshold value of δ. To reduce misleading results, the label mask $\mathring{\mathbf{m}}$ is obtained by calculating the Intersection-over-Union (IoU) of each instance in the pseudo-label set $\mathcal{G}_l^\tau$ with the current ground-truth $\mathcal{G}^\tau$. Objects with an IoU greater than a threshold ε are filtered out. Training loss of label KD can be expressed as follows:

$$\mathcal{L}_l = \mathcal{L}_c(\mathcal{K}^{\tau-1}, \mathring{\mathbf{m}}\mathcal{G}_l^\tau). \quad (9)$$

As illustrated in Figure3, the training objective of incremental learning is obtained as follows:

$$\mathcal{L}_i = \mathcal{L}_{ci} + \mathcal{L}_p + \mathcal{L}_l. \quad (10)$$

Without unnecessary imitation on pixels containing unknown objects, the incremental learner is capable of extracting features about unknown classes in these positions. Accurate label KD is achieved by removing misleading results utilizing $\mathring{\mathbf{m}}$, which is crucial for incorporating novel concepts.

F. Joint open-world training

At time step t, we leverage the base detector and the unknown identifier to build M^t. The training loss is calculated as follows:

$$\mathcal{L}^t = \lambda_1 \mathcal{L}_c(\mathcal{K}^t, \mathcal{G}^t) + \lambda_2 \mathcal{L}_s, \quad (11)$$

where λ_1 and λ_2 are hyperparameters that control the contribution of each loss item. At the subsequent time step $\tau \geq t+1$, we conduct open-world training to build M^τ. The base detector achieves its own self-evolution by integrating the incremental learner. A new unknown identifier is obtained to directly replace the original one. The joint loss function at time step τ is the combination of the self-supervised loss and the incremental loss, as follows:

$$\mathcal{L}^\tau = \beta_1 \mathcal{L}_s + \beta_2 \mathcal{L}_i, \quad (12)$$

where β_1 and β_2 are hyperparameters. Following [8], [18], a set of exemplars is stored and replayed after each open-world training step.

IV. EXPERIMENTS AND RESULTS

A. Open-world Evaluation Protocol

Data split: The nuScenes dataset [6] is a large-scale autonomous-driving dataset for 3D detection and tracking. It includes a total of 1,000 scenes, with 700 designated for training, 150 for validation, and 150 for testing, respectively. As shown in TableI, for the open-world 3D detection task, we group classes in the nuScenes into three sub-tasks to simulate real-world driving scenes, denoted as $\mathcal{T} = \{T^t, T^{t+1}, T^{t+2}\}$. This allows researchers to evaluate their models in a variety of conditions, providing a more comprehensive assessment of open-world performance. At each time step τ, classes in $T^t, \ldots, T^\tau$ are treated as known $\mathcal{K}^\tau$, categories in current

TABLE III

INCREMENTAL LEARNING RESULTS OF 5+5, 7+3, AND 9+1 SETTINGS. THE BEST RESULTS, EXCEPT FOR THE UPPER BOUND, ARE IN BOLD

Setting	5+5			7+3			9+1		
Method	mAP_{pre}	mAP_{cur}	mAP_{both}	mAP_{pre}	mAP_{cur}	mAP_{both}	mAP_{pre}	mAP_{cur}	mAP_{both}
Upper Bound	67.7	51.1	59.4	56.3	66.7	59.4	58.0	72.3	59.4
Finetuning	0	**47.8**	23.9	0	**64.2**	19.3	0	**72.2**	7.2
Frozen Backbone	67.9	8.3	38.1	55.7	13.9	43.2	57.4	20.5	53.7
OW-DETR	59.0	32.8	45.9	47.8	47.4	47.7	53.2	55.5	53.4
OW3Det	**63.0**	44.3	**53.6**	**54.3**	60.4	**56.2**	**56.8**	62.8	**57.4**

classes T^τ are viewed as $\mathcal{N}^\tau$, and the remaining classes in $T^\tau, \ldots, T^{t+2}$ are considered as unknown $\mathcal{U}^\tau$. During the training on the current task T^τ, labels from previous tasks are not provided. The training scenes for each task are obtained from the nuScenes train set, while the testing scenes are taken from the nuScenes validation set, as the nuScenes test set does not include annotations.

Baseline: We refer to some representative methods from the open-world 2D object detection domain and re-implement them in the 3D object detection domain as our baselines. The Upper Bound represents the model that has access to all annotations. The Finetuning setting refers to the method of directly training using data from the new task. Following the recent work [18], we implement their open-world methods in 3D object detection, denoted as OW-DETR, including the unknown identifier and examplar replay. In this work, researchers use a transformer-based detector to identify objects in a single-stage manner, which is more similar to the proposed OW3Det model than the two-stage method [8].

Evaluation metrics: In 2D open-set object detection, the Absolute Open-Set Error (ASOE) [20] is used to report the number of unknown objects that are wrongly classified as any of the known classes. In the case of 3D point cloud instances, small instances easily achieve 0 intersection over union (IoU) even when detected with a minor translation error. To address this issue, the nuScenes dataset defines a match by thresholding the 2D center distance d on the ground plane instead of IoU. Therefore, we report the mean Absolute Open-Set Error (mAOSE) which is calculated by taking the average of the AOSE over a set of matching thresholds $\mathbf{D} = \{0.5, 1, 2, 4\}$. To benchmark the incremental capability, we calculate the mean Average Precision (mAP) averaged over the thresholds $\mathbf{D}$. mAP_{pre} is utilized to measure the performance of previous classes, mAP_{cur} represents the results on novel classes, mAP_{both} reports overall performance on known classes.

B. Open World 3D Object Detection Results

To demonstrate the effectiveness of the proposed OW3Det, we conduct experiments on the open-world evaluation protocol, as shown in TableII. In task T^t, the upper bound setting with access to categories achieves the lowest mAOSE by reducing the assignment of unknown objects as known classes. But this, to some extent, suppresses the prediction of known objects, resulting in a bit lower mAP_{cur} in T^t. The model under the Finetuning setting lacks the ability to handle novel objects, thereby yielding suboptimal performance on mAOSE. Additionally, the Finetuning model's mAP_{pre} is 0, due to the unavailability of labels from previous tasks during training on a new task. The all-zero matrix is used as the training target for the previous heatmap during training, which deteriorates predictions for the original classes. OW-DETR attains improved mAOSE and mAP_{pre} by utilizing the unknown identifier and the examplar replay. However, directly utilizing the exemplar replay requires making trade-offs between the novel and previous knowledge. Using a large learning rate for the backbone network is beneficial for learning novel categories, but it also causes significant parameter changes. This leads to the backbone disregarding some crucial features of the old categories, making it difficult to recover the network's performance through replay. Our proposed method, compared to OW-DETR, attains superior results on all metrics. We improve mAP_{pre} by introducing knowledge distillation restrictions. By employing the unknown-driven pivotal mask, unnecessary enforcement is removed to achieve a better mAP_{cur}. We also utilize the proposed mask to minimize misleading information, which has been shown to provide a performance gain on mAOSE.

TABLE IV

ABLATION STUDY ON THE 7+3 INCREMENTAL LEARNING SETTING. WE USE THE TERM ER TO REPRESENT EXEMPLAR REPLAY, KD TO DENOTE KNOWLEDGE DISTILLATION, AND UPM TO REFER TO THE UNKNOWN-DRIVEN PIVOTAL MASK. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

ER	KD	UPM	$\text{mAP}_{pre}\uparrow$	$\text{mAP}_{cur}\uparrow$	$\text{mAP}_{both}\uparrow$
			0	**64.2**	19.3
✓			47.8	47.4	47.7
	✓		51.2	59.7	53.7
	✓	✓	51.9	60.4	54.4
✓	✓	✓	**54.3**	60.4	**56.2**

C. Incremental 3D Object Detection Results

We further investigate the incremental capability of the OW3Det by conducting incremental experiments in multiple settings, listed in TableI. The x+y setting means splitting classes in nuScenes into previous task x and current task y. Each model is obtained by training with previous x classes first and then learning the current y categories incrementally. As shown in TableIII, the upper bound has

access to all classes. When directly finetuning the static to novel classes, we can obtain a model with competitive performance on mAP_{cur}. However, this approach results in severe forgetting, which negatively impacts mAP_{pre}. Despite freezing backbone parameters facilitating preserving knowledge, the mAP_{cur} significantly drops because freezing parameters restricts the extraction of class-specific features. The OW-DETR achieved balanced performance by making trade-offs between the new and old knowledge. Leveraging knowledge distillation, the OW3Det consistently outperforms all baseline methods in terms of mAP_{both}. The unknown-driven pivotal mask enables the distillation enforcement to focus on regions containing previously learned classes. This allows the incremental learner to incorporate new knowledge without severe performance drops on mAP_{pre}.

D. Ablation Study

To verify the efficiency of the components in the proposed OW3Det, we progressively integrate them into the Funetuing model. A summary of the evaluation results is shown in TableIV. Employing the exemplar replay is beneficial for recovering mAP_{pre}, but it requires making trade-offs between the performance of previous and current tasks. Knowledge distillation gives a large performance boost by providing supervision signals on the previous classes. This facilitates OW3Det to preserve previous knowledge without reducing the learning rate of the backbone network. The proposed unknown-driven pivotal mask mitigates unnecessary restrictions and eliminates misleading predictions, resulting in an enhancement of mAP_{both}. Combining all components together, OW3Det is capable of incorporating novel tasks without severe forgetting.

V. Conclusion

In this paper, we formulate the open-world 3D object detection task for autonomous driving. To achieve open-world objectives, we propose OW3Det, which comprises a base detector, a self-supervised unknown identifier, and a knowledge-distillation-restricted incremental learner. We employ knowledge distillation to mitigate forgetting. The unknown-driven pivotal mask, including the heatmap mask and the label mask, is proposed to alleviate unnecessary enforcement. Extensive experiments and visualizations demonstrate that the proposed OW3Det attains state-of-the-art performance. Moreover, a novel evaluation protocol is designed to measure the efficacy of open-world 3D detectors. In future work, we will explore the potential of extending the proposed method to multi-modal 3D object detection.

ACKNOWLEDGMENT

The work is supported in part by the National Natural Science Foundation of China (No. 62176004, No. U1713217), Intelligent Robotics and Autonomous Vehicle Lab (RAV), Wuhan East Lake High-Tech Development Zone National Comprehensive Experimental Base of Governance of Intelligent Society, and High-performance Computing Platform of Peking University.

References

[1] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[2] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[3] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[4] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[5] 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.

[6] 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.

[7] A. Bendale and T. Boult, "Towards open world recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1893–1902.

[8] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5830–5840.

[9] G. Hinton, O. Vinyals, J. Dean, *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[10] C. H. Nguyen, T. C. Nguyen, T. N. Tang, and N. L. Phan, "Improving object detection by label assignment distillation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1005–1014.

[11] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[12] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779.

[13] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao, "Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection," *arXiv preprint arXiv:2201.06493*, 2022.

[14] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[15] J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, "Towards efficient 3d object detection with knowledge distillation," *arXiv preprint arXiv:2205.15156*, 2022.

[16] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[17] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[18] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "Ow-detr: Open-world detection transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9235–9244.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[20] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3243–3249.