VLM-PL: Advanced Pseudo Labeling approach for Class Incremental Object Detection via Vision-Language Model

[1,2]Junsu kim   [1,2]Yunhoe Ku[*]   [1]Jihyeon Kim[*]   [1]Junuk Cha   [1]Seungryul Baek

[1]UNIST   [2]MODULABS

Abstract

In the field of Class Incremental Object Detection (CIOD), creating models that can continuously learn like humans is a major challenge. Pseudo-labeling methods, although initially powerful, struggle with multi-scenario incremental learning due to their tendency to forget past knowledge. To overcome this, we introduce a new approach called Vision-Language Model assisted Pseudo-Labeling (VLM-PL). This technique uses Vision-Language Model (VLM) to verify the correctness of pseudo ground-truths (GTs) without requiring additional model training. VLM-PL starts by deriving pseudo GTs from a pre-trained detector. Then, we generate custom queries for each pseudo GT using carefully designed prompt templates that combine image and text features. This allows the VLM to classify the correctness through its responses. Furthermore, VLM-PL integrates refined pseudo and real GTs from upcoming training, effectively combining new and old knowledge. Extensive experiments conducted on the Pascal VOC and MS COCO datasets not only highlight VLM-PL's exceptional performance in multi-scenario but also illuminate its effectiveness in dual-scenario by achieving state-of-the-art results in both.

1. Introduction

The pursuit of artificial intelligence that mimics human-like continuous learning has led to significant exploration in Class Incremental Learning (CIL). CIL aims to develop methods that enable models to incorporate new classes while maintaining expertise in ones already learned. This endeavor seeks to tackle a central challenge: how to expand a model's knowledge base without eroding its current knowledge, a phenomenon known as catastrophic forgetting [73].

To address this challenge, the academic community has predominantly adopted three strategies: regulariza-
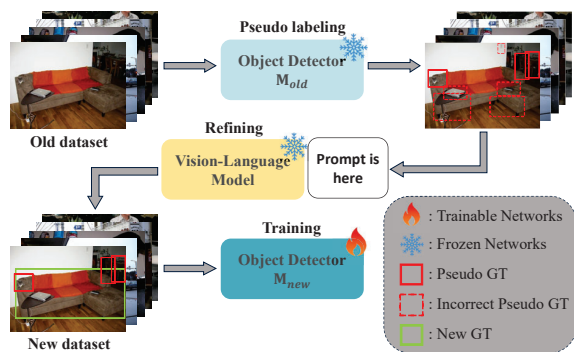
---

[*]Equal contribution.

Figure 1. *Workflow of our proposed method*: This schematic illustrates the sequential steps of our method. It starts with pseudo-labeling by a pre-trained model M_{old}, followed by refining through the Vision-Language Model. Custom-generated prompts are used for each pseudo ground-truth (GT). This refining process filters out incorrect pseudo GTs to yield reliable pseudo GTs. These annotations are then used to train a detector M_{new}, incorporating previous knowledge with the updated dataset.

tion, knowledge distillation, and replay. Regularization methods [34, 38, 59, 62, 90, 92] aim to preserve previous learning by penalizing changes to critical parameters. Knowledge distillation [27, 51, 61, 70, 79] techniques, on the other hand, focus on transferring knowledge from an older model version to its updated form, ensuring that the new model retains the ability to perform well on old tasks. Replay approaches, categorized into partial experience replay [8, 15, 28, 70, 75] and deep generative replay [13, 24, 31, 76, 83, 84], combat catastrophic forgetting [73]. The former retains a subset of previous tasks' data, utilizing it as a memory buffer while training new tasks. The latter uses generative models to re-experience past tasks' data.

Class incremental learning (CIL) strategies, designed to embrace new knowledge while retaining previous learnings, often encounter limitations when applied to the more intricate class incremental object detection (CIOD). CIOD, which is considered a separate domain, deals with the com-

plex challenge of detecting multiple labels within a scene. As a result, CIOD necessitates advanced methodologies to enhance detection capabilities across varied labels without compromising on the accuracy of identifying previously learned classes.

In the field of CIOD, pioneering researches [1, 23, 55, 78] have played a crucial role in advancing the field. These innovations extend methodologies initially crafted for single-class classification to the multifaceted challenges of CIOD, yielding notable advancements. Furthermore, the transition of detection frameworks from traditional CNN-based [48, 72] towards transformer-based [5, 44, 97], introducing research directions that capitalize on the attention mechanism's ability to enhance model generalization capabilities. In this base-frameworks transition, pseudo-labeling has been utilized to boost model performance in many CIOD strategies for alleviating forgetting. For example, techniques like class-wise distillation using pseudo ground-truth (GT) in CL-DETR [58], along with methods used in OW-DETR [26] and SDDGR [36], involve pseudo labeling approach using the confidence score from the classification branch of object queries. Despite their innovative applications, these techniques are fundamentally dependent on the performance of previously trained models. This reliance introduces significant limitations, especially in multi-incremental scenarios. As the complexity of scenarios increases, the retention of knowledge from earlier tasks weakens, leading to a discernible decline in performance. This decline is attributed to inaccuracies in pseudo GTs generated based on prior models' knowledge.

To address the problems mentioned earlier, we introduce VLM-PL, a new approach method that uses Vision-Language Models [18, 54, 87] (VLMs) to enhance pseudo-labeling by refining incorrect pseudo GTs. This method ensures the consistent use of accurate pseudo ground-truths (GTs) in various scenarios. Drawing inspiration from recent studies [16, 31, 36, 80] in pre-trained foundation models [50, 69, 74], our approach utilizes the vast knowledge of these models. This counters the significant performance drop that pseudo-labeling strategies often face due to reliance on previously trained models, especially in multi-class incremental object detection settings. Specifically, we employ prompt-tuning with VLMs to identify reliable pseudo GTs, eliminating the need for model retraining for classification tasks. This strategy effectively reduces error accumulation in complex scenarios and exhibits strong performance in both multi and dual scenarios. Moreover, it achieves state-of-the-art results for both Pascal and COCO datasets without the need for a replay strategy. The flow of the proposed approach is illustrated in Figure 1. Our contributions are organized as follows:

- To the best of our knowledge, we are pioneers in integrating VLM into CIOD, addressing challenges not primarily

tackled in this field before.
- Our method introduces effective prompt-tuning of VLM and input-output flows that accommodate scenarios with multiple incremental class additions, thus combating the usual performance declines seen in such challenging situations.
- Extensive experiments show that our approach excels not only in multi-incremental scenarios but also sets a new state-of-the-art in single incremental scenarios, thereby revealing the impact of VLM assistance on object detection.

2. Related works

2.1. Continual learning

Continual learning represents a broad spectrum of machine learning strategies designed to equip models with the capability to learn continuously, accumulating knowledge over time without catastrophic forgetting [73] the previously acquired information. Within this broad domain, class incremental learning (CIL) is a critical subset, focusing specifically on the model's ability to integrate new classes seamlessly.

Class incremental learning. CIL primarily addresses the challenge of classification, where models learn to identify new classes over time while retaining accuracy on previously learned classes. The essence of CIL is based on its methodologies, which can be broadly categorized into three key strategies: regularization, distillation, and replay. *Regularization* methods, such as [34, 38, 59, 62, 90, 92], aim to restrict the model's parameter updates to preserve knowledge across tasks. *Distillation* methods [27, 51, 61, 79] focus on the transfer of knowledge from a teacher model to a student model, where the teacher model, which encapsulates prior knowledge, guides the student model to retain old information while learning new. *Replay* involves reusing a subset of original data (partial experience replay [3, 7, 25, 39, 68, 70]) or generating new data samples mimicking the old data distribution (generative replay [13, 24, 76, 83, 84]), to prevent the learned knowledge forgetting. However, expanding the principles of CIL to object detection introduces a harder challenge: class incremental object detection(CIOD). Unlike CIL, where the focus is on classifying single objects within an image, CIOD involves detecting and classifying multiple object instances across various classes within the same image. Consequently, researchers have started investigating the specific strategy for CIOD.

Class incremental object detection. Recent developments in CIOD have explored both methodologies, such as CNN-based [4, 6, 10, 22, 53, 57, 63, 64, 72] and transformer-based [17, 26, 35–37, 58]. In general, transformer-based methods have been more focused on their superior gen-

eralization ability compared to CNN-based. Within this context, CL-DETR [58], OW-DETR [26], and Open-world DETR [17] have effectively utilized the architectural strengths of transformers, such as pseudo-labeling and class-wise distribution replay, while kang *et al.* [35] suggested both fine-grained distillations (*i.e.* DMD and IFD) strategy. Despite these innovations, such methods [17, 26, 58] often over-depend on the performance of the previous model when trained in multi incremental scenarios. To address this, we propose a new approach employing a Vision-Language Model (VLM), integrating the capabilities of a large language model and vision model, aiming to surpass the limitations inherent in reducing the dependency on the previously trained models within the pseudo-labeling strategy.

2.2. Vision-Language Models

In recent years, the performance of large language models (LLMs) has been advancing rapidly, leading to active research in vision models based on the superior performance of LLMs. Initially, models in the form of BLIP [45, 46] or GLIP [47] were investigated, focusing on simple captioning of images. Recently, research has progressed beyond image captioning to various tasks such as question answering [2, 12, 14, 19, 42, 54, 86, 96] using CLIP [69] and visual encoder [21, 45, 46]. Specifically, InternLM [19] achieved superior performance by proposing to train only modules that extract visual features using LoRA [30]. Alongside, the domain of region-based question answering [9, 11, 66, 82, 87, 87, 89, 93, 94] has witnessed considerable advancements, enabling detailed conversations about specific regions within an image. In particular, the Ferret [87] proposed a spatial-aware visual sampler to enable any-prompt visual input, regardless of the format of the prompt. We consider that the large-scale knowledge inherent in model [87] holds the potential to mitigate catastrophic forgetting, so we propose to utilize this potential with conventional method such as pseudo-labeling in our work.

3. Preliminaries

3.1. Transformer-based detector

The advent of transformer-based architectures [5, 56, 91, 97], especially the Detection Transformer (DETR) series, has significantly advanced the field of object detection. DETR models utilize an efficient attention [81] mechanism to diminish the inductive bias found in conventional detection frameworks [4, 71, 72], crucial for distinguishing foreground from background entities and extracting global feature relationships. Thus, it ultimately enhances the model's generalization capabilities [5]. Furthermore, a notable innovation is their ability to predict object categories and bounding boxes directly through self-attention and cross-attention

layers, eliminating the need for traditional post-processing techniques like non-maximum suppression. This process involves processing a predefined number of object queries, each predicting class probabilities and bounding box (bbox) coordinates. As a result, DETR-based detectors simplify the pipeline and reduce the dependency on handcrafted features, ensuring more robust detection outcomes. We use Deformable DETR [97] (D-DETR) for its efficient object query handling and reduced computational costs.

3.2. Region-specific conversation

Ferret [87] represents a significant advancement in Vision-Language Models (VLMs) by enabling precise, region-specific question-answering capabilities within images. It can detail conversation by employing a hybrid region representation, which integrates both discrete features (*e.g.* points and boxes) and continuous features (*e.g.* strokes, scribbles, or complex polygons). Ferret is utilizing the CLIP [69] model for visual features, achieves a detailed capture of image features by resizing images to 336×336 and extracting feature embeddings $Z \in \mathbb{R}^{H \times W \times C}$. Here, H is image height, W is image width and C is the image channel. Textual embeddings are derived using a pre-trained Language Model tokenizer, providing embeddings $T \in \mathbb{R}^{L \times D}$ where L and D denote sequence length and embedding dimension, respectively. The visual embeddings are projected to match the text embeddings dimension D, facilitating the seamless integration of visual and textual input. To integrate visual features into the language model's text prompts, Ferret employs a placeholder approach, indicated by <SPECIAL>. This allows the language model to process visual information alongside textual data. Specifically, to inform the language model of the exact location of a feature within an image, the question is formatted as "a region $[x_1, y_1, x_2, y_2]$ <SPECIAL>", where $[x_1, y_1, x_2, y_2]$ denotes the coordinates of the region of interest. This methodology enables the model to understand and respond to queries about specific areas within an image, significantly enhancing its ability to provide precise, context-aware answers.

4. Methods

4.1. Process configuration

In the domain of class incremental object detection (CIOD), the primary objective is seamlessly integrating new class labels into the model without diminishing its performance on classes that have already been learned. This process is divided into a series of separate tasks, represented as $\mathcal{T}_d$, where d ranges from 1 to N, indicating the total number of tasks. Each task, $\mathcal{T}_d$, is uniquely identified with a dataset $\mathcal{D}_d$ comprising a collection of images, denoted as $\mathcal{X}_d = \{x_d^1, x_d^2, \ldots, x_d^L\}$, and their respective annotations,
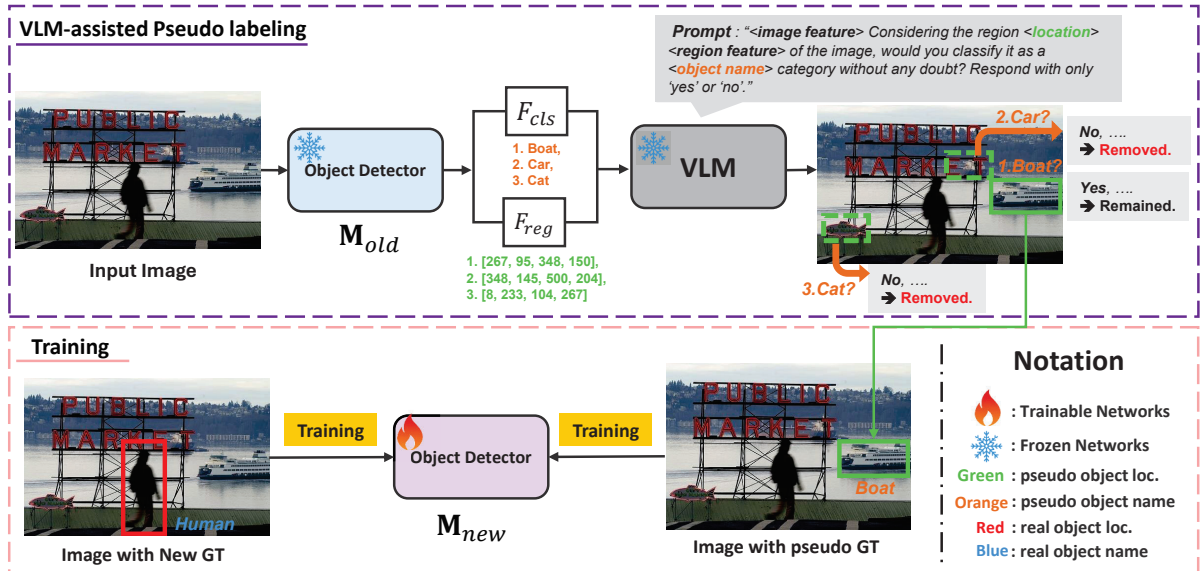
Figure 2. *Overview of the VLM-assisted Pseudo Labeling*: The sequence begins with the detector $\mathbf{M}_{old}$, applying pseudo labeling to identify potential objects (*e.g.* Boat, Car, and Cat within the input image), alongside their corresponding bounding box locations. Each identified object and its location are encapsulated into a prompt template. This template integrates placeholders for <image feature> and <region feature>, where the former is substituted with the overall image features and the latter with features corresponding to the specific region of interest. The prompts are classified by the VLM for reliability, using responses such as 'yes' or 'no' to verify each pseudo GT. Subsequently, the refined pseudo GTs are combined with new GTs from the new task for training the detector $\mathbf{M}_{new}$.

$\mathcal{Y}_d = \{\mathbf{y}_d^1, \mathbf{y}_d^2, \ldots, \mathbf{y}_d^L\}$. Here, L represents the number of data points for task $\mathcal{T}_d$, with d serving as the index for each task. Each annotation vector $\mathbf{y}_d^l$ for an image x_d^l can encapsulate multiple object instances, represented as $\mathbf{y}_d^l = \{(c_1, \mathbf{B}_1), \ldots, (c_K, \mathbf{B}_K)\}$, where K denotes the number of objects in an image l, and each c_k and $\mathbf{B}_k$ denote a category and bbox coordinates of the k-th object, respectively. This approach is aligned with the conventional CIOD configuration as prior works [23, 35, 36, 78].

Our VLM-PL consists of two primary components: 1) A method for extracting pseudo GTs from the pre-trained model (Section 4.2), and 2) A technique for refining these pseudo GTs through the VLM model (Section 4.3). A comprehensive overview can be seen in Figure 2.

4.2. Pseudo-labeling

The adaptation of the pseudo-labeling mechanism [17, 26, 36, 58] for transformer-based detectors [27, 91, 97] leverages the decoder's output from learned object queries. Each object query then goes through the classification branch ($\mathbf{F}_{cls}$) and regression branch ($\mathbf{F}_{reg}$), which include several MLPs at the final layer of the decoder. In the classification branch, queries calculate logits across all learned categories, including the background. These logits are then converted into scores ranging from 0 to 1 using a sigmoid function across all categories. This results in an output score matrix

$\hat{O} \in \mathbb{R}^{Q \times (\text{category}+1)}$ for class predictions and $\hat{\mathbf{B}} \in \mathbb{R}^{Q \times 4}$ for bboxes, where Q represents the total number of object queries. Building upon this framework, scores that exceed a predefined threshold, τ, are nominated for pseudo-labeling, ensuring the utilization of only the most reliable predictions in generating pseudo ground-truth (pseudo GT), denoted by $\hat{\mathbf{y}} = \{\hat{c}, \hat{\mathbf{B}}\}$. Here, $\hat{c}$ represents the predicted category name (*e.g.* 'person', 'kite' etc) from the highest-scoring category in each query, and $\hat{\mathbf{B}}$ provides the bbox coordinates for the corresponding query. In our study, drawing on the insights from previous research on [36, 58], we set the τ threshold to 0.3 as the optimal confidence score for maximizing performance.

Following the conventional pseudo-labeling strategy [36, 58, 78] for CIOD, we use the generated pseudo ground truth $\hat{y}$, in conjunction with the actual labels y, to train new coming categories. However, despite our efforts to alleviate forgetting, the reliance on predictions from previously trained models in pseudo-labeling can still potentially degrade performance. While this issue is less noticeable in dual-task scenarios, it becomes significantly more prominent when incrementally training on multiple tasks (*e.g.* 5+5+5+5, 10+5+5 in our experiments). Because, in multi-scenarios, the knowledge from previous models about earlier learned objects becomes increasingly blurred. This leads to reliance on incorrect predictions as pseudo GT for
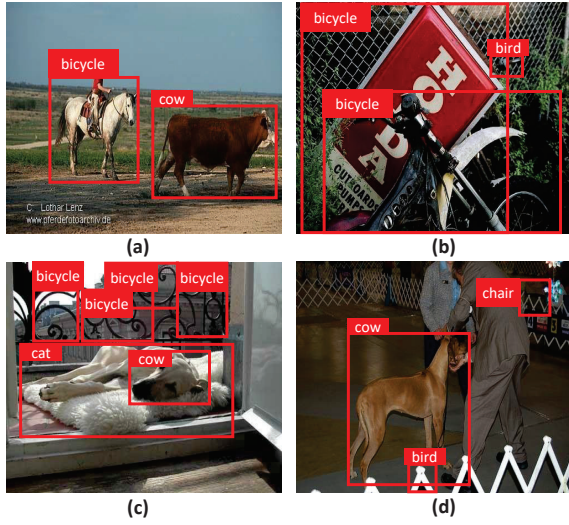
Figure 3. Illustration of incorrect pseudo GT examples generated during a multi-incremental learning scenario (*i.e.* 5+5+5+5) on the Pascal [20] dataset. (a) depicts an incorrect pseudo GT where a 'bicycle' is mislabeled; (b) shows both 'bicycle' and 'bird' misidentified; (c) highlights a case where all annotations are incorrect; and (d) indicates mislabeled 'cow' and 'bird' instances.

subsequent training. This not only creates a harmful cycle of error propagation, but also exacerbates performance degradation. Figure 3 shows incorrect pseudo GT examples of conventional pseudo-labeling strategies on the Pascal [20] dataset.

4.3. Vision-Language Model assistance

To address the aforementioned inherent challenges of pseudo-labeling in multi-task scenarios, we employ a Vision-Language Model (VLM), specifically Ferret [87]. Ferret's ability to perform question-answering tasks on specific image regions allows for a direct assessment of pseudo GT accuracy. This allows us to filter for consistent and accurate pseudo GT independent of the performance of previously used models.

Image Feature Extraction. Initially, the CLIP [69] model is employed to extract the overall image features, which are crucial for grasping the comprehensive context of the image. For regions requiring pseudo GT verification, we refer to the bbox coordinates, $\hat{B}$, and category names, $\hat{c}$, as described in Section 4.2. It is important to note that multiple instances of pseudo GT may exist within a single image. The bbox coordinates $\hat{B}$, originally in the normalized format of $[x, y, w, h]$, are transformed into real-size corner coordinates $[x_1, y_1, x_2, y_2]$. This accurately outlines each region of interest, making it easier to create binary masks (*i.e.* 0 is non-interesting region, 1 is interesting region) that emphasize the specific areas in the context of the overall image

generated through CLIP [69].

Prompt Formation. We use a predefined text prompt template to construct text queries for Ferret [87]. This template includes not only the text format but also the overall image visual feature and specific spatial information of the identified regions. We do this by using a *<placeholder>* as an LLM input. The template is: *"<image feature> Considering the region <location> <region feature> of the image, would you classify it as a <object name> category without any doubt? Respond with only 'yes' or 'no'."* This method represents a form of prompt-tuning, a technique that enables a Large Language Model (LLM) to perform classification tasks based on the provided context without additional model training [41, 49, 77]. In this template, the *<image feature>* placeholder represents the position of the overall image's CLIP visual feature. The *<location>* placeholder gets filled with the bbox coordinates, specifically $\hat{B}$ from $\hat{y}$, indicating the area of interest within the image. The *<object name>* corresponds to $\hat{c}$, signifying the object's category name. Finally, the *<region feature>* denotes the position of features extracted from the binary image mask, highlighting the part of the overall image feature that bbox $\hat{B}$ represents. The prompt-making process is iteratively done for each pseudo GTs in an image, modifying them accordingly.

Output. In the final stage, our method carefully selects pseudo GTs that Ferret validates with a 'yes'. We use this curated dataset to train the model on new coming classes. This approach ensures the training process uses consistently reliable pseudo GTs without the influence of a pre-trained model.

5. Experiments

5.1. Dataset and metrics

Dataset. Our research primarily uses the PASCAL VOC dataset [20]. Known for its 20 diverse object classes, it contains 9,963 images, split into 5,011 for training and 4,952 for testing. We also use the MS COCO 2017 dataset [52] as a additional experiment. The MS COCO [52], with its 80 object classes spread across 118,000 training images and 5,000 evaluation images, serves as a challenging benchmark.

Eval metrics. From PASCAL VOC, we use the mean average precision at the IOU threshold of 0.5 (AP_{50}). From MS COCO, we also use the mean average precision at various IOU thresholds (0.5:0.95, 0.5, 0.75) and object sizes: AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. Here, AP represents the mean value across all IOU thresholds from 0.5 to 0.95.

5.2. Implementation and experiments

Implementation details. In our study, we employ a methodology based on Deformable-DETR [97], utilizing

Table 1. Performance comparison of CIOD methods on Pascal VOC 2007 in *multi-scenario* settings, using AP_{50} (%) metric. Results for [6, 64, 78] are cited from ABR study [57]. The red arrow (↑) indicates performance improvement over the previous state-of-the-art (SoTA). Meanwhile, a red dash (-) indicates performance equivalent to the SoTA. The best results in each configuration are highlighted in **bold**.

Method	10+5+5 (3-tasks)			5+5+5+5 (4-tasks)		
	1-10 ($\mathcal{T}_1$)	11-20 ($\mathcal{T}_2 + \mathcal{T}_3$)	1-20	1-5 ($\mathcal{T}_1$)	6-20 ($\mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4$)	1-20
ILOD [78]	67.2	59.4	63.3	58.5	15.6	26.3
Faster ILOD [64]	68.3	57.9	63.1	55.7	16.0	25.9
MMA [6]	67.4	60.5	64.0	62.3	31.2	38.9
ABR [57]	68.7	67.1	**67.9**	64.7	56.4	58.4
DMD+IFD [35]	-	-	-	46.14	60.19	58.72
VLM-PL(Ours)	67.9	67.9	**67.9** -	64.5	68.4	**65.5** ↑6.78

Table 2. Performance comparison of CIOD methods on Pascal VOC 2007 in *dual-scenario* settings, utilizing the AP_{50} (%) metric. Results for [6, 26, 32, 33, 64, 78, 85, 95] are cited from the ABR study [57]. The red arrow (↑) signifies an improvement over the previous state-of-the-art. The best results in each configuration are highlighted in **bold**.

Method	19+1			15+5			10+10			5+15		
	1-19 ($\mathcal{T}_1$)	20 ($\mathcal{T}_2$)	1-20	1-15 ($\mathcal{T}_1$)	16-20 ($\mathcal{T}_2$)	1-20	1-10 ($\mathcal{T}_1$)	11-20 ($\mathcal{T}_2$)	1-20	1-5 ($\mathcal{T}_1$)	6-20 ($\mathcal{T}_2$)	1-20
ILOD [78]	69.8	64.5	69.6	72.5	58.5	68.9	69.8	53.7	61.7	61.0	37.3	43.2
Faster ILOD [64]	70.9	63.2	70.6	73.1	57.3	69.2	70.3	53.0	61.7	62.0	37.1	43.3
PPAS [95]	70.5	53.0	69.2	-	-	-	63.5	60.0	61.8	-	-	-
MVC [85]	70.2	60.6	69.7	69.4	57.9	66.5	66.2	66.0	66.1	-	-	-
MMA [6]	70.9	62.9	70.5	72.7	60.6	69.7	69.8	63.9	66.8	66.8	57.2	59.6
ORE [32]	69.4	60.1	68.9	71.8	58.7	68.5	60.4	68.8	64.6	-	-	-
OW-DETR [26]	70.2	62.0	69.8	72.2	59.8	69.1	63.5	67.9	65.7	-	-	-
Meta-ILOD [33]	70.9	57.6	70.2	71.7	55.9	67.8	68.4	64.3	66.3	-	-	-
ABR [57]	71.0	69.7	70.9	73.0	65.1	71.0	71.2	72.8	72.0	64.7	71.0	69.4
VLM-PL(Ours)	73.7	89.3	**73.6** ↑2.7	73.9	82.4	**72.4** ↑1.4	80.3	76.3	**78.3** ↑6.3	79.4	83.5	**81.0** ↑11.6

the pretrained ResNet-50 [29] as a backbone for extracting multi-scale features. We establish the number of object queries Q at 300. The AdamW [60] optimizer is utilized for model training, with a learning rate of 0.0002 and a weight decay parameter of 0.0001. We also set the gradient clipping parameter to 0.1. Other hyper-parameters are in line with our baseline [97]. All experimental procedures are conducted using 4 NVIDIA A100 GPUs, each with a batch size of 8. For evaluation, We use a single GPU. In the VLM assistance stage, we use the CLIP-ViT-L/14 [69] image encoder to extract image feature embeddings. We also use the Ferret-13B model [87] as a VLM, which is trained on public datasets [40, 67, 88] using the Vicuna [12] LLM. **Scenario setup.** In the *multi-scenarios* setting, the process begins with initial training on a base categories set $\mathcal{T}_1$. This is followed by subsequent training phases that introduce additional categories subsets $\mathcal{T}_{2:n}$, gradually expanding the model's knowledge. For instance, scenarios like $5(\mathcal{T}_1)+5(\mathcal{T}_2)+5(\mathcal{T}_3)+5(\mathcal{T}_4)$ and $10(\mathcal{T}_1)+5(\mathcal{T}_2)+5(\mathcal{T}_3)$ are referred to as 4-task and 3-task settings in our study. In these settings, we evaluate the cumulative learning performance after n phases across all integrated classes $\mathcal{T}_{1:n}$, and assess the performance on the baseline ($\mathcal{T}_1$) and the added tasks ($\mathcal{T}_2+\mathcal{T}_3$) and ($\mathcal{T}_2+\mathcal{T}_3+\mathcal{T}_4$) separately. In the *Dual-scenario* setting, training begins on subset $\mathcal{T}_1$, then introduces $\mathcal{T}_2$, represented as $\mathcal{T}_1+\mathcal{T}_2$. Evaluation covers all knowledge like

$1-20$ across all settings (*e.g.* $19+1, 15+5, 10+10, 5+15$), including an assessment of base knowledge (*i.e.* 19, 15, 10, 5). For COCO [52], the data configuration follows $\mathcal{T}_1 + \mathcal{T}_2$, focusing evaluation solely on comprehensive model performance.

5.3. Results and analysis

Multi-scenario. Table 1 demonstrates the outstanding performance of our method on the PASCAL [20] dataset, particularly in the complex 4-tasks scenario. Here, we achieved a 65.5% accuracy, marking a significant 6.78% improvement over the previous state-of-the-art (SoTA), DMD+IFD [35]. Furthermore, in the 10+5+5 scenario, we matched the ABR's [57] performance without the need for partial replay. These results underscore our model's ability to counteract catastrophic forgetting by leveraging VLM knowledge, thereby maintaining the robust performance of the pseudo-labeling strategy across various scenarios.

Dual-scenario. Despite our goal of solving *multi-scenario* challenges, our VLM-assisted pseudo-labeling strategy (VLM-PL) also demonstrates significant performance in single incremental task situations, as shown in Table 2 and Table 3. In Table 2, compared to OW-DETR [26], which uses the same baseline Deformable DETR [97], our method surpasses it in both initial class performance and overall (*i.e.* 1-20 result) outcome on pascal [20]. It also has out-

Table 3. Performance comparison of CIOD methods on COCO in *dual-scenario* settings of 70+10, using COCO AP metrics. Results for [6, 23, 43, 51, 65] are cited from the CL-DETR study [58]. The red arrow (↑) signifies an improvement over the previous state-of-the-art. The best performance is highlighted in **bold** at each evaluation. † denotes the version without the replay strategy (*e.g.* partial replay, synthetic replay).

Scenarios	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
70 + 10	LWF [51]	7.1	12.4	7.0	4.8	9.5	10.0
	RILOD [43]	24.5	37.9	25.7	14.2	27.4	33.5
	MMA [6]	30.2	52.1	31.5	-	-	-
	ABR [57]	31.1	52.9	32.7	-	-	-
	SID [65]	32.8	49.0	35.0	17.1	36.9	44.5
	ERD [23]	34.9	51.9	37.4	18.7	38.8	45.5
	CL-DETR† [58]	35.8	53.5	39.5	19.4	41.5	46.1
	SDDGR† [36]	38.6	56.2	42.1	22.3	42.1	50.6
	VLM-PL(Ours)	**39.8** ↑1.2	**58.2** ↑2.0	**43.3** ↑1.2	**23.3** ↑1.0	**43.5** ↑1.4	**51.6** ↑1.0

Table 4. Ablation experiment results of different pseudo-labeling strategies on the Pascal [20] dataset with *multi-scenario* (*i.e.* 5+5+5+5) are presented. The values are mean Average Precision (AP_{50}, %), and we use the ferret [87] as a VLM. The best performance is highlighted in **bold** at each evaluation.

Method	Number of Classes			
	5	10	15	20
Original pseudo labeling	82.6%	76.8%	73.6%	62.4%
VLM-assist pseudo labeling	82.6%	**79.4%**	**77.9%**	**65.5%**

Table 5. Ablation results of different VLMs [18, 87] on the Pascal [20] dataset with *multi-scenario* (*i.e.* 5+5+5+5) are presented. The values are mean Average Precision (AP_{50}, %). The best performance is highlighted in **bold** at each evaluation.

Method	Number of Classes			
	5	10	15	20
InternLM2 [18]	82.6%	75.7%	76.5%	63.9%
Ferret [87]	82.6%	**79.4%**	**77.9%**	**65.5%**

Table 6. Ablation study of the influence of VLM [87] capacity on the Pascal [20] dataset with *multiple scenarios* (*i.e.* 5+5+5+5) is presented. The values represent mean Average Precision (AP_{50}, %). The best performance is highlighted in **bold** for each evaluation.

Method	Number of Classes			
	5	10	15	20
Ferret 7B [87]	82.6%	78.7%	77.0%	64.7%
Ferret 13B [87]	82.6%	**79.4%**	**77.9%**	**65.5%**

evaluated task (*e.g.* T3 model evaluation includes T1, T2, and T3 classes).

Pseudo-labeling. In Table 4, we evaluate the effectiveness of using our VLM-PL approach, as compared to conventional pseudo-labeling methods. We observed that with the progression of tasks, the application of the Ferret model for refining pseudo GTs led to improvements in performance about all each tasks, demonstrating gains of 3.4%, 4.3%, and 3.1% for T2, T3, and T4 respectively. These enhancements indicate that as tasks accumulate, there is a tendency for the generation of incorrect pseudo GTs, likely due to a blurring of the knowledge initially trained into the model. Nevertheless, the VLM [87] has consistently demonstrated its ability to correct these inaccuracies, emphasizing its crucial role in refining the pseudo-labeling strategy.

Comparative analysis of VLMs. VLMs employ various approaches, one of which is the Ferret [87] model. This model allows detailed conversation about a specific location based on the whole image context. As indicated in Table 5, we compared the Ferret model with another VLM model, InternLM2 [18]. It has delivered state-of-the-art results in question-answering tasks using entire images only. To use InternLM2 [18] for classifying a specific image region, we cropped the image based on bboxes, identical to Ferret's approach in ours. Consequently, in the incremental tasks T2, T3, and T4, Ferret [87] outperforms InternLM2 [18] by

standing performance against the recent SoTA, ABR [57]. Our performance in scenarios 19+1, 15+5, 10+10, and 5+15 (73.7%, 73.9%, 80.3%, and 79.4%, respectively) demonstrates our method's effectiveness in mitigating forgetting, even without using directly any previous data (replay-free). Furthermore, Table 3 shows that our VLM-PL method outperforms various other approaches on COCO [52], including those using the same baseline [97] and excluding replay strategy. Consequently, we can check the adaptability of our strategy in both dual and multi-scenario contexts.

5.4. Ablations

In our ablation study, we evaluate our components in multi-scenario settings. Specifically, we use a 4-tasks scenario on the Pascal with the metric AP_{50}. It's important to note that each result represents the cumulative performance up to the
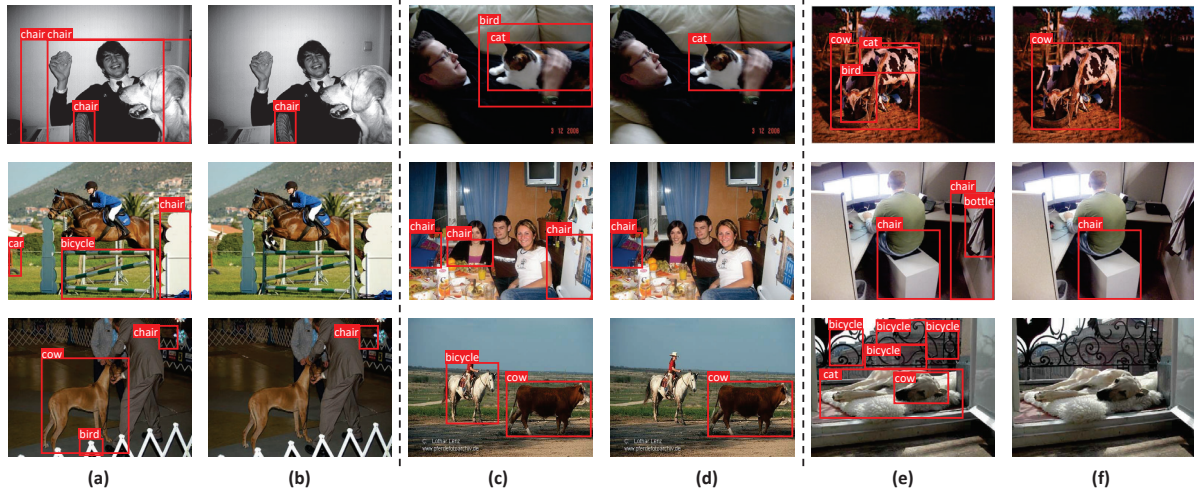
Figure 4. Qualitative results of both conventional pseudo labeling, as used in OW-DETR [26] and SDDGR [36], and VLM-assisted pseudo labeling in a multi-incremental scenario (for example, 5(T1)+5(T2)+5(T3)+5(T4) on the Pascal dataset) are presented here. The effects of VLM assistance can be observed from (a) to (b), (c) to (d), and (e) to (f). This is especially noticeable in (a) to (b) of the second row and (e) to (f) of the third row, which indicate that all pseudo GTs are incorrect.

3.7%, 1.4%, and 1.6% respectively. This indicates Ferret's effectiveness, which allows querying both the entire image feature and specific parts.

VLM capacity impact. VLM generally relies on the performance of the LLM. Generally, LLM's larger capacities and more comprehensive training data lead to better outcomes in VLM. We investigate the impact of model capacity on the refining process of pseudo GTs. Table 6 shows the results from experiments using Ferret [87] 7B and Ferret [87] 13B, each using Vicuna [12] 7B and Vicuna [12] 13B, respectively. Importantly, the larger 13B model provides modest but consistent improvements across all tasks, with performance increases of 0.7%, 0.9%, and 0.8%. These improvements mean the LLM capacity can affect VLM performance. As a result, we chose the 13B model for our study.

Qualitative results. As illustrated in Figure 3, the advantage of our method is particularly noticeable in multi-scenario settings, where incorrect pseudo GTs are prevalent. The images in columns (b), (d), and (f) clearly demonstrate the thorough removal of all incorrect pseudo GTs through VLM assitance. Furthermore, in row 1, images a, c, and e vividly depict this challenge, showcasing multiple pseudo GTs for a single object. Nevertheless, through the VLM-assisted refining process, our approach successfully consolidates these into a single accurate GT for each object, demonstrating the precision and reliability of VLM assistance in the pseudo GT refining process.

6. Conclusions And Limitations

In this study, we introduced VLM-PL, a new approach that leverages Vision-Language Model to address the rapid performance decline in pseudo-labeling strategy under the influence of previously trained models in multiple incrementally class incremental object detection. Our VLM-PL not only demonstrates robust performance across multi and dual scenarios but also achieves state-of-the-art results in conditions without a replay strategy.

Limitations. While our approach effectively addresses pseudo-labeling reliability in various scenarios, it does have limitations. Its performance may be limited when there are few images available for pseudo-labeling, or when new images contain few objects from previously learned categories. Despite these challenges, our approach significantly outperforms not only techniques that do not utilize pseudo-labeling but also surpasses prior works [36, 58] that generate pseudo GT leveraging trained model knowledge. Furthermore, we believe it also has the potential to improve methods of preventing catastrophic forgetting when used in conjunction with partial replay techniques.

7. Acknowledgements.

References

[1] Manoj Acharya, Tyler L Hayes, and Christopher Kanan. Rodeo: Replay for online object detection. In *BMVC*, 2020. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 3

[3] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *CVPR*, 2021. 2

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. In *arXiv*, 2020. 2, 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3

[6] Fabio Cermelli, Antonino Geraci, Dario Fontanel, and Barbara Caputo. Modeling missing annotations for incremental learning in object detection. In *CVPRW*, 2022. 2, 6, 7

[7] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018. 2

[8] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. In *arXiv*, 2019. 1

[9] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. 3

[10] Kean Chen, Weiyao Lin, Jianguo Li, John See, Ji Wang, and Junni Zou. Ap-loss for accurate one-stage object detection. In *TPAMI*, 2020. 2

[11] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3, 6, 8

[13] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. In *NeurIPS*, 2020. 1, 2

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[15] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *ICCV*, 2021. 1

[16] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. In *arXiv*, 2022. 2

[17] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Open world detr: transformer based open world object detection. In *arXiv*, 2022. 2, 3, 4

[18] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. In *arXiv*, 2024. 2, 7

[19] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 3

[20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *IJCV*, 2010. 5, 6, 7

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 3

[22] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, 2021. 2

[23] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *CVPR*, 2022. 2, 4, 7

[24] Rui Gao and Weiwei Liu. Ddgr: Continual learning with deep diffusion-based generative replay. In *ICML*, 2023. 1, 2

[25] Yunhui Guo, Mingrui Liu, Tianbao Yang, and Tajana Rosing. Improved schemes for episodic memory-based lifelong learning. In *NeurIPS*, 2020. 2

[26] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *CVPR*, 2022. 2, 3, 4, 6, 8

[27] Yu Hao, Yanwei Fu, Yu-Gang Jiang, and Qi Tian. An end-to-end architecture for class-incremental object detection with knowledge distillation. In *ICME*, 2019. 1, 2, 4

[28] Chen He, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exemplar-supported generative reproduction for class incremental learning. In *BMVC*, 2018. 1

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[31] Quentin Jodelet, Xin Liu, Yin Jun Phua, and Tsuyoshi Murata. Class-incremental learning using diffusion model for distillation and replay. In *ICCVW*, 2023. 1, 2

[32] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 6

[33] KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. In *TPAMI*, 2021. 6

[34] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. In *arXiv*, 2016. 1, 2

[35] Mengxue Kang, Jinpeng Zhang, Jinming Zhang, Xiashuang Wang, Yang Chen, Zhe Ma, and Xuhui Huang. Alleviating catastrophic forgetting of incremental object detection via within-class and between-class knowledge distillation. In *ICCV*, 2023. 2, 3, 4, 6

[36] Junsu Kim, Hoseong Cho, Jihyeon Kim, Yihalem Yimolal Tiruneh, and Seungryul Baek. Sddgr: Stable diffusion-based deep generative replay for class incremental object detection. In *arXiv*, 2024. 2, 4, 7, 8

[37] Junsu Kim, Sumin Hong, Chanwoo Kim, Jihyeon Kim, Yihalem Yimolal Tiruneh, Jeongwan On, Jihyun Song, Sunhwa Choi, and Seungryul Baek. Class-wise buffer management for incremental object detection: An effective buffer training strategy. In *ICASSP*, 2024. 2

[38] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. In *PNAS*, 2017. 1, 2

[39] Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. In *arXiv*, 2021. 2

[40] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2017. 6

[41] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv*, 2021. 5

[42] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023. 3

[43] Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry Heck. Rilod: Near real-time incremental learning for object detection at the edge. In *SEC*, 2019. 7

[44] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 2

[45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *arXiv*, 2023. 3

[47] 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 *CVPR*, 2022. 3

[48] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020. 2

[49] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv*, 2021. 5

[50] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 2

[51] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *TPAMI*, 2017. 1, 2, 7

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 6, 7

[53] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3

[55] Liyang Liu, Zhanghui Kuang, Yimin Chen, Jing-Hao Xue, Wenming Yang, and Wayne Zhang. Incdet: In defense of elastic weight consolidation for incremental object detection. In *TNNLS*, 2020. 2

[56] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2022. 3

[57] Yuyang Liu, Yang Cong, Dipam Goswami, Xialei Liu, and Joost van de Weijer. Augmented box replay: Overcoming foreground shift for incremental object detection. In *ICCV*, 2023. 2, 6, 7

[58] Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *CVPR*, 2023. 2, 3, 4, 7, 8

[59] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, 2017. 1, 2

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *arXiv*, 2017. 6

[61] Yichen Lu, Mei Wang, and Weihong Deng. Augmented geometric distillation for data-free incremental person reid. In *CVPR*, 2022. 1, 2

[62] Inyoung Paik, Sangjun Oh, Taeyeong Kwak, and Injung Kim. Overcoming catastrophic forgetting by neuron-level plasticity control. In *AAAI*, 2020. 1, 2

[63] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *CVPR*, 2019. 2

[64] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. In *Pattern Recognition*, 2020. 2, 6

[65] Can Peng, Kun Zhao, Sam Maksoud, Meng Li, and Brian C Lovell. Sid: Incremental learning for anchor-free object detection via selective and inter-related distillation. In *CVIU*, 2021. 7

[66] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[67] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 6

[68] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*, 2020. 2

[69] 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 *ICML*, 2021. 2, 3, 5, 6

[70] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 1, 2

[71] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. In *arXiv*, 2018. 3

[72] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 3

[73] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. In *Connection Science*, 1995. 1, 2

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[75] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, 2017. 1

[76] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, 2017. 1, 2

[77] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv*, 2020. 5

[78] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017. 2, 4, 6

[79] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *CVPR*, 2021. 1, 2

[80] Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. In *arXiv*, 2022. 2

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[82] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[83] Chenshen Wu, Luis Herranz, Xialei Liu, Joost Van De Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *NeurIPS*, 2018. 1, 2

[84] Ye Xiang, Ying Fu, Pan Ji, and Hua Huang. Incremental learning using conditional adversarial networks. In *NeurIPS*, 2019. 1, 2

[85] Dongbao Yang, Yu Zhou, Aoting Zhang, Xurui Sun, Dayan Wu, Weiping Wang, and Qixiang Ye. Multi-view correlation distillation for incremental object detection. In *Pattern Recognition*, 2022. 6

[86] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 3

[87] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024. 2, 3, 5, 6, 7, 8

[88] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 6

[89] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*, 2023. 3

[90] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017. 1, 2

[91] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2022. 3, 4

[92] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *WACV*, 2020. 1, 2

[93] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[94] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 3

[95] Wang Zhou, Shiyu Chang, Norma Sosa, Hendrik Hamann, and David Cox. Lifelong object detection. In *arXiv*, 2020. 6

[96] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *arXiv*, 2023. 3

[97] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2, 3, 4, 5, 6, 7