WI3D: Weakly Incremental 3D Detection via Vision Foundation Models

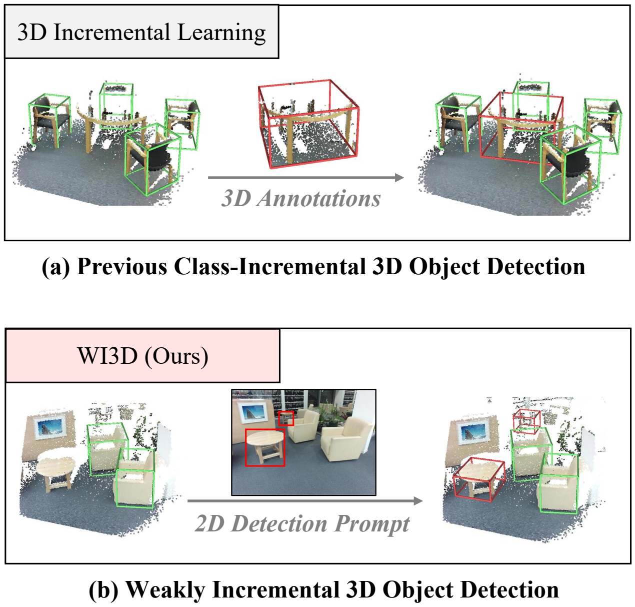
Mingsheng Li[1] Sijin Chen[1] Shengji Tang[1] Hongyuan Zhu[2] Yanyan Fang[1]
Xin Chen[3] Zhuoyuan Li[1] Fukun Yin[1] Tao Chen[1,†]
[1]Fudan University [2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore [3]Tencent PCG
[†] Corresponding Author

Abstract

Class-incremental 3D object detection demands a 3D detector to locate and recognize novel categories in a stream fashion while preserving its base detection ability. However, existing methods require delicate 3D annotations for learning novel categories, resulting in significant labeling costs. To this end, we explore a label-efficient approach called ***W****eakly* ***I****ncremental* ***3D*** ***D****etection (****WI3D****), which teaches a 3D detector to learn incrementally with off-the-shelf vision foundation models. We propose a novel dual-teaching framework incorporating both intra-modal and inter-modal knowledge from pseudo labels and feature space. Specifically, our framework features a class-agnostic pseudo-label refinement module, designed for generating high-quality 3D pseudo labels. This module is built on a lightweight transformer that models the spatial relationships between pseudo labels and their interactions with rich contextual information in point clouds. Additionally, we introduce a cross-modal knowledge transfer module to enhance the representation learning of novel classes, along with a reweighting knowledge distillation strategy that dynamically assesses and distills knowledge from previously learned categories. Extensive experiments show that our approach can efficiently learn novel concepts while preserving knowledge of base classes in WI3D scenarios, and surpass baseline approaches on both SUN-RGBD and ScanNet.*

1. Introduction

Existing 3D detectors [32, 37, 41, 45, 47] have achieved remarkable performance in detecting objects within a predefined category set from the 3D environment. However, novel-class objects will emerge when deploying 3D detectors in wild and dynamic scenarios. To adapt a well-trained 3D detector to novel classes, a straightforward approach would be to directly tune the model on novel-class samples. However, the direct tuning typically results in the degradation of a detector's ability to detect base classes, a phenomenon known as catastrophic forgetting [7, 12, 27, 50].



(a) Previous Class-Incremental 3D Object Detection

(b) Weakly Incremental 3D Object Detection

Figure 1. **Illustration of previous class-incremental 3D object detection (a) and WI3D (b).** Previous methods for class-incremental 3D object detection rely heavily on the continual provisions of human annotations for novel classes in the point cloud scene. In contrast, we explore WI3D, a new task that introduces novel concepts to a 3D detector through 2D images to reduce the cost of annotating the point cloud.

Meanwhile, an alternative strategy involves combining base datasets with novel ones and retraining the model from scratch. Nevertheless, this approach becomes impractical when frequent updates are necessary, as training on the entire dataset would be time-consuming [3]. Recently, incremental learning [23, 31, 39, 46], which studies how to incorporate novel classes by training only on novel-class samples while preventing catastrophic forgetting issues, has become eminent in various 2D and 3D vision tasks [9, 10, 22, 44].

Prior works [24, 55, 58] have made initial attempts in the field of class-incremental 3D detection using delicate 3D annotation for novel-class objects. However, acquiring a large number of well-annotated 3D scene data is prohibitively expensive in both 3D data collection and labeling [21, 40]. In light of these considerations, it is meaningful to study whether the capabilities of well-trained 3D detectors can be extended to recognize new categories without continuous annotation in the point cloud. In this work, we introduce **W**eakly **I**ncremental **3D** **D**etection (WI3D), which incrementally updates the base 3D detector through cost-effective vision foundation models [18, 20, 26, 51–53], avoiding the need to revisit 3D labels for novel classes, as shown in Fig. 1. To the best of our knowledge, we are the *first* attempt to leverage pre-trained foundation models for addressing WI3D, an unexplored yet important problem.

However, it is non-trivial to adapt existing methods for WI3D. The main difficulty lies in: **1)** how to introduce novel classes to a 3D detector continually using foundation models, and **2)** how to retain base classes knowledge without revisiting any 3D annotations. Recent studies [30, 35] have made initial attempts to directly generate 3D pseudo labels from 2D predictions using projection matrix, and utilize these pseudo labels as supervisory signals. However, these approaches fail to adequately address the significant noise in the 3D pseudo label derived from the 2D plane given the additional degrees of freedom. The existence of such pseudo labels severely deteriorates the detection performance in WI3D. Furthermore, the widely adopted knowledge distillation techniques [55, 58] treat different regions of interests equally, leading to the failure to learn discriminative region features among the sparse and cluttered point cloud scenes.

To address these issues, we propose an effective framework for WI3D, benefiting from the dual teaching of both intra-modal and inter-modal teachers. The intra-modal teacher is a base 3D detector trained on a fixed set of categories, while the inter-modal teacher is an off-the-shelf yet powerful 2D foundation model. Our framework is supervised by 1) pseudo labels generated by intra-modal and inter-modal teachers and 2) concept representation space for both base and novel classes. In practice, we use the inter-modal teacher to detect novel classes from images as visual prompts, and then project them into 3D space using the projection matrix. To obtain more accurate pseudo labels, we propose a novel pseudo-label refinement module that utilizes the coordinates of coarse proposals and context information from point clouds to refine bounding boxes of novel classes in 3D scenarios. By learning both the intrinsic relationships among pseudo labels and interactions with their context, our pseudo-label refinement module significantly enhances the accuracy and reliability of the pseudo labels.

In addition to incrementally teaching a well-trained 3D detector to detect novel categories explicitly, we also leverage an implicit way of supervision by learning in feature space. We propose an auxiliary cross-modal knowledge transfer for WI3D, which leverages bipartite matching to transfer texture-aware information from regions of images to enhance the 3D object representation. Finally, we explore a reweighting knowledge distillation strategy that can discern and select valuable knowledge from existing classes, leading to further improvements in performance.

To summarize, our contributions are listed as follows:

- We introduce Weakly Incremental 3D Detection (WI3D), a new task that generalizes a well-trained base 3D detector to learn novel classes with the aid of off-the-shelf foundation models.
- We analyze the challenges in WI3D and propose an effective framework that infuses a class-agnostic pseudo-label refinement module for high-quality pseudo-label generation and concept representation learning in feature space for both base and novel classes.
- Extensive experiments on two benchmark datasets, SUN RGB-D and ScanNet, illustrate the effectiveness of our methods under the low-cost setting of WI3D scenarios.

2. Related Work

In this section, we first briefly review existing methods for class-incremental detection in 2D and 3D. Then, we introduce work on weakly-supervised 3D detection and the design of existing 3D object detectors.

Class-Incremental Detection explores the task of incrementally learning and detecting new classes over time while preserving the original capabilities of the detector as much as possible. [14, 28, 34, 49] have made great efforts to class-incremental image object detection. Concurrently, several attempts for class-incremental 3D detection are proposed. SDCoT [55] proposes a static-dynamic co-teaching method for class-incremental 3D object detection. DA-CIL [58] proposes a 3D domain adaptive class-incremental object detection framework with a dual-domain copypaste augmentation method to adapt the domain gradually. Recent work I3DOD [24] proposes a task-shared prompts mechanism to learn the matching relationships between the object localization information and category semantic information for class-incremental 3D object detection. In this paper, we explore a new paradigm, WI3D, to study how 2D knowledge enables a 3D detector to learn novel objects continually, without the reach for labor-consuming 3D annotations for the novel classes.

Weakly-Supervised 3D Detection studies a way to train a 3D detector without detailed instance annotations. BR [48] proposes to train 3D detectors using a few key points, such as object centers. WyPR [40] develops an approach that relies solely on scene-level class labels for training 3D detection models. Additionally, SESS [56] further proposes a semi-supervised 3D detection framework with a novel perturbation scheme. Recently, OV-3DET [30] and CoDA [1] introduce open-vocabulary 3D object detection, which uti-

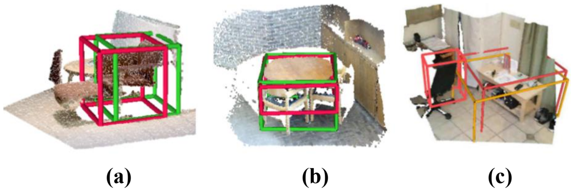


Figure 2. **The challenge of generating accurate 3D pseudo-labels from 2D predictions.** (1) *Projection Migration* occurs due to the background pixels within a 2D bounding box, causing the displacement of the 3D bounding box compared with the ground-truth label (marked as green). (2) *Scale Ambiguity.* Due to the sparse representation of object surfaces, the pseudo labels (marked as red) generated cannot encompass the entire table legs. (3) *Overlapped Boxes* arise when aggregating pseudo labels from multi-view images.

lizes a pre-trained 2D model to generate pseudo labels for 3D detectors. However, both OV-3DET [30] and CoDA [1] focus on associating each class-agnostic bounding box with a relevant text prompt by computing similarities in the classification head, which can't handle the problems of *incremental localization and recognition* of emerging objects in the scene. In addition, how to acquire *accurate* 3D pseudo labels from 2D predictions remains unexplored in these approaches. In this paper, we study the potential of utilizing the visual foundation model for weakly incremental 3D detection by learning from denoised pseudo labels and regional concept representation.

3D Object Detectors requires a model to localize objects of interest from a 3D scene input. [4, 29, 32, 37, 54] manage to operate directly on the point clouds for 3D object detection. VoteNet [37] and H3DNet [54] achieve end-to-end 3D object detection based on sampling, grouping, and voting operators designed especially for point clouds. 3DETR [32] and GroupFree3D [29] extend the transformer [43] architecture to 3D object detection. In our paper, we adopt the modified VoteNet [37] proposed by SDCoT [55] as our detection backbone and explore how to extend a base 3D detector with the ability to detect objects of novel classes through the off-the-shelf foundation models.

3. Methodology

In Sec. 3.1, we define the task setting of WI3D and analyze the noise of 3D pseudo labels directly generated from 2D predictions. Then, we provide the overview of our dual-teaching framework for WI3D in Sec. 3.2, which supervises the student with both the denoised pseudo labels from teachers (Sec. 3.3) and concept representation learning in feature space (Sec. 3.4). Finally, we offer the training objectives in Sec. 3.5.

3.1. Problem Definition

Task Definition. Given a well-trained 3D detector capable of localizing and recognizing base category set C_{base} from point cloud, WI3D extends its capacity to detecting a larger category set $\mathcal{C}_{all} = \mathcal{C}_{base} \cup \mathcal{C}_{novel}$ with only visual prompts for C_{novel} from off-the-shelf 2D models.

Coarse Pseudo Labels Generation. To generate novel-class labels for $\mathcal{S}^{3D}$ without point-level annotations, we employ a strategy similar to that described in [30], leveraging predictions from a cost-free 2D teacher $\mathcal{T}^{2D}$. Specifically, we first project the point cloud onto the image plane via the projection matrix and select points within each 2D bounding box. DBSCAN [13] is then adopted to segment points within each 2D box into multiple clusters, based on the density of points. After that, we drop the clusters that contain fewer points than $1/10$ of the number of points in that 2D box, which ensures the generation of a tight 3D instance mask. The cluster with the largest population is selected, and PCA is used to calculate a coarse bounding box, including its center, size, and rotation angle.

Noise Analysis. However, it is worth noting that straightforward clustering cannot precisely distinguish targets from noisy points, resulting in the following challenges when generating 3D pseudo labels from 2D predictions: *Projection Migration*: As shown in Fig. 2(a), background pixels within 2D bounding boxes lead to the displacement of the 3D bounding box. (2) *Scale Ambiguity*: As shown in Fig. 2(b), the scale ambiguity problem often arises because 3D sensors capture only sparse points on an object's surface, leading to inaccurate dimension estimations for the entire object. (3) *Overlapped Boxes*: As shown in Fig. 2(c), duplicated estimations on the same instance will occur when fusing multi-frame predictions (e.g. the red and yellow pseudo labels represent the predicted results from two consecutive frames, respectively). To reduce noise in the generated 3D bounding boxes, we propose a novel pseudo label refinement module that utilizes coarse pseudo labels from the clustering algorithm and context information from the point cloud.

3.2. Pipeline Overview

Our pipeline is initialized with a base 3D detector $\mathcal{T}^{3D}$, which is capable of detecting $\mathcal{C}_{base}$. As is shown in Fig. 3, in order to train a 3D detector $\mathcal{S}^{3D}$ for both incrementally detecting *novel* classes and retaining *base* knowledge without any reach of 3D annotations, we seek supervision from both pseudo labels and feature space. More accurate 3D pseudo labels become within reach for $\mathcal{S}^{3D}$ by further adopting the proposed pseudo label refinement module in Sec. 3.3. Additionally, cross-modal knowledge transfer and intra-modal knowledge distillation in Sec. 3.4 serve as great feature-level supervision for concept learning across both base and

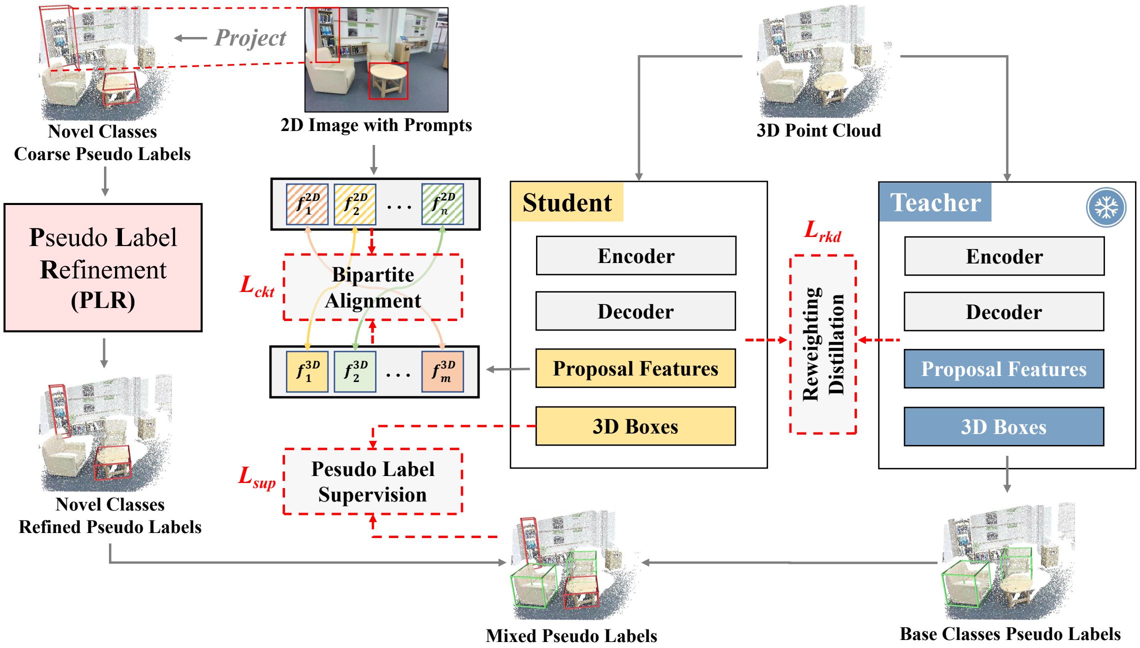


Figure 3. **The overview of our framework.** We train a 3D student detector $\mathcal{S}^{3D}$ on (1) 3D pseudo labels and (2) visual concept representations generated by both the inter-modal teacher $\mathcal{T}^{2D}$, and intra-modal teacher $\mathcal{T}^{3D}$. Specifically, 3D point cloud and paired images are used as inputs, with the vision foundation model $\mathcal{T}^{2D}$ employed to detect novel classes in the images. The 3D pseudo labels, which directly supervise $\mathcal{S}^{3D}$, are generated by denoising and blending the predictions of $\mathcal{T}^{2D}$ and $\mathcal{T}^{3D}$. Concurrently, the visual concept representation learning includes cross-modal knowledge transfer for novel class and reweighting knowledge distillation for base classes. During inference, $\mathcal{S}^{3D}$ takes the point cloud as input and predicts objects for both base and novel classes. Color is used for visualization only.

novel classes.

3.3. Pseudo Label Refinement

To generate more precise 3D pseudo labels from 2D predictions, we propose a novel class-agnostic **P**seudo **L**abel **R**efinement(PLR) module. As depicted in Fig. 4, PLR takes coarse bounding boxes and the encoded contextual information from the point cloud as input and generates refined pseudo labels. This module focuses on modeling intrinsic relationships among coarse pseudo labels and interacting with contextual information. By incorporating relative dependencies from pseudo labels and rich information within the context, PLR enhances object localization in 3D scenarios.

Box-aware Feature Aggregation (BFA). Inspired by the query learning in DETRs [2, 5, 6, 32], our model learns the positional relationships among different proposals using a multi-head self-attention network. In practice, we utilize a box encoder, $\mathcal{E}_{box}$, composed of several fully connected layers, to extract positional information for all bounding boxes. These box-aware features are then transformed into query, key, and value inputs through linear layers. The self-attention network then computes an attention map that dynamically models the interrelationships between boxes, allowing the model to aggregate features for enhancing the representation of proposals.

Box-Context Interaction (BCI). Given that context from point clouds can provide rich spatial information and global features, our box-context querying benefits from the multi-head cross-attention mechanism, which facilitates information passing between the box-aware features and the context embedding. Specifically, we use a lightweight point cloud encoder [36], $\mathcal{E}_{PN}$, to encode global information from the input scene $\tilde{p}$. The output of $\mathcal{E}_{PN}$ is then transformed by two separate linear layers to form the key and value inputs. By using box-aware features as queries interacting with context embedding, crucial context features for proposals are effectively captured to refine the pseudo labels.

The output from the feed-forward network (FFN) is then fed into a residual predictor, consisting of fully connected layers, to estimate the residual coordinates for each proposal. To address the issue of *box overlap*, we further propose a method to improve the reliability of each 3D pseudo label by incorporating an additional **B**inary **C**lassification **H**eader (BCH). The BCH takes the output of the FFN as input and generates a binary probability, determining the validity of each pseudo label. During training, we employ the Hungarian algorithm [19] to match each annotation with a corresponding pseudo label. Proposals that successfully match an annotation are assigned a binary probability of presence (labeled as 1) and absence (labeled as 0). Conversely, unmatched pseudo labels are marked with a probability of 0 for presence and 1 for absence. During inference, the bounding box is considered valid only when the probability of presence exceeds the probability of absence.

Since PLR is class agnostic, we can train it on base classes and apply it to novel classes during incremental learning, without additional training.

3.4. Concept Representation Learning

Beyond the denoising module for generating high-quality 3D pseudo boxes explicitly, we introduce auxiliary objectives to enhance the student's representation learning capability in an implicit way.

Cross-modal Knowledge Transfer. Compared to sparse point clouds, images possess rich texture features, offering significant advantages in expressing visual semantic information. However, directly projecting 3D proposals onto a single-view image plane to construct 3D-2D region pairs can result in different 3D boxes pointing to the same or nearby image regions, making it difficult to learn distinctive feature representations when they serve as a source of supervision. To this end, we propose **C**ross-modal **K**nowledge **T**ransfer (CKT) to help the student learn robust feature representations. Inspired by [25], we frame the cross-modal feature transfer as a matching problem and use bipartite matching to align region-level features across point clouds and images.

In practice, we project the 3D bounding box generated by $\mathcal{S}^{3D}$ onto the corresponding image $B_i^{3D\to 2D}$ and build the matching matrix by calculating IoU between the projected 3D box with 2D predictionsB_j^{2D} generated by $\mathcal{T}^{2D}$. The cost function for bipartite matching can be formulated as follows:

$$\begin{cases} max \sum_i \sum_j m_{ij} * IoU(B_i^{3D\to 2D}, B_j^{2D}) \\ \sum_i m_{ij} = 1, \end{cases} \tag{1}$$

where $m_{ij} \in \{0,1\}$ indicates whether it matches, and IoU represents the intersection over union. Then a pre-trained image encoder $\mathcal{E}^{2D}$ (i.e. CLIP [38]) is used to extract the features of $\mathcal{T}^{2D}$'s predictions B_i^{2D} from images, denoted as:

$$F_j^{2D} = \mathcal{E}^{2D}(B_j^{2D}). \tag{2}$$

For 3D proposal B_i^{3D} paired with corresponding B_j^{2D}, we feed the proposals' features F_i^{3D} in B_i^{3D} into an MLP-based projection head $\mathcal{H}^{3D}$, to encode the 3D proposal features into the same feature space of F_j^{2D}, denoted as

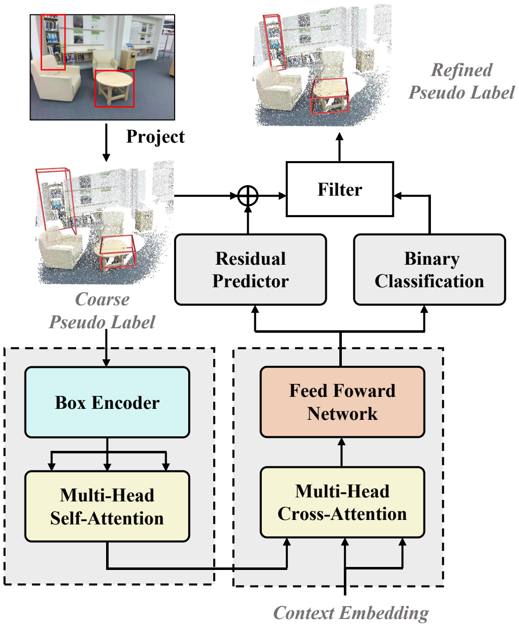


Figure 4. **The design of Pseudo Label Refinement(PLR).** PLR encodes initial coarse 3D bounding boxes and contextual information from the point cloud to generate refined pseudo labels. It utilizes a lightweight transformer specifically designed for aggregating box-aware features and facilitating box-context interaction. Additionally, PLR includes a residual predictor to decouple features and predict coordinate offsets for proposals, as well as a binary classification header to ensure the reliability of pseudo labels. Color is used for visualization only.

$F_i^{3D'} = \mathcal{H}^{3D}(F_i^{3D})$. Finally, we design a cross-modal knowledge transferring loss based on the negative cosine similarity [8]:

$$\mathcal{L}_{ckt} = -\sum_{i\in I, j\in J} \frac{F_i^{3D'}}{\|F^{3D'}\|_2} * \frac{F_j^{2D}}{\|F^{2D}\|_2}, \tag{3}$$

where I and J represent the number of 3D proposals and 2D predictions, respectively.

Intra-modal Base Knowledge Distillation. To alleviate forgetting issues, existing works [55, 58] use knowledge distillation [16] to preserve learned knowledge. However, previous work usually utilizes all the predicted responses and treats knowledge equally, failing to capture discriminative proposal features in sparse and cluttered point cloud scenes. In this work, we propose **R**eweighting **K**nowledge **D**istillation(RKD), which selectively distills features from the old teacher model to address the challenge of catastrophic forgetting.

Specifically, we employ an MLP-based classification head, which takes features from regions of interest as inputs and produces objectness scores o_i for all proposals. We introduce reweighting modulation factors, α_i, which dynamically adjusts the contribution of each proposal to the distillation loss based on the objectness scores o_i. This modulation factors α_i is defined as:

$$\alpha_i = \frac{e^{o_i}}{\sum_{i=1}^{K} e^{o_i}}, \tag{4}$$

where K is the total number of proposals. In RKD, α_i is utilized to modulate the traditional knowledge distillation loss, which compares the predictions of student model against those of teacher model. The distillation loss, $\mathcal{L}_{rkd}$, is computed as follows:

$$\mathcal{L}_{rkd} = \frac{1}{K}\left[\sum_{i\in\Phi_B} \alpha_i(||F_i^S - F_i^T||_2 + ||l_i^S - l_i^T||_2)\right], \tag{5}$$

where Φ_B is the set of indices of base-class proposals; F_i and l_i are the features and classification logits of the i_{th} proposal, respectively; the superscripts S and T denote the student and teacher models.

3.5. Training Objectives

Base Training. We train the modified VoteNet [55] (i.e. teacher model $\mathcal{T}^{3D}$) on base class annotations with the detection loss $\mathcal{L}_{det}$ [37], which is defined as

$$\mathcal{L}_{det} = \alpha_1\mathcal{L}_{vote} + \alpha_2\mathcal{L}_{obj} + \alpha_3\mathcal{L}_{box} + \alpha_4\mathcal{L}_{sem-cls}. \tag{6}$$

Here, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are set as $1, 0.5, 1, 0.2$, and $\mathcal{L}_{vote}$, $\mathcal{L}_{obj}$, $\mathcal{L}_{box}$, $\mathcal{L}_{sem-cls}$ stands for vote regression, proposal objectness classification, box regression, and proposal semantic classification respectively. Note that we also train PLR on C_{base} in this stage, where $\mathcal{L}_{PLR} = \mathcal{L}_{box}$.

Weakly Incremental Learning. The supervision of the WI3D comes in two folds: explicit detection training on the pseudo labels generated by $\mathcal{T}^{2D}$ and $\mathcal{T}^{3D}$ with $\mathcal{L}_{det}$, and concept representation learning in the feature space, which includes novel-class knowledge transfer, denoted as $\mathcal{L}ckt$, and base-class knowledge distillation, represented by $\mathcal{L}rkd$. The loss function can be defined as

$$\mathcal{L} = \beta_1\mathcal{L}_{det} + \beta_2\mathcal{L}_{ckt} + \beta_3\mathcal{L}_{rkd}. \tag{7}$$

Here, $\beta_1, \beta_2, \beta_3$ are set as $1, 10, 5$ heuristically.

4. Experiments

We first introduce the datasets, metrics, and implementation details for weakly incremental 3D object detection in Sec. 4.1. Then, we compare our methods with different baseline approaches and prior arts in Sec. 4.2. Afterward, we take out ablation studies to study the effectiveness of the proposed components in Sec. 4.3. Finally, we showcase some visualization results in Sec. 4.4.

4.1. Datasets, Metrics, and Implementation Details

Datasets. Following previous works on class-incremental 3D detection [24, 55, 58], we conduct experiments on two widely used datasets, SUN RGB-D [42] and ScanNet [11]. SUN-RGBD consists of 10,335 single-view RGB-D scans, where 5,285 are used for training, and 5,050 are for validation. Each scan is annotated with rotated 3D boxes. ScanNet includes 1,201 training samples and 312 validation samples reconstructed from RGB-D sequences. We split the full category set into two non-overlapped subsets into $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ according to [55].

Metrics. To compare the performance of different approaches under incremental settings, we adopt mAP_{base}, mAP_{novel}, and mAP_{all} as abbreviations for **m**ean **A**verage **P**recision (mAP) under an IoU threshold of 0.25 for base classes, novel classes, and overall performance.

Modification on VoteNet. While VoteNet [37] is recognized for its efficiency in 3D object detection, its reliance on random sub-sampling and fixed prediction scores during training make it unsuitable for continually incorporating novel classes. To address these constraints, we modify the original VoteNet by the principles of SDCoT [55]. The key modifications include: 1) reusing indices of sampled points in the base model; 2) altering the final layer to separate category predictions from spatial predictions and dynamically updating the classifier's weights for novel classes.

More Implementation Details. The input of our student and intra-modal teacher is a point cloud $P \in \mathbb{R}^{N\times3}$ representing a 3D scene, where N is set as 20,000 and 40,000 respectively for SUN RGB-D and ScanNet. We use Grounding Dino [26], a robust zero-shot image detector capable of generating high-quality boxes and labels with free-form text, as the inter-modal teacher. We feed the image and novel-class text prompts (category 1, ..., category N) into Grounding Dino [26] to detect novel-class objects, selecting the proposals with box confidence greater than 0.35 and class confidence above 0.25 as pseudo labels. The PLR comprises just one layer, incorporating multi-headed attention with four heads and a two-layer MLP with 128 hidden dimensions. Following [55], the base training lasts for 150 epochs using an Adam optimizer [17] with a batch size of 8, and a learning rate of 10^{-3} decaying to 10^{-4} and 10^{-5} at the $80th$ and $120th$ epoch respectively. During weakly incremental learning, we copy weights from $\mathcal{T}^{3D}$ to initialize the student model $\mathcal{S}^{3D}$, and optimize $\mathcal{S}^{3D}$ under the supervision of both refined pseudo labels and feature space. During both training stages, we evaluate $\mathcal{S}^{3D}$ every 10 epochs. Each experiment is conducted on a single RTX3090 GPU.

Table 1. **Weakly-incremental 3D object detection (mAP@0.25) on SUN RGB-D validation set. All methods listed are first trained on base classes $|C_{base}| = 10 - |C_{novel}|$ before incremental learning novel classes $|C_{novel}|$. ↑ means the higher, the better.**

Method	$\lvert\mathcal{C}_{novel}\rvert = 3$			$\lvert\mathcal{C}_{novel}\rvert = 5$			$\lvert\mathcal{C}_{novel}\rvert = 7$		
	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
base-training	53.84	-	-	58.54	-	-	50.88	-	-
fine-tuning	1.02	35.41	11.34	1.11	32.98	17.05	0.13	27.25	19.12
freeze-and-add	53.05	9.99	40.13	56.29	5.99	31.21	47.11	1.95	15.50
	39.64	45.38	41.36	49.27	34.08	41.68	49.75	25.83	33.00
Ours:									
WI3D	41.82	**51.01**	**44.58**	51.79	**43.07**	**47.43**	51.26	**35.37**	**40.14**
Upper Bounds:									
3D-GT	44.82	67.69	51.68	54.27	58.89	56.58	55.10	56.73	56.24

Table 2. **Weakly-incremental 3D object detection (mAP@0.25) on ScanNet validation set.** All methods listed are first trained on base classes $|C_{base}| = 18 - |C_{novel}|$ before incremental learning novel classes $|C_{novel}|$. ↑ means the higher, the better.

Method	$\lvert\mathcal{C}_{novel}\rvert = 6$			$\lvert\mathcal{C}_{novel}\rvert = 9$			$\lvert\mathcal{C}_{novel}\rvert = 12$		
	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
base-training	51.01	-	-	58.37	-	-	64.70	-	-
fine-tuning	1.66	27.42	10.24	2.42	20.72	11.57	3.96	17.32	12.87
freeze-and-add	50.33	1.96	34.21	58.10	2.08	30.09	63.30	1.56	22.14
	38.97	23.45	33.80	47.46	20.07	33.77	51.99	16.83	28.55
Ours:									
WI3D	41.75	**31.49**	**38.33**	49.18	**29.75**	**39.47**	52.34	**26.71**	**34.85**
Upper Bounds:									
3D-GT	52.85	61.31	55.67	59.40	51.73	55.56	63.59	51.40	55.46

4.2. Comparison with Existing Methods

Since there is no prior method that directly works around WI3D, we mainly compare our method with several baselines, including: 1) ***Base-training*** directly train the 3D detector on base classes. 2) ***Fine-tuning*** tune the whole model (except the base classifier) and a new classifier (randomly initialized) for the C_{novel}. 3) ***Freeze-and-add*** freeze the backbone, followed by adding a new classification head and training only the new head on novel classes. Additionally, we modify the training of ***SDCoT*** [55] to fit our weakly incremental learning setting. In practice, we additionally introduce the process of clutter and projecting to SDCoT [55] to generate 3D pseudo labels from 2D predictions for novel-class learning. For a fair comparison, all the training settings, e.g., learning rate, optimizer, batch size, etc., are the same for all experiments.

To make thorough evaluations, we compare our method with all the mentioned methods under different weakly class-incremental settings on SUN RGB-D (Tab. 1) and ScanNet (Tab. 2), which include: a) $|\mathcal{C}_{novel}| < |C_{base}|$; b) $|\mathcal{C}_{novel}| = |C_{base}|$; c) $|\mathcal{C}_{novel}| > |C_{base}|$. One shall notice that under different settings, the baseline methods either lead to catastrophic forgetting or failure to learn novel concepts. For instance, when we evaluate the methods on SUN RGB-D with $|\mathcal{C}_{novel}| = 5$, ***fine-tuning*** only achieves 1.11 mAP$_{base}$, and ***freeze-and-add*** achieves 5.99 mAP$_{novel}$. The former suffers from severe catastrophic forgetting on base classes, while the latter cannot learn new classes effectively. Additionally, it can be shown that our method can also surpass ***SDCoT***, which achieves 49.27% mAP$_{base}$, 34.08% mAP$_{novel}$ and 41.68%mAP$_{all}$ when $|\mathcal{C}_{novel}| = 5$, while our framework achieves 51.79% mAP$_{base}$, 43.07%mAP$_{novel}$(+8.99%), 47.43% mAP$_{all}$(+5.75%) under the same task setting on SUN RGB-D dataset. Compared to SDCoT, which experiences significant performance degradation when introducing novel classes, our method maintains a balance between base and novel classes, achieving superior performance across a variety of class-incremental scenarios. These phenomena are prevalent and can be observed through experiments conducted on both datasets in Tab. 1 and Tab. 2.

4.3. Ablation Study and Analysis

In this section, we organize ablation studies to study the effectiveness of the proposed components. Without further specification, the following experiments are conducted on SUN RGB-D under the $|\mathcal{C}_{novel}| = 5$ setting.

Robustness to Different 2D Foundation Models. To validate the robustness of our approach for different 2D teachers, we employed four distinct teachers in our framework: "Faster R-CNN" [15], "FastSAM" [57], "Grounding Dino" [26], and 2D human annotations ("2D Oracle"). Specifically, we train "Faster R-CNN" [15] using 2D object bounding boxes from the SUN RGB-D dataset [42]. Vision foundation models such as "FastSAM" [57] and "Ground-

Table 3. **Robustness to different 2D foundation models.** We organize ablation studies to validate the robustness of our method to different 2D teachers. "Vanilla" denotes the baseline without PLR (details in Sec. 3.3), CKT and RKD (details in Sec. 3.4).

Model	Bacgbone	Vanilla			Ours		
		mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
Faster RCNN [15]	RN50	47.72	24.82	36.27	51.21	34.14	42.68
Fast-SAM [57]	YOLOv8	47.63	29.71	38.67	51.13	41.71	46.42
Grounding Dino [26]	Swin-B	47.78	32.16	39.97	51.79	43.07	47.43
2D-Oracle	-	50.38	34.58	42.48	53.14	46.79	49.97

Table 4. **Effectiveness of PLR.** We analyze whether the removal of the pseudo label refinement module affects weakly incremental learning performance on the SUN RGB-D dataset. "-" denotes the absence of PLR.

Pseudo Label Denosing	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
-	50.44	35.62	43.03
NMS [33]	50.78	36.35	43.57
PLR w/o BCH	51.32	41.46	46.39
PLR	51.79	**43.07**	**47.43**

Table 5. **Analysis of BFA and BCI.** The model achieves the best results only when both the box-aware feature aggregation (BFA) and box-context interaction (BCI) are taken into account.

BFA	BCI	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
✓	×	50.11	39.28	44.70
×	✓	50.45	41.37	45.91
✓	✓	51.79	**43.07**	**47.43**

ing Dino" [26] are directly used to infer on these images without any fine-tuning. "2D Oracle" represents the results annotated by human experts on images. The results in Tab. 3 illustrate improvements achieved by our approach with different 2D teachers, particularly in recognizing new classes of objects. For instance, when using existing detectors such as Faster RCNN [15], our approach shows a +3.49% improvement on base classes, +9.32% on novel classes, and +6.41% across all categories. For general-purpose foundation models like Grounding Dino [26], it achieves a +4.01% improvement on the base classes and a +10.91% on novel classes. Furthermore, our method achieves a +7.49% improvement across all categories when applied to manually annotated 2D labels, demonstrating that continuous training of the 3D detector using vision foundation models is a feasible option.

Analysis of PLR. To make a better comparison, we include several baseline methods, including directly training with coarse pseudo labels ("-"), **N**on-**M**aximum **S**uppression [33] ("NMS"), and Pseudo Label Refinement without the Binary Classification Head ("PLR w/o BCH") in Tab. 4. It can be seen that the full model of our proposed PLR efficiently improves the detection performance of novel classes (+7.45% mAP$_{novel}$) compared to using coarse pseudo labels. Since NMS [33] is initially designed to drop duplicated box estimations, it cannot handle the challenge of noisy pseudo labels generated from 2D box estimations well. Additionally, BCH can efficiently select pseudo-labels with higher quality, and further improve the detection performance (+1.61% mAP$_{novel}$ and +1.04% mAP$_{all}$).

Analysis of BFA and BCI. In Tab. 5, we investigate the roles of the Box-aware Feature Aggregation and Box-Context Interaction in pseudo label refinement. We notice that using either BFA or BCI alone will severely downgrade the detection performance, as each component is insufficient to provide adequate information to refine the coarse 3D pseudo boxes. The designed PLR module effectively leverages spatial relationships between proposals provided by BFA, as well as contextual information from the point cloud via BCI, resulting in optimal detection performance. By integrating both components, we observe significant improvements compared to using a single one, with an increase of +3.79% and +1.7% in mAP$_{novel}$ respectively.

Analysis of CKT. In Tab. 6, we compare our proposed Cross-modal Knowledge Transfer with two variants: one that solely relies on projection without instances matching between image and point cloud ("CKT w/o Match"), and a baseline method ("-") without $\mathcal{L}_{ckt}$. One can see that the strategy relying merely on projection ("CKT w/o Match") performs worse than the baseline method, with a decrease of 0.47% in mAP$_{novel}$. This decline is attributed to overlapping 3D projections onto a single-view image plane, which hampers learning of distinctive features when used for supervision. Meanwhile, our proposed matching-based CKT is able to help $\mathcal{S}^{3D}$ learn robust novel knowledge representations (+1.34% mAP$_{novel}$).

Analysis of RKD. We conduct experiments in Tab. 7 to compare the effectiveness of our proposed Reweighting Knowledge Distillation with other commonly used knowledge distillation strategies. To be specific, [16] computes the **K**ullback-**L**eibler (KL) divergence, while [55] computes the l_2 distance of the semantic logits for each proposal between the teacher and student model. As shown in Tab. 7,

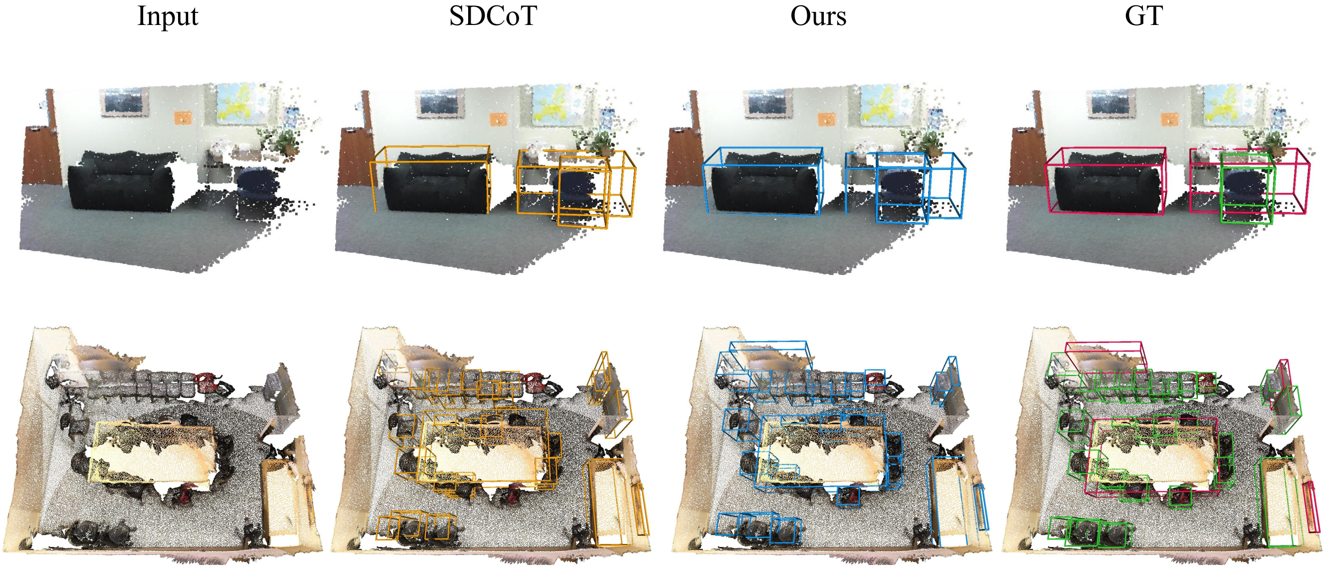


Figure 5. **Visualization of detection results.** Our proposed method is able to generate tight bounding boxes for both novel classes and base classes in these complex and diverse scenes. 3D ground truth annotations for these scenes are marked for base (marked as green) and novel (marked as red) classes respectively.

Table 6. **The performance of cross-modal knowledge transfer (CKT) by bipartite matching.** We compare the CKT utilizing bipartite matching with the unmatched approach. “-” denotes the absence of CKT.

Strategy	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
-	51.57	41.73	46.65
CKT w/o Match	51.65	41.26	46.46
CKT	51.79	**43.07**	**47.43**

our proposed RKD achieves a higher performance for both base (51.79 mAP$_{base}$) and novel (43.07 mAP$_{novel}$) classes.

Table 7. **Effectiveness of reweighting knowledge distillation (RKD) for weakly incremental 3D object detection.** We compare our proposed RKD with other commonly used knowledge distillation manner. “-” denotes that no distillation technology is used.

Distillation	mAP$_{base}$ ↑	mAP$_{novel}$ ↑	mAP$_{all}$ ↑
-	50.01	40.76	45.39
Hinton et. al.	50.89	41.12	46.01
Zhao et. al.	51.07	42.37	46.72
RKD	**51.79**	**43.07**	**47.43**

4.4. Qualitative Results

We showcase some qualitative results of our proposed methods on SUN RGB-D [42] and ScanNet [11] in Fig. 5. One can see that our proposed method is capable of generating tight bounding boxes for both novel and base classes.

5. Conclusions and Limitations

In this paper, for the *first* time, we attempt to address weakly incremental 3D object detection, dubbed WI3D, which is a new approach introducing both the **continuous** *localization* and *recognization* ability of novel classes to a well-trained 3D detector through off-the-shelf vision foundation models. By learning from both inter-modal and intra-modal teachers, we propose (1) a novel pseudo-label refinement module to improve the quality of 3D pseudo labels, and (2) concept representation learning in feature space for both base and novel classes. Experiments on SUN-RGBD and ScanNet demonstrate that our proposed framework surpasses all baselines, including the previous approach to class-incremental 3D object detection. We fervently aspire that our endeavors in the realm of label-efficient 3D class-incremental learning tasks will spark inspiration and fuel future explorations in this community.

References

[1] Yang Cao, Zeng Yihan, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[3] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental learning in semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4371–4381, 2022. 1

[4] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. *arXiv preprint arXiv:2311.18651*, 2023. 3

[5] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11124–11133, 2023. 4

[6] Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang YU, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2024. 4

[7] Wei Chen, Haoyang Xu, Nan Pu, Yu Liu, Mingrui Lao, Weiping Wang, Li Liu, and Michael S. Lew. Lifelong fine-grained image retrieval. *IEEE Transactions on Multimedia*, 25:7533–7544, 2023. 1

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020. 5

[9] Wei Cong, Yang Cong, Jiahua Dong, Gan Sun, and Henghui Ding. Gradient-semantic compensation for incremental semantic segmentation. *IEEE Transactions on Multimedia*, 26: 5561–5574, 2024. 1

[10] Yawen Cui, Wanxia Deng, Xin Xu, Zhen Liu, Zhong Liu, Matti Pietikäinen, and Li Liu. Uncertainty-guided semi-supervised few-shot class-incremental learning with knowledge distillation. *IEEE Transactions on Multimedia*, 25: 6422–6435, 2023. 1

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 9

[12] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Bridging non co-occurrence with unlabeled in-the-wild data for incremental object detection. *Advances in Neural Information Processing Systems*, 34:30492–30503, 2021. 1

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[14] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9427–9436, 2022. 2

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 7, 8

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5, 8

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[19] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[21] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. M3dbench: Let's instruct large models with multi-modal 3d prompts. *arXiv preprint arXiv:2312.10763*, 2023. 2

[22] Mingsheng Li, Lin Zhang, Mingzhen Zhu, Zilong Huang, Gang Yu, Jiayuan Fan, and Tao Chen. Lightweight model pre-training via language guided knowledge distillation. *IEEE Transactions on Multimedia*, pages 1–11, 2024. 1

[23] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1

[24] Wenqi Liang, Gan Sun, Chenxi Liu, Jiahua Dong, and Kangru Wang. I3dod: Towards incremental 3d object detection via prompting. *arXiv preprint arXiv:2308.12512*, 2023. 1, 2, 6

[25] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 5

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023. 2, 6, 7, 8

[27] Yu Liu, Xiaopeng Hong, Xiaoyu Tao, Songlin Dong, Jingang Shi, and Yihong Gong. Model behavior preserving for class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[28] Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23799–23808, 2023. 2

[29] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 3

[30] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. *arXiv preprint arXiv:2304.00788*, 2023. 2, 3

[31] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022. 1

[32] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 1, 3, 4

[33] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 8

[34] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters*, 140:109–115, 2020. 2

[35] Liang Peng, Senbo Yan, Boxi Wu, Zheng Yang, Xiaofei He, and Deng Cai. Weakm3d: Towards weakly supervised monocular 3d object detection. In *International Conference on Learning Representations*, 2022. 2

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[37] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1, 3, 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[39] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 1

[40] Zhongzheng Ren, Ishan Misra, Alexander G Schwing, and Rohit Girdhar. 3d spatial recognition without spatially labeled 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13213, 2021. 2

[41] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: fully convolutional anchor-free 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 477–493. Springer, 2022. 1

[42] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 6, 7, 9

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[44] Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Han-Jia Ye, Yatao Bian, De-Chuan Zhan, and Peilin Zhao. Beef: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[45] Haiyang Wang, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, Liwei Wang, et al. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35: 29975–29988, 2022. 1

[46] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 1

[47] Tao Xie, Li Wang, Ke Wang, Ruifeng Li, Xinyu Zhang, Haoming Zhang, Linqi Yang, Huaping Liu, and Jun Li. Farp-net: Local-global feature aggregation and relation-aware proposals for 3d object detection. *IEEE Transactions on Multimedia*, 26:1027–1040, 2024. 1

[48] Xiuwei Xu, Yifan Wang, Yu Zheng, Yongming Rao, Jie Zhou, and Jiwen Lu. Back to reality: Weakly-supervised 3d object detection with shape-guided label enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8438–8447, 2022. 2

[49] Binbin Yang, Xinchi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, and Xiaodan Liang. Continual object detection via prototypical task correlation guided gating mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9255–9264, 2022. 2

[50] Yang Yang, Zhen-Qiang Sun, Hengshu Zhu, Yanjie Fu, Yuanchun Zhou, Hui Xiong, and Jian Yang. Learning adaptive embedding considering incremental class. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2736–2749, 2021. 1

[51] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 2

[52] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023.

[53] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 2

[54] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 311–329. Springer, 2020. 3

[55] Na Zhao and Gim Hee Lee. Static-dynamic co-teaching for class-incremental 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3436–3445, 2022. 1, 2, 3, 5, 6, 7, 8

[56] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 2

[57] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. 7, 8

[58] Ziyuan Zhao, Mingxi Xu, Peisheng Qian, Ramanpreet Pahwa, and richard chang. Da-cil: Towards domain adaptive class-incremental 3d object detection. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 1, 2, 5, 6